

PR #31289 完整报告

sgl-project/sglang

[CI] Lower GLM-5.2 NVFP4 MTP speed threshold

合并时间: 2026-07-15 15:39

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31289>

执行摘要

- 一句话: 降低 GLM-5.2 NVFP4 MTP CI 速度阈值
- 推荐动作: 变更合理且必要, 直接解决 CI 误报问题。建议阅读者关注 PR#31001 的回退计划 (flashinfer v0.5.15), 届时可能需要再次调整阈值或恢复原值。

功能与动机

PR#31001 合入后, `test_dsa_glm52_nvfp4_tp_mtp.py` 在 CI 中实测 tps 约 270, 最低达 254.43, 接近或低于现有阈值 300, 导致 CI 测试不稳定。开发者在本地非容器环境验证通过但接近阈值, 而 CI 容器因 debug env vars (如 `SGLANG_ENABLE_ASYNC_ASSERT`) 和容器化调度引入额外开销, 实际 tps 更低。为消除误报, 需要将阈值下调至 250。

实现拆解

1. 修改测试配置: 在 `test/registered/models_e2e/test_dsa_glm52_nvfp4_tp_mtp.py` 中将 `bs_1_speed_thres` 从 300 改为 250。
2. 验证方式: 开发者分别在带 CI env vars 和无 CI env vars 环境下运行同一测试, 测得 tps 分别为 313.23 和 357.05, 并确认阈值 250 能覆盖 CI 容器环境的最差情况 (低点 254.43)。
3. 影响范围: 仅调整一个配置常量, 无需修改核心逻辑或测试用例结构。

关键文件:

- `test/registered/models_e2e/test_dsa_glm52_nvfp4_tp_mtp.py` (模块 测试; 类别 test; 类型 test-coverage): 唯一变更文件, 将速度阈值从 300 下调至 250, 以适配近期性能和 CI 环境变化。

关键符号: 未识别

关键源码片段

`test/registered/models_e2e/test_dsa_glm52_nvfp4_tp_mtp.py`

唯一变更文件, 将速度阈值从 300 下调至 250, 以适配近期性能和 CI 环境变化。

```
class TestGLM52NVFP4TPMTP(  
    DsaMtpServerBase, DsaMtpEvalConfigDefaults, GSM8KMixin, SpecDecodingMixin  
):
```

```
model = "nvidia/GLM-5.2-NVFP4"
tp_size = 4
mem_fraction_static = 0.8
# 下调速度阈值：由于 PR#31001 降低了加速 token 长度，
# 且 CI 容器环境引入额外开销（如 SGLANG_ENABLE_ASYNC_ASSERT），
# 实际 tps 约 270，最低 254.43，因此将阈值从 300 降至 250
bs_1_speed_thres = 250
extra_server_args = [
    "--moe-runner-backend",
    "flashinfer_trtllm",
    "--quantization",
    "modelopt_fp4",
]
```

评论区精华

关键讨论点：

- kpham-sgl 指出 PR#31001 导致加速 token 长度降低（4.284 vs 4.599），是 tps 下降的原因之一；并提及待 flashinfer v0.5.15 发布后可考虑回退 #31001。
- mmangkad 补充强调 SGLANG_ENABLE_ASYNC_ASSERT 和 SGLANG_IS_IN_CI 等 debug 环境变量在 decode 热路径中增加额外 GPU 检查，是容器环境下 tps 偏低的重要因素。
- 结论：250 tps 是合理的折中阈值，已获得批准。
- 阈值调整依据 (other)：审核者同意 250 tps 合理。

风险与影响

- 风险：风险极低。仅下调测试阈值，不涉及任何生产代码变更。可能的轻微风险：若未来性能进一步退化至 250 以下，新阈值仍可能无法完全避免误报，但相比原 300 阈值已有明显松弛。
- 影响：影响范围小，仅影响 CI 中 GLM-5.2 NVFP4 MTP 测试的通过条件。不会影响用户行为、系统性能或生产稳定性。但该测试作为 B200 平台上 NVFP4 量化模型的 CI 护网，阈值调整体现了对容器环境性能差异的适应。
- 风险标记：测试阈值调整，仅测试文件变更

关联脉络

- PR #31001（待查证，推测为影响 perf 的合入）：PR#31001 的合入导致加速 token 长度降低（4.5→4.2），是本次阈值下调的直接原因之一；评论提到该 PR 应在 flashinfer v0.5.15 后回退。