PR #26666 完整报告

sgl-project/sglang

[UnifiedTree]: Split unified tree kl ci into multiple files to reduce GPU usage.

合并时间：2026-05-29 17:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26666

执行摘要

- 一句话：拆分 UnifiedRadixTree KL CI 测试文件以减少 GPU 占用量。
- 推荐动作：适合 CI 维护者和测试架构学习者精读，了解如何通过测试文件拆分和资源配置优化 CI 成本。设计决策值得注意：每个测试文件按模型类型和资源需求独立配置，而非统一使用最高配置。

功能与动机

原测试文件（如 test_unified_radix_cache_kl_mamba.py 和 test_unified_radix_cache_kl_hicache.py）包含了多种模型类型的测试，导致单次 CI 运行需要较高端 GPU（如 8*H200）。拆分后，每种模型类型独立运行，可以选用更经济的 GPU 配置，从而减少 CI 资源占用和排队时间。

实现拆解

1. 整合 Mamba 相关测试：新建 test_unified_radix_cache_kl_mamba.py 到 unified_radix_tree/ 目录，将原 test_unified_radix_cache_kl_mamba.py 中的 TestUnifiedMambaRadixCache、原 test_unified_radix_cache_kl_hicache.py 中的 TestUnifiedMambaHiCache 以及原 test_unified_radix_cache_kl_hicache_part2.py 中的 TestUnifiedMambaHiCacheL3 合并，统一使用 runner_config "4-gpu-h100" 替代之前的 "8-gpu-h200"。
2. DSV4 测试精简：原 test_unified_radix_cache_kl_hicache.py 同时包含 Mamba 和 DSV4 测试。本 PR 移除其中的 Mamba 部分，只保留 TestUnifiedDeepSeekV4FlashHiCache，并重命名为 test_unified_radix_cache_kl_dsv4.py，放入 unified_radix_tree/，GPU 配置从 "8-gpu-h200" 降为 "4-gpu-h100"。
3. Full Attention 和 SWA 测试降配：将 test_unified_radix_cache_kl_full.py 和 test_unified_radix_cache_kl_swa.py 移至 unified_radix_tree/ 子目录，TP 大小从 4 改为 2，runner_config 从 "4-gpu-h100" 改为 "2-gpu-large"，以匹配模型需求并降低 GPU 占用。
4. 删除旧文件：删除合并后不再需要的 test_unified_radix_cache_kl_mamba.py 和 test_unified_radix_cache_kl_hicache_part2.py。test_unified_radix_cache_kl_hicache_nightly.py 无内容变更仅重命名迁移。

关键文件：

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py（模块 Mamba 测试；类别 test；类型 test-coverage；符号 TestUnifiedMambaRadixCache, setUpClass, tearDownClass, TestUnifiedMambaHiCache）：新增文件，整合了所有 Mamba 相关测试（纯 Radix、HiCache L2/L3），是本次拆分的核心合并产物。
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_dsv4.py（模块 DSV4 测试；类别 test；类型 rename-or-move；符号 TestUnifiedMambaHiCache, setUpClass, tearDownClass）：从原混合文件拆分出的 DSV4 测试，移除了 Mamba 部分并调整 GPU 配置。
- test/registered/radix_cache/test_unified_radix_cache_kl_mamba.py（模块 测试删除；类别 test；类型 deletion；符号 TestUnifiedMambaRadixCache, setUpClass, tearDownClass）：被删除的旧 Mamba 测试文件，内容已整合到新文件中。
- test/registered/radix_cache/test_unified_radix_cache_kl_hicache_part2.py（模块 测试删除；类别 test；类型 deletion；符号 TestUnifiedMambaHiCacheL3, setUpClass, tearDownClass）：被删除的 HiCache 第二部分测试，内容已整合到新 Mamba 文件中。
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_full.py（模块 Full 测试；类别 test；类型 rename-or-move）：重命名并调整配置（tp-size 4→2, runner 改为 2-gpu-large）。
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_swa.py（模块 SWA 测试；类别 test；类型 rename-or-move）：重命名并调整配置（tp-size 4→2, runner 改为 2-gpu-large）。
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hicache_nightly.py（模块 Nightly 测试；类别 test；类型 rename-or-move）：仅重命名迁移，无内容变更。

关键符号：setUpClass, tearDownClass, TestUnifiedMambaRadixCache, TestUnifiedMambaHiCache, TestUnifiedMambaHiCacheL3, TestUnifiedDeepSeekV4FlashHiCache

