

PR #36281 完整报告

sgl-project/sglang

[Unified Cache]: add glm5.2 per commit ci

合并时间: 2026-08-26 10:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36281>

执行摘要

- 一句话: GLM5.2 缓存精度测试前移至每提交 CI, 抽取双遍精度 Mixin
- 推荐动作: 值得快速浏览, 重点学习两点: 一是如何通过 `register_cuda_ci(stage="extra-b")` 将模型级精度测试提升到 per-commit 阶段; 二是把跨模型复用的测试逻辑抽到 `python/sglang/test/kits/` 的共享 Mixin 组织方式。对于需要新增 unified cache 相关精度测试的工程师有直接参考价值; 不关心测试基建的读者可以跳过。

功能与动机

PR body 只有模板, 未给出明确动机描述, 但从变更内容可推断: UnifiedRadixTree 与 HiCache 属于相对活跃且易回归的缓存路径, GLM-5.2 的精度保障此前只在 nightly 阶段运行, 回归发现周期过长。本次将其提升为 per-commit CI (`stage="extra-b"`), 并结合共享 Mixin 降低多模型测试的维护成本, 让精度回归在每次提交时即被发现。

实现拆解

1. 抽取共享 Mixin: 在 `python/sglang/test/kits/unified_radix_cache_kit.py` 中新增 `AccuracyTwoPassMixin`, 包含 `_run_gsm8k`、`_flush_cache`、`_two_pass` 三个辅助方法, 以及 `test_gsm8k_two_passes`、`test_l3_prefetch_full_prefix_hit_after_flush` 两个测试方法; 配置文件里引入了 `requests` 依赖。
2. 删除旧 nightly 文件: 整个删除 `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py` (258 行), 其中原 `TestGLM5HiRadixCacheL3Accuracy` 测试类与 Mixin 一并移除, 避免两处定义漂移。
3. 新增 GLM-5.2 per-commit 测试: 新增 `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_glm52.py`, 通过 `register_cuda_ci(est_time=900, stage="extra-b", runner_config="8-gpu-h200")` 注册到每次提交的 extra-b 阶段; 测试类 `TestGLM5UnifiedRadixCacheL3Accuracy` 继承 `AccuracyTwoPassMixin`, 在 `setUpClass` 中用 `zai-org/GLM-5.2-FP8` 启动 TP=8 服务, 开启 HiCache L3 (file 后端) 与 EAGLE 投机解码, 并设置 `SGLANG_ENABLE_UNIFIED_RADIX_TREE=1`。
4. 修正 Mamba/DSV4 导入: `test_unified_radix_cache_kl_mamba.py` 与 `test_unified_radix_cache_kl_dsv4.py` 从直接从 nightly 文件导入 `AccuracyTwoPassMixin`, 改为从 `sglang.test.kits.unified_radix_cache_kit` 导入, 同时保留 `UnifiedRadixTreeTestMixin` 的原有引用。

5. 无独立配置 / 部署配套：CI 注册完全通过测试文件内的 `register_cuda_ci` 声明完成，没有新增 workflow 或配置文件。

关键文件：

- `python/sglang/test/kits/unified_radix_cache_kit.py` (模块 测试套件；类别 `test`；类型 `test-coverage`；符号 `AccuracyTwoPassMixin`, `_run_gsm8k`, `_flush_cache`, `_two_pass`)：核心变更文件，新增 `AccuracyTwoPassMixin` 共享 Mixin，包含双遍 GSM8K 精度对比与 L3 预取全前缀命中校验，是本次测试复用的基础。
- `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_glm52.py` (模块 GLM5 测试；类别 `test`；类型 `test-coverage`；符号 `TestGLM5UnifiedRadixCacheL3Accuracy`, `setUpClass`, `tearDownClass`)：新增的 GLM-5.2 per-commit 测试入口，将精度验证注册到 extra-b 阶段，并配置 HiCache L3 与 EAGLE 投机解码。
- `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py` (模块 缓存测试；类别 `test`；类型 `deletion`；符号 `AccuracyTwoPassMixin`, `_run_gsm8k`, `_flush_cache`, `_two_pass`)：整个 nightly 测试文件被删除 (258 行)，其中 `AccuracyTwoPassMixin` 与 GLM5 HiCache 测试类被迁移到 kit 与新测试文件，是本 PR 结构调整的核心动作。
- `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`)：导入路径从删除的 nightly 文件改为共享 kit，避免因文件删除导致的 `ImportError`。
- `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_dsv4.py` (模块 DSV4 测试；类别 `test`；类型 `test-coverage`)：同步修正 `AccuracyTwoPassMixin` 导入路径，保持 DSV4 测试与新的共享 Mixin 组织一致。

关键符号：`AccuracyTwoPassMixin`, `_run_gsm8k`, `_flush_cache`, `_two_pass`,
`test_gsm8k_two_passes`, `test_l3_prefetch_full_prefix_hit_after_flush`,
`TestGLM5UnifiedRadixCacheL3Accuracy.setUpClass`,
`TestGLM5UnifiedRadixCacheL3Accuracy.tearDownClass`

