PR #26927 完整报告

sgl-project/sglang

[UnifiedTree]: Add HiCache Nightly CI For GLM5

合并时间：2026-06-02 19:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26927

执行摘要

- 一句话：为 GLM5 添加 UnifiedTree Nightly CI 测试
- 推荐动作：值得作为测试基础设施扩展的示例：通过环境变量切换不同实现，在 nightly 中保持对候选方案的长期验证。注意阈值调整需结合真实精度数据，避免过度放松标准。

功能与动机

为 GLM5 模型的 HiCache L3 精度测试增加 UnifiedRadixTree 覆盖，同时调整测试参数以提高 CI 运行稳定性。

实现拆解

1. 文件重命名：将 test_unified_radix_cache_kl_hicache_nightly.py 重命名为 test_unified_radix_cache_kl_nightly.py，去除 hicache 前缀以反映同时覆盖多种缓存后端。
2. 调整测试参数：在 AccuracyTwoPassMixin 中将 mmlu_threshold 从 0.75 降至 0.65，num_mmlu_examples 从 200 减至 100，以降低测试资源消耗并减少随机性导致的失败。
3. 重构旧测试类：将 TestGLM5HiCacheL3Accuracy 重命名为 TestGLM5HiRadixCacheL3Accuracy，明确其使用 HiRadixTree，启动参数不变。
4. 新增 UnifiedRadixTree 测试类：新增 TestGLM5UnifiedRadixCacheL3Accuracy，继承同一 AccuracyTwoPassMixin，启动时额外设置环境变量 SGLANG_ENABLE_UNIFIED_RADIX_TREE=1，其余参数与旧类一致。新增类添加了 tearDownClass 方法（实际与旧类共享逻辑）。

关键文件：

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py（模块 测试；类别 test；类型 rename-or-move；符号 TestGLM5HiCacheL3Accuracy, TestGLM5HiRadixCacheL3Accuracy, tearDownClass, TestGLM5UnifiedRadixCacheL3Accuracy）：唯一变更文件，包含类重命名、参数调整和新增 UnifiedRadixTree 测试类，是 nightly CI 的核心测试逻辑。

关键符号：TestGLM5HiRadixCacheL3Accuracy.setUpClass, TestGLM5UnifiedRadixCacheL3Accuracy.setUpClass, AccuracyTwoPassMixin

关键源码片段

test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py

唯一变更文件，包含类重命名、参数调整和新增 UnifiedRadixTree 测试类，是 nightly CI 的核心测试逻辑。

```
# 调整后的 MMLU 参数
class AccuracyTwoPassMixin:
    gsm8k_threshold: float = 0.90
    num_gsm8k_questions: int = 200
    gsm8k_parallel: int = 40
    # 降低了阈值和样本数，减少测试资源消耗
    mmlu_threshold: float = 0.65
    num_mmlu_examples: int = 100
    mmlu_num_threads: int = 32
    max_accuracy_diff: float = 0.02

    def _two_pass(self, name, run_fn, threshold):
        # 略……

class TestGLM5HiRadixCacheL3Accuracy(AccuracyTwoPassMixin, CustomTestCase):
    """GLM-5.1-FP8 + HiCache L3 (file backend), with HiRadixTree (旧树)."""
    @classmethod
    def setUpClass(cls):
        # 与之前完全相同，使用 HiRadixTree（默认）
        cls.model = GLM5_MODEL
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.hicache_dir = tempfile.mkdtemp(prefix="hicache_l3_")
        cls.process = popen_launch_server(
            cls.model, cls.base_url, timeout=GLM5_LAUNCH_TIMEOUT,
            other_args=[
                "--trust-remote-code", "--tp-size", "8", "--page-size", "64",
                "--mem-fraction-static", "0.85",
                '--model-loader-extra-config', '{"enable_multithread_load": true, "num_threads": 64}',
                "--enable-hierarchical-cache", "--hicache-ratio", "2",
                "--hicache-write-policy", "write_through",
                "--hicache-storage-prefetch-policy", "wait_complete",
                "--hicache-io-backend", "direct", "--hicache-mem-layout", "page_first_direct",
                "--hicache-storage-backend", "file",
            ],
            env={"SGLANG_HICACHE_FILE_BACKEND_STORAGE_DIR": cls.hicache_dir},
        )

class TestGLM5UnifiedRadixCacheL3Accuracy(AccuracyTwoPassMixin, CustomTestCase):
    """GLM-5.1-FP8 + HiCache L3 (file backend), with UnifiedRadixTree (新树)."""
    @classmethod
    def setUpClass(cls):
        cls.model = GLM5_MODEL
        cls.base_url = DEFAULT_URL_FOR_TEST
```

```
        cls.hicache_dir = tempfile.mkdtemp(prefix="hicache_l3_")
        cls.process = popen_launch_server(
            cls.model, cls.base_url, timeout=GLM5_LAUNCH_TIMEOUT,
            other_args=[
                # 与 HiRadixTree 参数完全相同
                "--trust-remote-code", "--tp-size", "8", "--page-size", "64",
                "--mem-fraction-static", "0.85",
                '--model-loader-extra-config', '{"enable_multithread_load": true, "num_threads": 64}',
                "--enable-hierarchical-cache", "--hicache-ratio", "2",
                "--hicache-write-policy", "write_through",
                "--hicache-storage-prefetch-policy", "wait_complete",
                "--hicache-io-backend", "direct", "--hicache-mem-layout", "page_first_direct",
                "--hicache-storage-backend", "file",
            ],
            env={
                "SGLANG_HICACHE_FILE_BACKEND_STORAGE_DIR": cls.hicache_dir,
                "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1", # 启用新树
            },
        )
```

评论区精华

无实质性人工 review 讨论，仅有一个 bot 代码审查摘要。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：仅涉及测试文件变更和参数调整。但 MMLU 阈值下调（0.75→0.65）和样本数减少可能降低测试的严格性，若模型实际精度退步时可能漏报。同时，新测试类依赖环境变量 SGLANG_ENABLE_UNIFIED_RADIX_TREE，若该变量在 CI 环境中未正确设置或代码本身有 bug 可能导致测试失败。
- 影响：影响范围限于 nightly CI 测试（8-gpu-h200）。现在 nightly 测试同时覆盖 HiCache L3 的两种树形结构（HiRadixTree 和 UnifiedRadixTree），每次运行额外增加一组测试（估计时间约 900s）。对用户无直接影响，但增强了内部对缓存一致性的验证。
- 风险标记：测试阈值放宽可能降低严格度，新环境变量依赖

关联脉络

- 暂无明显关联 PR