PR #27070 完整报告

sgl-project/sglang

Relax mamba unified cache kl threshold

合并时间：2026-06-02 23:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27070

执行摘要

- 一句话：放宽 Mamba 缓存 KL 阈值 0.003 → 0.005
- 推荐动作：该 PR 是典型的测试阈值微调，无架构或逻辑变更，仅需了解。对于关注 CI 测试稳定性的团队值得注意。

功能与动机

PR body 指出 TestUnifiedMambaRadixCache.test_multiturn_prefill_cache_hit_branching 测试因 FP8 运行噪声而出现 flaky 失败。作者统计了超过 100 次固定输入运行，prefill 缓存命中 KL 超过 0.003 阈值 6 次，但从未超过 0.005，因此将阈值提升至 0.005 以匹配同一文件中的 HiCache 变体。

实现拆解

1. 修改测试配置常量：在 test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py 中，将 TestUnifiedMambaRadixCache 类的 kl_threshold 属性值从 0.003 改为 0.005。
2. 添加注释说明：在原值位置增加注释，解释放宽原因——FP8 缓存命中 KL 噪声导致原阈值 0.003 出现 flaky 失败，且新值与 HiCache 变体一致。

关键文件：

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，将 KL 阈值从 0.003 提高至 0.005 以消除 FP8 精度噪声导致的 flaky 失败。

关键符号：未识别

关键源码片段

test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py

唯一变更文件，将 KL 阈值从 0.003 提高至 0.005 以消除 FP8 精度噪声导致的 flaky 失败。

```
class TestUnifiedMambaRadixCache(UnifiedRadixTreeTestMixin, CustomTestCase):
    """Mamba hybrid + UnifiedRadixCache."""

    # fp8 cache-hit KL is noisy and flaked at 0.003, match the HiCache variant
```

```
    kl_threshold = 0.005 # 原值 0.003，根据 100+ 次运行统计未超过 0.005 而放宽
    prefill_cache_assert = staticmethod(
        make_mamba_prefill_assert(chunk_size=MAMBA_CHUNK_SIZE)
    )
    decode_cache_assert = staticmethod(
        make_mamba_decode_assert(track_interval=MAMBA_TRACK_INTERVAL)
    )

    @classmethod
    def setUpClass(cls):
        cls.model = MAMBA_MODEL
        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-size",
                "4",
                "--chunked-prefill-size",
                "2048",
                "--mem-fraction-static",
                "0.85",
                "--mamba-scheduler-strategy",
                "extra_buffer",
                "--mamba-track-interval",
                str(MAMBA_TRACK_INTERVAL),
            ],
            env={"SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1"},
        )
        cls.input_ids = get_input_ids(cls.model, num_samples=18)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。阈值放宽幅度很小（0.003→0.005），且基于统计分析（100+ 次运行未超过 0.005），不会遗漏真实回归问题。但若未来 FP8 精度有重大变化，需重新评估阈值是否仍适用。
- 影响：仅影响单个测试文件中的 KL 断言阈值，CI 测试稳定性提升，减少假阳性失败。不影响任何生产代码或模型行为。
- 风险标记：仅测试变更

关联脉络

- PR #27064 Fix stale import after kl_nightly rename: 同一个测试文件 test_unified_radix_cache_kl_mamba.py 的同类调整，均涉及 Mamba unified radix cache 的 KL 测试。
- PR #26927 [UnifiedTree]: Add HiCache Nightly CI For GLM5: PR body 提到新阈值与 HiCache 变体匹配，HiCache nightly CI 包含类似阈值设置。