关键源码片段

test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py

新增文件，整合了所有 Mamba 相关测试（纯 Radix、HiCache L2/L3），是本次拆分的核心合并产物。

```
import os
import shutil
import tempfile
import unittest

# 从 nightly 混入类导入双精度测试基类
from test_unified_radix_cache_kl_hicache_nightly import AccuracyTwoPassMixin

from sglang.srt.utils import kill_process_tree
```

```
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.kits.unified_radix_cache_kit import UnifiedRadixTreeTestMixin
from sglang.test.kl_multiturn_utils import (
    get_input_ids,
    make_mamba_decode_assert,
    make_mamba_prefill_assert,
)
from sglang.test.test_utils import (
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

# 注册 CI 舞台，估计时间 768s，运行于 4-gpu-h100 机器
register_cuda_ci(est_time=768, stage="base-c", runner_config="4-gpu-h100")

# 模型与参数常量
MAMBA_MODEL = "Qwen/Qwen3-Next-80B-A3B-Instruct-FP8"
MAMBA_CHUNK_SIZE = 64
MAMBA_TRACK_INTERVAL = 128


class TestUnifiedMambaRadixCache(UnifiedRadixTreeTestMixin, CustomTestCase):
    """Mamba hybrid + UnifiedRadixCache 测试类"""
    kl_threshold = 0.003
    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):
        # 启动模型服务器，4*TP、启用 unified radix tree
        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)

    @classmethod
    def tearDownClass(cls):
        kill_process_tree(cls.process.pid)
```

test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_dsv4.py

从原混合文件拆分出的 DSV4 测试，移除了 Mamba 部分并调整 GPU 配置。

```
import unittest

from sglang.srt.utils import kill_process_tree
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.kits.unified_radix_cache_kit import UnifiedRadixTreeTestMixin
from sglang.test.kl_multiturn_utils import get_input_ids
from sglang.test.test_utils import (
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    is_in_ci,
    popen_launch_server,
)

DSV4_FLASH_MODEL = "sgl-project/DeepSeek-V4-Flash-FP8"
DSV4_FLASH_LAUNCH_TIMEOUT = 3600

# 调整 CI 配置：8-gpu-h200 -> 4-gpu-h100
register_cuda_ci(est_time=768, stage="base-c", runner_config="4-gpu-h100")


def _assert_dsv4_decode_cached_tokens(result, history_len, output_len, label):
    """断言解码缓存 token 数在合理范围内。"""
    expected = history_len + output_len
    actual = result["meta_info"]["cached_tokens"]
    lower = max(0, expected - 256)
    assert actual >= lower, f"{label}: expected cached_tokens>={lower}, got {actual}"


class TestUnifiedDeepSeekV4FlashHiCache(UnifiedRadixTreeTestMixin, CustomTestCase):
    """DeepSeek V4 Flash FP8 + HiCache + UnifiedRadixCache 测试类"""
    hicache_io_backend = "direct"
    hicache_mem_layout = "page_first_direct"
    max_running_requests = 4
    kl_threshold = 0.005
    sampling_temperature = 0
    decode_hit_request_batch_size = 3
    decode_hit_inter_batch_delay_s = 0.5
```

```
    decode_cache_assert = staticmethod(_assert_dsv4_decode_cached_tokens)
    gsm8k_threshold = 0.90
    num_gsm8k_questions = 100

    # 跳过 Multiturn 测试（CI 中不执行）
    @unittest.skipIf(is_in_ci(), "To reduce the CI execution time.")
    def test_multiturn_logprobs_match(self):
        pass

    @classmethod
    def setUpClass(cls):
        cls.model = DSV4_FLASH_MODEL
        cls.base_url = DEFAULT_URL_FOR_TEST
        # 启动服务器配置省略（与之前一致但改用 4-gpu-h100）
```

评论区精华

- gemini-code-assist[bot] 指出 test_unified_radix_cache_kl_full.py 中缺少 --tp-size 参数，可能影响 32B 模型的运行。该问题在随后提交中已修复。
- hnyls2002 批准并评论 "Good job. Once the lint is fixed, we can merge it"。
- 缺失 --tp-size 参数问题 (correctness): 后续提交中已修复，添加了 --tp-size 2。
- 审核批准 (other): PR 已合并。

风险与影响

- 风险：
  - 测试覆盖风险：拆分过程中可能遗漏某些测试用例，但通过导入确认所有测试类已迁移至新文件。
  - 资源配置风险：TP 大小和 runner 配置调整可能在某些 GPU 环境下导致 OOM 或推理超时，需确保配置与模型内存需求匹配。
  - 路径变更风险：旧文件被删除，依赖绝对路径的 CI 脚本可能需要更新。
  - 影响：主要影响 CI 流程：降低单次测试 GPU 资源需求，提升资源利用率和并行度。对最终用户无功能影响。团队 CI 等待时间有望缩短。
  - 风险标记：配置变更可能影响测试可靠性，测试覆盖完整性需验证，旧路径删除需同步 CI 脚本

关联脉络

- 暂无明显关联 PR