

PR #36493 完整报告

sgl-project/sglang

chore: bump mooncake version to 0.3.13

合并时间: 2026-08-26 22:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36493>

执行摘要

- 一句话: 升级 Mooncake 至 0.3.13, 并适配 TCP 队列限制
- 推荐动作: 值得关注, 尤其是 Mooncake 新版本的 TCP admission 控制行为对测试的影响。建议阅读测试文件中的注释和变更, 理解如何通过环境变量适配。

功能与动机

Mooncake v0.3.13 已正式发布, 需要升级固定版本, 确保 Docker 镜像和 CI 使用正式版。同时, 新版本引入了 TCP admission 控制, 可能导致特定测试 (Kimi-Linear PD+DCP4) 因默认队列上限过小而失败, 需要适配。

实现拆解

1. 升级版本声明: 在 docker/Dockerfile 中修改 ARG MOONCAKE_VERSION 为 0.3.13; 在 scripts/ci/cuda/ci_install_dependency.sh 中修改 MOONCAKE_VERSION 变量, 保持一致。
2. 适配测试环境: 在 test/registered/disaggregation/test_kimi_linear_pd_dcp4.py 的 setUpClass 中, 于启动服务器前设置环境变量 MC_TCP_MAX_QUEUED_TRANSFERS_PER_PEER 和 MC_TCP_MAX_PENDING_ADMISSIONS_PER_PEER 为 65535, 以应对 TP4/EP4 到 TP4/DCP4 的 fan-out 超过默认限制的问题。并在 tearDownClass 中清理这些环境变量, 避免污染其他测试。
3. 无配置或部署配套改动: 除上述硬编码环境变量外, 无其他配置变化。

关键文件:

- test/registered/disaggregation/test_kimi_linear_pd_dcp4.py (模块 测试; 类别 test; 类型 test-coverage; 符号 tearDownClass): 适配 Mooncake 0.3.13 的 TCP 行为变更, 是本次升级中唯一涉及逻辑调整的文件。
- docker/Dockerfile (模块 部署脚本; 类别 infra; 类型 infrastructure): 升级 Mooncake 版本, 影响 Docker 镜像构建。
- scripts/ci/cuda/ci_install_dependency.sh (模块 CI; 类别 infra; 类型 infrastructure): 升级 CI 依赖安装脚本中的 Mooncake 版本, 确保 CI 使用新版本。

关键符号: tearDownClass

关键源码片段

test/registered/disaggregation/test_kimi_linear_pd_dcp4.py

适配 Mooncake 0.3.13 的 TCP 行为变更，是本次升级中唯一涉及逻辑调整的文件。

```
@classmethod
def setUpClass(cls):
    super().setUpClass()
    # Mooncake >=0.3.13 对于 TCP
    传输增加了每对端队列上限，并在溢出时直接失败（而不是退避）。
    # 本测试的 TP4/EP4 -> TP4/DCP4 扇出会超过默认的 1024+1024 限制，导致 KV 传输失败。
    # 因此这里提高限制，并在 launch_all() 之前设置，确保服务器子进程继承。
    os.environ["MC_TCP_MAX_QUEUED_TRANSFERS_PER_PEER"] = "65535"
    os.environ["MC_TCP_MAX_PENDING_ADMISSIONS_PER_PEER"] = "65535"
    cls._collect_monolithic_references()
    cls.launch_all()

@classmethod
def tearDownClass(cls):
    # 清理环境变量，避免对其他测试造成影响。
    os.environ.pop("MC_TCP_MAX_QUEUED_TRANSFERS_PER_PEER")
    os.environ.pop("MC_TCP_MAX_PENDING_ADMISSIONS_PER_PEER")
    super().tearDownClass()
```

评论区精华

无正式 review 评论，但 issue 评论中作者多次触发 rerun 相关测试（disaggregation、hcache、kimi linear pd dcp4），均通过，说明测试验证是针对新版本行为的重要保证。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 回归风险：Mooncake 0.3.13 的 TCP 行为变更可能影响其他使用 Mooncake 的场景（如 hcache、disaggregation），需要关注 CI 中相关测试是否全部通过。
 2. 测试环境变量作用域：在测试中设置环境变量可能影响其他测试（通过在 tearDownClass 中清理，风险已降低）。
 3. 版本兼容性：升级可能带来潜在的 API 变化，但本 PR 未涉及源码调用，风险较低。 - 影响：影响范围：部署脚本和 Docker 镜像（版本升级），以及特定测试（Kimi-Linear PD+DCP4）的稳定性。对用户和系统影响较小，主要影响 CI 和部署环境的一致性。团队需确保所有使用 Mooncake 的场景与 0.3.13 兼容。 - 风险标记：依赖升级，测试环境变量，CI 回归风险

关联脉络

- 暂无明显关联 PR