

PR #33462 完整报告

sgl-project/sglang

[AMD] Bump mori to latest in sglang

合并时间: 2026-08-05 06:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33462>

执行摘要

- 一句话: 升级 ROCm 镜像 MORI pin, 修复 EP16 分离启动不稳定
- 推荐动作: 该 PR 适合快速浏览, 重点不在代码而在上下游联动: 阅读 ROCm/mori#493 可理解 bootstrap 超时为何是启动不稳定的根因。值得关注的设计决策是让默认超时从约 10 秒提升到 300 秒, 并统一四个超时点, 这是分布式 bootstrap 超时配置的一个合适示例。若团队维护 AMD/ROCm 镜像, 应留意多节点验证补跑。

功能与动机

PR body 说明: 当前 pin 早于 bootstrap-timeout 修复 (ROCm/mori#493), 该修复通过 `MORI_BOOTSTRAP_TIMEOUT` 使 MORI rendezvous 超时可配置并解决间歇性 EP16 bootstrap 失败。上游 issue #493 指出 `socket_bootstrap.cpp` 中四个硬编码超时点合计约 10 秒, 根 rank 慢启动时会令非根 rank 耗尽重试并报错 `failed to connect to root after 50 retries`; 提交信息补充: EP16 启动器已设置 `MORI_BOOTSTRAP_TIMEOUT=300`, 但旧 pin 完全忽略该变量, 提升到 `12d1bc32` 后配置才真正生效。

实现拆解

1. 变更入口: 仅修改 `docker/rocm.Dockerfile` 中 ARG MORI_COMMIT, 这是 ROCm 镜像构建时固定 MORI 源码版本的参数。
2. 关键变更: MORI_COMMIT 从 `f7e6ac68...` 更新为 `12d1bc32...`, 后者包含 ROCm/mori#493。
3. 原因与效果: 旧 pin 中 bootstrap 超时硬编码为约 10 秒, 不可外部配置; 新 pin 将所有超时点统一从 MORI_BOOTSTRAP_TIMEOUT 读取 (默认 300 秒, 非法值回退), 并修正与实际行为不符的注释, 从而让 EP16 启动器设定的 300 秒预算生效。
4. 影响范围: 仅影响 `ENABLE_MORI=1` 的 AMD/ROCm 分离部署, 源码和测试逻辑零改动。
5. 验证配套: 无新增测试文件, 依赖 CI 回归; 单节点 MORI 回归在 head `fc35c2bd` 通过, 多节点 DSV4 2P1D 验证在合并前未重新触发。

关键文件:

- `docker/rocm.Dockerfile` (模块 镜像配置; 类别 infra; 类型 configuration): 唯一变更文件, 通过 MORI_COMMIT 将 MORI 依赖 pin 提升到包含可配置 bootstrap 超时修复的提交, 解决 AMD EP16 分离启动失败。

关键符号：未识别

关键源码片段

docker/rocm.Dockerfile

唯一变更文件，通过 `MORI_COMMIT` 将 MORI 依赖 pin 提升到包含可配置 bootstrap 超时修复的提交，解决 AMD EP16 分离启动失败。

```
# MORI 是 AMD 环境下的分离通信后端（负责 MoE all-to-all 与 KV 传输）。
# 该 ARG 固定 MORI 源码 commit；本次提升到 12d1bc32，包含 ROCm/mori#493，
# 使 bootstrap rendezvous 超时改为可配置环境变量 MORI_BOOTSTRAP_TIMEOUT（默认 300 s），
# 修复 EP16 启动时慢节点因旧版约 10 s 硬编码超时而报 "Failed to setup communication ring"
的间歇性失败。
ARG ENABLE_MORI=0
ARG NIC_BACKEND=none
ARG MORI_REPO="https://github.com/ROCm/mori.git"
ARG MORI_COMMIT="12d1bc32d0c93dcd5062e74f4e0f772e36e1aac4"
```

评论区精华

唯一实质讨论来自关联 issue 评论，bingxche 报告验证状态：单节点 MORI 回归套件在精确 head `fc35c2bd` 上通过（MORI EP small + MI35X MORI disaggregation）；多节点 DSV4 2P1D 验证因分支重写被取消，尚未重新触发。审查人 Duyi-Wang 直接 APPROVED，没有 review 评论。其余评论为已失效的 gemini-code-assist 机器人提示。

- MORI pin 的验证状态与多节点回归 (testing)：单节点复杂度已通过，多节点验证留待后续补跑，PR 已合并。

风险与影响

- 风险：（1）依赖升级风险：f7e6ac68 到 12d1bc32 之间除 #493 外可能存在其他上游提交，且多节点 DSV4 2P1D 未在合并前复验，可能有残余回归。（2）超时行为变化：MORI_BOOTSTRAP_TIMEOUT 默认 300 秒，真实连接失败时错误路径感知延迟变长；上游声明为有界等待，不会无限挂起。（3）镜像构建影响：需要重建 ROCm 镜像才能生效，已发布镜像在重建前不受影响。（4）影响范围：仅限 ENABLE_MORI=1 的 AMD 分离部署，CUDA/NIXL 后端不受影响。
- 影响：用户侧：AMD/ROCm 上使用 EP16 分离（如 DeepSeek 宽 EP16）的部署将不再间歇性遭遇 bootstrap 竞争失败，启动更稳定。系统侧：ROCm Docker 镜像的依赖被更新，需重新构建验证。团队侧：为后续 AMD 分离链路的 bootstrap 超时调试提供了可配置手段，降低同类问题排查成本。整体影响程度中等，主要影响 AMD 用户群，但代码改动极小。
- 风险标记：依赖 pin 升级，AMD/ROCm 路径，多节点验证未重新运行

关联脉络

- PR #33523 [npu] [bugfix] Fix PD-disaggregation error: 同属分离链路稳定性修复，但作用于 NPU 连接层；本 PR 从 ROCm 镜像依赖角度解决 AMD 上的同类 bootstrap 问题，

反映多硬件分离链路的维护路线。