

PR #27483 完整报告

sgl-project/sglang

[UnifiedTree]: Reduce HiCache PP KL test concurrency to avoid decode OOM

合并时间: 2026-06-07 17:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27483>

执行摘要

- 一句话: 降低 HiCache PP KL 测试并发避免 OOM
- 推荐动作: 简单有效的修复, 值得合并。建议后续考虑更可靠的 OOM 处理机制或动态并发控制。

功能与动机

PR body 指出修复 <https://github.com/sgl-project/sglang/pull/27285#issuecomment-4641329128> 中测试发生的 OOM 问题。在 HiCache + PP 场景下, 并发 4 请求时 decode 阶段显存不足。

实现拆解

1. 定位问题: 在 HiCache + Pipeline Parallelism 的 KL 蒸馏测试中, `max_running_requests=4` 导致 decode 阶段显存溢出 (OOM)。
2. 调整并发数: 将 `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_hicache_pp_kl.py` 中 `max_running_requests` 从 4 改为 2。
3. 验证: 通过 `/rerun-test` 触发 CI, 单测通过 (4-gpu-h100)。

关键文件:

- `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_hicache_pp_kl.py` (模块 测试配置; 类别 `test`; 类型 `configuration`): 测试配置文件, 将 `max_running_requests` 从 4 改为 2, 解决 HiCache PP decode OOM 问题。

关键符号: 未识别

关键源码片段

`test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_hicache_pp_kl.py`

测试配置文件, 将 `max_running_requests` 从 4 改为 2, 解决 HiCache PP decode OOM 问题。

```
# 文件 : test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_hicache_pp_kl.py
```

```
class TestUnifiedQwen3HiCachePP(UnifiedRadixTreeTestMixin, CustomTestCase):
```

```
    """
```

```
    Qwen3-30B-A3B-FP8 + HiCache + PP + UnifiedRadixCache.
```

```
    并发数从 4 降至 2，避免 decode 阶段因显存不足 OOM。
```

```
    """
```

```
    hicache_io_backend = "direct"
```

```
    hicache_mem_layout = "page_first_direct"
```

```
    max_running_requests = 2 # 先前为 4，调整为 2 以减少并发显存压力
```

```
    kl_threshold = 0.012
```

```
    gsm8k_threshold = 0.7
```

```
    num_gsm8k_questions = 50
```

```
    mmlu_threshold = 0.7
```

```
    decode_cache_assert = staticmethod(_assert_pp_decode_cached_tokens)
```

```
    # ... 其余方法保持不变
```

评论区精华

无 review 讨论。仅自动化 bot 回复 rerun 结果。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试参数，不涉及任何生产代码。降低并发后可能减弱测试强度，但 GSM8K/MMLU 阈值未变，准确性验证依然足够。
- 影响：仅影响 HiCache PP KL 测试用例。并发从 4 降至 2，减少显存压力，避免 OOM 导致测试不稳定。对其他模块无影响。
- 风险标记：降低测试并发，可能降低覆盖强度

关联脉络

- PR #27285 [HiCache] PP KL test fails with OOM: 该 PR 修复了 #27285 中 OOM 问题的关联 issue comment。