PR #34405 完整报告

sgl-project/sglang

Fix flaky decode cache-hit check in Inkling test

合并时间：2026-08-11 21:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34405

执行摘要

- 一句话：修复 Inkling 测试中 decode cache-hit 检查的 flaky 问题
- 推荐动作：该 PR 值得精读，尤其适合负责测试维护或关注 mamba/SWA 缓存机制的工程师。它展示了一个由配置不对齐导致的 flaky 测试的典型排查过程，以及如何通过将对齐直觉转化为显式断言（0.99）来提升测试的信噪比。但本身不涉及核心推理逻辑，重要性中等。

功能与动机

PR body 指出：test_input_output_logprobs_match_decode_cache_hit 间歇性红色，报 Too few decode cache hits: 16/32，而 avg_kl_div 保持 0.0。命中数并非噪声——decode checkpoint 仅在滑动窗口完整存留时可用，而窗口外的 SWA 槽位按页释放，因此最深的可复用位置是页边界；默认 mamba_track_interval=256 配合 --page-size 128 导致只有一半 prompt 能命中，期望值恰好 16/32，而测试需要 17 个，所以从落地起就一直差一个。

实现拆解

1. 根因定位：在 test/registered/models_e2e/test_inkling_small_nvfp4.py 中，服务启动参数使用 --page-size 128，但 --mamba-track-interval 未显式设置，沿用默认 256。decode checkpoint 的可复用性取决于页边界，间隔与页大小不对齐导致只有约一半序列长度满足条件。
2. 修改测试配置：新增常量 KL_TRACK_INTERVAL = 128，并在 setUpClass 的服务启动参数中加入 --mamba-track-interval str(KL_TRACK_INTERVAL)，使 checkpoint 间隔与页大小一致，让每个页边界都成为 checkpoint，从而所有 prompt 都能命中 decode 缓存。
3. 泛化 helper：在 python/sglang/test/kl_test_utils.py 的 test_input_output_logprobs_match_decode_cache_hit_helper 中新增 min_cache_hit_ratio=0.5 参数，替换原先硬编码的 0.5，保持默认行为不变，同时允许调用方按需收紧断言。
4. 收紧断言：在 test_input_output_logprobs_match_decode_cache_hit 中传入 min_cache_hit_ratio=0.99，将命中率从“过半数”提升到“几乎全部”，使单次 miss 即视为状态复用回归，测试覆盖范围从原来仅一半请求翻倍。
5. 配套说明：调整 _run 方法支持透传 **kwargs，并更新注释解释对齐逻辑，避免未来维护者误改配置。

关键文件：

- test/registered/models_e2e/test_inkling_small_nvfp4.py（模块 端到端测试；类别 test；类型 test-coverage；符号 _run）：修复 flaky 的主战场：新增 KL_TRACK_INTERVAL 常量、在服务启动参数中显式设置 --mamba-track-interval 128，并将 decode cache-hit 测试的命中率阈值提升到 0.99，从根本上消除间歇性失败。
- python/sglang/test/kl_test_utils.py（模块 测试工具；类别 test；类型 test-coverage）：为 decode cache-hit helper 引入 min_cache_hit_ratio 参数，替代硬编码的 0.5，使调用方可以按需收紧断言阈值，是支撑本 PR 收紧测试的关键配套改动。

