

PR #32302 完整报告

sgl-project/sglang

chore: bump mooncake version to 0.3.12.post1

合并时间: 2026-07-25 14:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32302>

执行摘要

- 一句话: Mooncake 版本从 0.3.11.post1 升级至 0.3.12.post1
- 推荐动作: 值得合并, 属于常规依赖升级。建议关注 4-gpu-gb300 CI 失败原因, 确认非本 PR 引入。

功能与动机

保持 Mooncake 依赖版本与 upstream 同步, 获取最新 bugfix 和性能改进。PR body 未提供详细动机, 但结合历史可见 Mooncake 是 SGLang 分解式推理的核心组件。

实现拆解

1. 修改 scripts/ci/cuda/ci_install_dependency.sh: 将 MOONCAKE_VERSION 从 "0.3.11.post1" 改为 "0.3.12.post1"。
2. 修改 docker/Dockerfile: 将 MOONCAKE_VERSION 参数从 0.3.11.post1 改为 0.3.12.post1。

关键文件:

- scripts/ci/cuda/ci_install_dependency.sh (模块 CI 脚本; 类别 infra; 类型 infrastructure) : CI 安装脚本中 Mooncake 版本号的定义位置, 直接影响 CI 环境中的依赖版本。
- docker/Dockerfile (模块 Docker; 类别 infra; 类型 infrastructure) : Docker 镜像构建中定义 Mooncake 版本, 影响所有基于该 Dockerfile 的容器镜像。

关键符号: 未识别

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

CI 安装脚本中 Mooncake 版本号的定义位置, 直接影响 CI 环境中的依赖版本。

```
# scripts/ci/cuda/ci_install_dependency.sh
install_extra_deps() {
    # 更新 Mooncake 版本到 0.3.12.post1
    MOONCAKE_VERSION="0.3.12.post1"
    NIXL_VERSION="1.3.0"
```

```
# ...  
}
```

评论区精华

无 review 评论， 仅作者触发 CI rerun 测试群组（disaggregation、hichache）。CI 结果中 4-gpu-gb300 平台测试持续失败， 但未进入讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅涉及版本号字符串变更， 无逻辑改动。但若 Mooncake 0.3.12 引入兼容性问题， 可能导致分解式推理或 hiCache 功能异常。CI 中 4-gpu-gb300 测试失败需关注是否与新版本相关。
- 影响：影响范围小， 仅限使用 Mooncake 的分解式推理和 hiCache 功能； 对所有用户透明， 升级后自动使用新版本。
- 风险标记：CI 测试持续失败（需确认是否相关）

关联脉络

- PR #31144 [PD] Prevent decode scheduler from blocking on ZMQ sends to a stalled prefill peer: 共享 Mooncake 相关修改历史， 都是分解式推理的依赖调整。
- PR #29326 feat(hichache): Add shared memory allocator for host KV cache: hiCache 依赖 Mooncake 存储后端， 版本升级可能影响其功能。