

PR #33609 完整报告

sgl-project/sglang

[Tiny] Little enhancement of Kimi-K3 test

合并时间: 2026-08-05 06:36

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33609>

执行摘要

- 一句话: Kimi-K3 B300 测试小增强: 缩短 CI 预估时长并更新配方参数
- 推荐动作: 本 PR 为纯测试配置调整, 不值得精读。若关注 Kimi-K3 B300 配方演进, 建议将本 PR 与 #33612 一起查看, 可观察到“文档配方 — CI 测试 — 底层实现”三者的同步演化模式; 另外 `est_time` 随实测收敛的做法也值得在 CI 维护中借鉴。

功能与动机

PR body 未填写具体动机 (仅保留模板), 需从变更与关联 PR 推断:

1) #33612 已将文档配方中的 `--disable-custom-all-reduce` 移除, 说明 B300 上 `custom all-reduce` 已可用, 测试需要同步保持一致; 2) 显式加入 `--enable-linear-replayssm-spec` 以避免依赖隐式默认; 3) 将 `est_time` 从 1800 秒下调至 900 秒以优化 CI 调度。Issue 评论中作者两次触发 `/rerun-test`, 第一次失败、第二次通过, CI 最终转绿。

实现拆解

- 第一步: 调整 CI 注册 (`test/registered/models_e2e/test_kimi_k3_b300.py` 第 19 行)。将 `register_cuda_ci` 的 `est_time` 从 1800 秒改为 900 秒, `stage` 与 `runner_config` 不变。这是纯调度元数据修改, 使 CI 对 Kimi-K3 B300 测试的时长预估与实际更接近。
- 第二步: Low Latency 配方显式开启 linear ReplaySSM 投机。在 `--speculative-dspark-block-size 7` 之后追加 `--enable-linear-replayssm-spec`, 使 DSPARK 投机路径与 ReplaySSM 线性模式显式配对, 避免依赖未来的隐式默认值。
- 第三步: Balanced 配方移除 `--disable-custom-all-reduce`。与文档配方 PR#33612 对齐, 说明该禁用参数已不再需要, `custom all-reduce` 路径在 B300 上已可用。
- 第四步: 验证。通过 Issue 命令 `/rerun-test` 重跑两次, 第二次通过; 未新增测试用例与断言, 无源码、配置或部署配套改动。

关键文件:

- `test/registered/models_e2e/test_kimi_k3_b300.py` (模块 端到端测试; 类别 `test`; 类型 `test-coverage`; 符号 `TestKimiK3B300LowLatency`, `TestKimiK3B300Balanced`): 唯一的变更文件, 集中体现了本 PR 的全部三处调整: CI 预估时长减半、显式开启 linear ReplaySSM 投机、移除 `custom all-reduce` 禁用参数。