关键符号：_run, test_input_output_logprobs_match_decode_cache_hit_helper

关键源码片段

test/registered/models_e2e/test_inkling_small_nvfp4.py

修复 flaky 的主战场：新增 KL_TRACK_INTERVAL 常量、在服务启动参数中显式设置 --mamba-track-interval 128，并将 decode cache-hit 测试的命中率阈值提升到 0.99，从根本上消除间歇性失败。

```
# 与下方 --page-size 128 对齐：SWA 窗口外的槽位按页释放，
# 只有落在页边界上的 checkpoint 才能保留完整滑动窗口数据，
# 默认 256 间隔会使一半的序列长度错过边界、丢掉 decode 前缀。
KL_TRACK_INTERVAL = 128

class TestInklingSmallNvfp4(CustomTestCase):
    @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=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--tp", "4",
                "--trust-remote-code",
                "--quantization", "modelopt_fp4",
                "--attention-backend", "fa4",
                "--page-size", "128",
                "--fp4-gemm-backend", "flashinfer_trtllm",
                "--moe-runner-backend", "flashinfer_trtllm_routed",
                "--mamba-radix-cache-strategy", "extra_buffer",
                "--swa-full-tokens-ratio", "0.1",
                "--mamba-full-memory-ratio", "0.1",
                "--mem-fraction-static", "0.85",
                # 关键修复：checkpoint 间隔与页大小对齐，确保每个页边界都有 checkpoint
                "--mamba-track-interval", str(KL_TRACK_INTERVAL),
                "--enable-deterministic-inference",
            ],
            env={**os.environ, "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1"},
```

```
        )

    def _run(self, helper, **kwargs):
        helper(
            self.base_url,
            {self.model: {"kl_div": KL_DIV_THRESHOLD}},
            self.model,
            max_samples=32,
            max_new_tokens=KL_MAX_NEW_TOKENS,
            trust_remote_code=True,
            **kwargs,
        )

    def test_input_output_logprobs_match_decode_cache_hit(self):
        # 0.99 意味着要求几乎全部 prompt 命中：间隔与页对齐后复用成为无条件行为，
        # 因此任何一次 miss 都代表状态复用回退，而不再是几何巧合。
        self._run(assert_logprobs_match_decode_cache_hit, min_cache_hit_ratio=0.99)
```

python/sglang/test/kl_test_utils.py

为 decode cache-hit helper 引入 min_cache_hit_ratio 参数，替代硬编码的 0.5，使调用方可以按需收紧断言阈值，是支撑本 PR 收紧测试的关键配套改动。

```
def test_input_output_logprobs_match_decode_cache_hit_helper(
    base_url,
    ACC_THRESHOLDS,
    model_name,
    max_samples=None,
    max_new_tokens=8192,
    trust_remote_code=False,
    min_cache_hit_ratio=0.5,
):
    # ... 第一轮生成与第二轮生成逻辑 ...
    for i, result in enumerate(results):
        if result["meta_info"]["cached_tokens"] <= len(first_turn_input_ids[i]) + 1:
            print(f"Decode cache miss for prompt {i}, skipping")
            continue
        new_input_ids.append(second_turn_input_ids[i] + result["output_ids"])
        output_logprobs.append(_extract_output_logprobs(result))

    if not os.environ.get("SGLANG_TEST_SKIP_CACHE_HIT_ASSERT"):
        # 页对齐的 SWA 保留策略决定了哪些 prompt 能命中，因此默认阈值只用于排除“全
        miss”的空跑；
        # 调用方若通过配置让每个 prompt 都命中，可提高该阈值来获得更严格的回归检测。
        assert len(new_input_ids) > min_cache_hit_ratio * len(
            second_turn_input_ids
        ), f"Too few decode cache hits: {len(new_input_ids)}/{len(second_turn_input_ids)}"
```

评论区精华

本 PR 没有 review 评论，仅有一条 issue 评论由作者触发 /rerun-test，第二次 rerun 通过（4-gpu-b200 上的测试从失败转为成功）。没有公开的设计讨论或争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险主要在于测试行为收紧：min_cache_hit_ratio=0.99 会把任何一次非预期的 cache miss 变成失败，若未来 mamba 或 SWA 的内存管理策略发生变化（如页大小调整、释放粒度变化），可能导致误报。另外，--mamba-track-interval 128 是测试专用配置，不会影响生产路径；helper 的参数默认值保持向后兼容，其他调用方不受影响。整体风险较低，但需要注意该测试对几何对齐的敏感性。
- 影响：影响范围限于 CI 测试：修复了 TestInklingSmallNvfp4 在 B200 4-GPU 上的间歇性红色，提升了测试稳定性；同时通过提高命中率要求，使该测试能更可靠地捕获 decode 阶段的状态复用回归。kl_test_utils.py 的改动对所有使用该 helper 的测试透明，默认行为不变，并提供了更灵活的阈值控制。对用户和运行时无影响。
- 风险标记：测试阈值收紧，可能暴露新的缓存回退问题，依赖 mamba-track-interval 与 page-size 的隐式耦合

关联脉络

- PR #34301 Fix CI server warmup progress logging: 同为 CI 测试稳定性修复，针对多模态生成测试中的 flaky 日志 / 断言问题，说明团队持续治理测试稳定性，与本 PR 目标一致。