PR #34476 完整报告

sgl-project/sglang

[AMD][DI][CI] Add GLM-5.2 MXFP4 wide-EP16 2P1D nightly recipes

合并时间：2026-08-12 18:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34476

执行摘要

- 一句话：新增 GLM-5.2 MXFP4 wide-EP16 2P1D 夜间测试配方
- 推荐动作：值得快速浏览，作为了解 SGLang 在 AMD 上 wide-EP 和 MTP 配置的参考。重点关注 wide_ep 参数、MoRI 环境变量和 MTP 配置块。注意检查模型路径一致性。

功能与动机

扩展 GLM-5.2 MXFP4 在 AMD 平台的测试覆盖，参考 DSV4-Pro / Kimi Oren 的 wide-EP 配置经验，验证 wide-EP16 在 decode 阶段的收益，并补充 MTP（NextN）场景下的准确性与性能数据。PR body 中明确提到这是对 #32120 中 1P1D GLM-5.2 recipes 的扩展。

实现拆解

1. 新增配方文件：在 scripts/ci/slurm/recipes/mi355x-fp4/glm52/1k1k/ 下创建 2p1d-ep16.yaml（82 行）和 2p1d-ep16-mtp.yaml（90 行）。两个文件均定义 2 prefill workers + 1 decode worker 的 4 节点拓扑，prefill 使用 TP8/EP8/DP8，decode 使用 TP16/EP16/DP16。
2. 核心配置项：runtime 部分设置 moe_a2a_backend: mori 和 kv_transfer_backend: mori，并配置 wide_ep 块（kv_cache_dtype: fp8_e4m3、prefill/decode 的 mem_fraction_static 等）。模型相关参数包括 --reasoning-parser glm45、--tool-call-parser glm45、--disable-shared-experts-fusion，并显式声明 DSA attention 自动选择（无需 --attention-backend）。
3. MTP 扩展：MTP 版本额外启用 mtp 块（num_steps: 3、eagle_topk: 1、num_draft_tokens: 4），用于验证内置 NextN 头的推测解码场景。
4. 注册夜间任务：在 scripts/ci/slurm/nightly-configs.yaml 末尾追加 glm52-fp4-mi355x-ep16-sglang 和 glm52-fp4-mi355x-ep16-mtp-sglang 两个配置块，指定模型路径、运行器、多节点与 disagg 标志、seq-len 配置（ISL/OSL 均为 1024）以及并发列表。
5. 测试配套：无源码或单元测试变更，仅依赖夜间 CI 的 GSM8K few-shot 准确率门禁（1319 题，阈值 0.91）。实测基础版 0.949、MTP 版 0.947，均超过阈值。

关键文件：

- scripts/ci/slurm/recipes/mi355x-fp4/glm52/1k1k/2p1d-ep16.yaml（模块 CI 配置；类别 infra；类型 configuration）：新增的 wide-EP16 2P1D 基础配方，定义了 prefill EP8 +

decode EP16 的 4 节点拓扑和 MoRI 配置，是本次变更的核心。

- scripts/ci/slurm/recipes/mi355x-fp4/glm52/1k1k/2p1d-ep16-mtp.yaml（模块 CI 配置；类别 infra；类型 configuration）：在基础配方上增加 MTP（NextN）推测解码配置，验证 wide-EP16 与 MTP 组合场景。
- scripts/ci/slurm/nightly-configs.yaml（模块 CI 配置；类别 infra；类型 configuration）：在夜间任务注册表中新增两个配置块，使新配方进入 CI 调度。

关键符号：未识别

评论区精华

该 PR 无 review 评论，HaiShaw 直接批准（APPROVED）。PR body 中说明了配置与 Kimi EP16 配方的差异：DSA attention 自动选择、GLM 解析器、禁用共享专家融合。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 模型路径疑似不一致：nightly-configs.yaml 中 model 字段为 amd/GLM-5.2-MXFP4，但 model_path 指向 models--amd--GLM-5.1-MXFP4，可能为笔误或镜像映射，需确认实际加载的模型版本。
  2. 阈值覆盖风险：GSM8K 阈值 0.91 低于实测（0.949/0.947），留有一定余量，但若未来硬件或软件回归，可能出现误报通过。
  3. CI 资源消耗：4 节点（MI355X）配置资源占用大，夜间任务失败可能影响其他 CI 任务排队。
  4. 配置维护成本：新增两个配方依赖特定镜像标签和 MoRI 网络环境，若镜像或网络拓扑变化，需同步更新。 - 影响：对用户无直接影响（纯 CI 基础设施）。对团队而言，增强了 AMD MI355X 平台上 GLM-5.2 模型在 wide-EP16 2P1D 场景下的回归覆盖，为后续优化提供基准数据。对系统影响有限，仅增加夜间任务数量。 - 风险标记：模型路径疑似版本不一致，新增 CI 资源消耗，阈值余量需关注

关联脉络

- PR #32120 Add GLM-5.2 MXFP4 1P1D recipes: 本 PR 明确说明是对 #32120 中 1P1D GLM-5.2 recipes 的扩展，引入了 2P1D wide-EP16 变体。