PR #34089 完整报告

sgl-project/sglang

[CI] Add Kimi-K3 low-latency performance check

合并时间：2026-08-09 04:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34089

执行摘要

- 一句话：Kimi-K3 B300 CI 新增低延迟性能阈值检查
- 推荐动作：值得关注：这是性能保障模式（SpecDecodingMixin + 类属性阈值）在 Kimi-K3 低延迟配方上的落地，可了解 sglang.test.kits.spec_decoding_kit 的断言机制。设计决策是用声明式类属性配置性能门槛，简单可扩展。对维护 K3 性能的工程师有参考价值，但对一般读者只需了解其把关作用即可。

功能与动机

该测试原本只用 GSM8K 准确率守护 Kimi-K3 服务配方（低延迟 DSPARK 与 Balanced DCP/HiCache）的模型质量，无法发现"准确率不变但低延迟路径变慢"的性能回退。本 PR 在文档字符串中明确了新增目标：Low Latency recipe 还需保证单请求 decode 性能。PR body 未提供详细背景，以上动机主要依据代码变更推断：引入 SpecDecodingMixin 并声明阈值，是为了让 CI 能在每提交阶段捕获 DSPARK 配置下的解码速度与推测接受长度下降。

实现拆解

1. 引入性能断言能力：在文件头新增 from sglang.test.kits.spec_decoding_kit import SpecDecodingMixin，并将 TestKimiK3B300LowLatency 的继承列表改为 GSM8KMixin, SpecDecodingMixin, CustomTestCase。SpecDecodingMixin 提供与推测解码相关的性能断言工具，是复用既有测试 kit 的标准做法。
2. 声明性能阈值：在类属性中新增 accept_length_thres = 6.6 与 bs_1_speed_thres = 440。从命名推断，前者用于断言平均接受长度不低于 6.6（反映推测解码有效性），后者用于断言 batch size 1 时解码速度不低于 440 token/s（反映低延迟路径吞吐）。阈值在第二个 commit「Set Kimi-K3 performance thresholds」中确定，说明先跑通检查再定标。
3. 同步更新模块 docstring：在文件顶部说明中补充"Low Latency recipe 还必须保持单请求 decode 性能"，使测试意图与代码一致。
4. 测试与 CI 配套：该文件通过 register_cuda_ci(est_time=900, stage="base-c", runner_config="8-gpu-b300") 注册到 B300 8 卡 CI，本 PR 未改动任何 workflow、配置或生产代码。PR 上作者两次发起 /rerun-test（test/registered/models_e2e/test_kimi_k3_b300.py），均在 8-gpu-b300 上通过。

关键文件：

- test/registered/models_e2e/test_kimi_k3_b300.py（模块 性能巡检；类别 test；类型 test-coverage；符号 TestKimiK3B300LowLatency）：唯一变更文件，为 Kimi-K3 B300 低延迟 DSPARK 测试引入 SpecDecodingMixin 并声明性能阈值，是性能回归检查的核心落地位置。

关键符号：TestKimiK3B300LowLatency

关键源码片段

test/registered/models_e2e/test_kimi_k3_b300.py

唯一变更文件，为 Kimi-K3 B300 低延迟 DSPARK 测试引入 SpecDecodingMixin 并声明性能阈值，是性能回归检查的核心落地位置。

```
# 文件： test/registered/models_e2e/test_kimi_k3_b300.py（Kimi-K3 B300 每提交 CI）
# 本 PR 为该类混入 SpecDecodingMixin，并声明性能阈值，用于低延迟 DSPARK
配置的性能回归把关
class TestKimiK3B300LowLatency(GSM8KMixin, SpecDecodingMixin, CustomTestCase):
    """TP8 Low Latency recipe with DSPARK linear ReplaySSM speculation."""

    # 质量门槛：GSM8K 200 例准确率不低于 0.95，沿用原有精度守护
    gsm8k_score_threshold = 0.95
    gsm8k_num_examples = 200

    # 性能门槛（本次新增）：
    # accept_length_thres：DSPARK 平均接受长度应 ≥ 6.6，用于捕获推测解码质量的回退
    # bs_1_speed_thres：batch size 1 单请求解码速度应 ≥ 440 token/
    s，用于捕获低延迟路径性能回退
    accept_length_thres = 6.6
    bs_1_speed_thres = 440

    @classmethod
    def setUpClass(cls):
        # 使用 B300 节点固定快照权重，TP=8 + DSPARK + 线性 ReplaySSM 推测解码
        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",
            "--enable-linear-replayssm-spec",
        ],
    )
```

评论区精华

本 PR 没有实质性的 review 讨论：唯一审核记录是 Fridge003 的 APPROVED，评论为空。值得留意的是作者在 PR 上两次通过 /rerun-test 让 CI 重跑该测试（8-gpu-b300 上均通过），说明在合入前已确认性能断言在目标环境上稳定成立，没有出现阈值过紧导致的偶发失败。

> mmangkad：/rerun-test test/registered/models_e2e/test_kimi_k3_b300.py
> github-actions[bot]：8-gpu-b300 运行通过 ✅

- K3 B300 CI 运行稳定性验证 (other): 性能阈值在目标 CI 环境上验证通过，未发现 flaky 迹象。

风险与影响

- 风险：风险较低但存在偶发失败隐患：bs_1_speed_thres = 440 与 accept_length_thres = 6.6 是在当前 B300 环境的固定模型快照上标定的阈值；若 CI 节点 GPU 时钟波动、NVLink 状态异常或模型权重快照更新，可能出现误报失败。仅覆盖单一配置：该检查只针对 TP8 + DSPARK + ReplaySSM 的 B300 低延迟配方，对 DCP/HiCache 配方、其他 GPU 规格或非推测解码路径无保护。与 K3 相关改动联动：近期 PR#33764 修复 K3 router GEMM 精度、PR#33981 新增 K3 DSpark MLA verify 内核等改动可能影响该测试的指标表现，阈值需随性能预期调整。
- 影响：影响范围：仅 CI 测试用例，对生产服务和用户无直接影响。对团队的价值是建立持续性能回归防线，防止低延迟路径静默变慢；对维护 Kimi-K3 的工程师而言，后续涉及 DSPARK、MLA 或 Mamba 相关改动时，本测试会自动提示性能波动。影响程度较小但具有持续防护价值。
- 风险标记：性能阈值依赖环境稳定，仅覆盖 B300 单配置，无生产代码变更

关联脉络

- PR #33764 Fix the router GEMM inaccuracy when using _front_w in Kimi-K3: 同模型 Kimi-K3 的精度修复，涉及 router GEMM 与专家选择，可能影响 DSPARK 低延迟配置的接受长度与速度指标；本 PR 为其提供性能回归覆盖。
- PR #33981 [AMD] Add K3 verified mla kernel for DSpark on triton backend: 与 K3 DSPARK 推测解码相关的 MLA verify 内核改动，是本 PR 性能阈值所守护的核心路径之一。