关键符号: TestKimiK3B300LowLatency.setUpClass,
TestKimiK3B300Balanced.setUpClass

关键源码片段

test/registered/models_e2e/test_kimi_k3_b300.py

唯一的变更文件, 集中体现了本 PR 的全部三处调整: CI 预估时长减半、显式开启 linear ReplaySSM 投机、移除 custom all-reduce 禁用参数。

```
# B300 Kimi-K3 per-commit CI 覆盖:
# Low Latency DSPARK 配方 + Balanced DCP/HiCache 配方
# 预估时长从 1800 秒下调到 900 秒 (base-c 阶段, 8-gpu-b300 runner)
register_cuda_ci(est_time=900, stage="base-c", runner_config="8-gpu-b300")
```

```
class TestKimiK3B300LowLatency(GSM8KMixin, CustomTestCase):
    """TP8 Low Latency recipe with DSPARK linear ReplaySSM speculation."""

    gsm8k_score_threshold = 0.95
    gsm8k_num_examples = 200

    @classmethod
    def setUpClass(cls):
        cls.model = MODEL_PATH
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=SERVER_LAUNCH_TIMEOUT,
            other_args=[
                "--trust-remote-code",
                "--tp-size", "8",
                "--mem-fraction-static", "0.85",
                "--weight-loader-prefetch-checkpoints",
                "--reasoning-parser", "kimi_k3",
                "--tool-call-parser", "kimi_k3",
                "--mamba-full-memory-ratio", "0.86",
                "--speculative-algorithm", "DSPARK",
                "--speculative-draft-model-path", DSPARK_DRAFT_MODEL,
                "--speculative-dspark-block-size", "7",
                # 显式启用 linear ReplaySSM 投机, 避免依赖隐式默认行为
                "--enable-linear-replayssm-spec",
            ],
        )
```

```
class TestKimiK3B300Balanced(GSM8KMixin, CustomTestCase):
    """TP8/DCP8 Balanced recipe with hierarchical cache."""

    gsm8k_score_threshold = 0.95
    gsm8k_num_examples = 200
```

```

@classmethod
def setUpClass(cls):
    cls.model = MODEL_PATH
    cls.base_url = DEFAULT_URL_FOR_TEST
    cls.process = popen_launch_server(
        cls.model,
        cls.base_url,
        timeout=SERVER_LAUNCH_TIMEOUT,
        other_args=[
            "--trust-remote-code",
            "--tp-size", "8",
            "--dcp-size", "8",
            # 已移除 "--disable-custom-all-reduce":
            # B300 上 custom all-reduce 已稳定，与文档配方（PR#33612）保持一致
            "--mem-fraction-static", "0.85",
            "--weight-loader-prefetch-checkpoints",
            "--reasoning-parser", "kimi_k3",
            "--tool-call-parser", "kimi_k3",
            "--mamba-full-memory-ratio", "7.21",
            "--enable-hierarchical-cache",
        ],
    )

```

评论区精华

本 PR 无 review 评论（0 条 review comments），仅在 Issue 中有 CI 重跑交互：

Fridge003: /rerun-test test/registered/models_e2e/test_kimi_k3_b300.py

github-actions[bot]: 第一次运行 , 重跑后 

两次重跑触发 2 次，最终绿。失败根因未在评论区说明，推测为 B300 runner 偶发环境问题或参数调整后的瞬时波动。

- CI 重跑：Kimi-K3 B300 测试首次失败后重跑通过 (test)：重跑后通过，PR 合入；失败被认为是偶发环境问题或参数调整后的瞬时波动。

风险与影响

- 风险：
 - CI 超时风险：est_time 从 1800 秒降至 900 秒，若真实执行时间因新参数超过 900 秒，可能引发调度异常或排队偏差（取决于 CI 对该字段的语义，若仅作调度提示则风险较低）。
 - 回归风险：移除 --disable-custom-all-reduce 后，Balanced 配方会启用 custom all-reduce 路径；若 B300 上该实现仍有边界缺陷，可能导致偶发失败（本 PR 首轮 rerun 失败或与此相关，但无法证实）。
 - 投机行为变化：显式启用 --enable-linear-replayssm-spec 可能改变投机解码路径，从而影响 GSM8K 分数（阈值 0.95）。

- 影响面：仅限 base-c 阶段的 8-gpu-b300 测试，不触及生产源码。
- 影响：
 - 用户：无直接影响。
 - CI 系统：base-c 套件中 Kimi-K3 B300 测试的预估时长减半，调度更紧凑；测试配方与推荐配置 #33612 对齐，避免 CI 与文档不一致。
 - 团队：Kimi-K3 B300 回归反馈更快，但需留意 `est_time=900` 是否贴近真实运行时长，以及新参数下测试的稳定性。
 - 风险标记：`est_time` 减半可能引发超时，`custom all-reduce` 路径回归风险，首轮 CI 失败根因未定位

关联脉络

- PR #33612 Remove custom all-reduce disable from Kimi-K3 B300 recipe: 本 PR 在测试中同步移除了 `--disable-custom-all-reduce`，与 #33612 中删除文档配方该参数一一对应，共同构成 Kimi-K3 B300 配方一致性演进。
- PR #32541 [Kimi] Support kimi-k3: Kimi-K3 的 Day-0 支持 PR，本测试文件即为其配套回归测试，本次调整属于该模型支持链的后续小修阶段。