关键源码片段

`test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_glm52.py`

新增的 GLM-5.2 per-commit 测试入口，将精度验证注册到 extra-b 阶段，并配置 HiCache L3 与 EAGLE 投机解码。

```
class TestGLM5UnifiedRadixCacheL3Accuracy(AccuracyTwoPassMixin, CustomTestCase):
    """GLM-5.2-FP8 + HiCache L3 (file 后端) 配合 UnifiedRadixTree 的精度测试。"""

    @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=[
        "--trust-remote-code",
        "--tp-size", "8",
        "--page-size", "64",
        "--mem-fraction-static", "0.8",
        "--model-loader-extra-config",
        '{"enable_multithread_load": true, "num_threads": 64}',
        "--enable-hierarchical-cache",
        "--hichache-ratio", "2",
        "--hichache-write-policy", "write_through",
        "--hichache-storage-prefetch-policy", "wait_complete",
        "--hichache-io-backend", "kernel",
        "--hichache-mem-layout", "page_first",
        "--hichache-storage-backend", "file",
        "--speculative-algorithm", "EAGLE",
        "--speculative-num-steps", "3",
        "--speculative-eagle-topk", "1",
        "--speculative-num-draft-tokens", "4",
    ],
    env={
        # L1 缓存目录指向测试专属临时目录，便于测试结束清理。
        "SGLANG_HICACHE_FILE_BACKEND_STORAGE_DIR": cls.hicache_dir,
        # 显式开启 unified radix tree 路径，覆盖本次 CI 的目标组合。
        "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1",
    },
)

@classmethod
def tearDownClass(cls):
    # 结束服务进程并清掉 HiCache 临时文件，避免污染后续测试。
    kill_process_tree(cls.process.pid)
    if os.path.isdir(cls.hicache_dir):
        shutil.rmtree(cls.hicache_dir, ignore_errors=True)

```

评论区精华

该 PR 没有人类 review 评论，评论区只有作者本人发起的 CI 重跑指令与 bot 的响应。可见的运维动作包括：首次对 `test_unified_radix_cache_kl_glm52.py` 发起 `/rerun-test`，在 8-gpu-h200 上重跑通过；随后对 `radix_cache/unified_radix_tree/` 目录发起 `/rerun-group`，4-gpu-h100 上的 4 个测试全部通过。没有出现设计争议或未解决的疑虑。

- GLM-5.2 测试重跑 (other): 重跑全部通过，未触发代码修改，属于 CI 运维操作而非设计争议。

风险与影响

- 风险:

1. 删除文件可能遗留导入引用: `test_unified_radix_cache_kl_nightly.py` 被整体删除, 虽已同步修改 Mamba 与 DSV4 两个测试文件的导入, 但若仓库内还有其他测试或脚本仍从该文件导入 `AccuracyTwoPassMixin`, 将触发 `ImportError`。
2. CI 资源成本显著上升: 新测试注册在 `extra-b` 阶段, 每次提交都会在 8 卡 H200 上运行约 900 秒, 长时间占用高性能 CI 资源, 可能延长主流水线排队时间。
3. 外部依赖与稳定性风险: 测试依赖 `zai-org/GLM-5.2-FP8` 外部模型下载, 配合 EAGLE 投机解码与 HiCache L3 文件后端; 模型仓库网络问题、磁盘空间或环境差异都可能导致非功能性误标红。
4. AMD ROCm 槽位失败: PR 的 CI 状态显示 AMD ROCm 7.2 槽位为红色, 虽然本次改动未触及 AMD 路径, 但需要确认该失败是否与本 PR 引入的测试注册方式有关。
 - 影响: 对开发者而言, 每次提交都会额外触发 GLM-5.2 在 `UnifiedRadixCache + HiCache L3` 路径上的精度验证, 回归发现周期从 `nightly` 缩短到 `commit` 级别; 对系统资源而言, 8 卡 H200 的 15 分钟测试成为常驻成本; 对团队而言, `AccuracyTwoPassMixin` 进入共享测试套件后, 后续新增模型只需继承并配置启动参数即可复用双遍精度与 L3 预取校验逻辑, 降低了测试编写成本。整体影响面集中在 CI 与测试组织, 不影响运行时行为。
 - 风险标记: 删除文件可能遗留导入引用, 每提交 8 卡 H200 高成本测试, 依赖外部模型下载与 EAGLE 配置, AMD ROCm 槽位 CI 失败

关联脉络

- PR #34668 `fix(test): stabilize nightly precision regression`: 同为精度测试稳定性话题, 且都涉及 `nightly` 测试与 CI 工具链的调整, 本 PR 进一步将模型精度验证前移到 `per-commit`。
- PR #36284 [CI] `Add Kimi-K3 MMMU-Pro accuracy coverage`: 同为将模型级精度覆盖接入 CI 的变更, 展示该项目持续把精度保障前置到提交阶段的演进方向。
- PR #36232 `Refactor HiCache host pool management`: 同一 HiCache/UnifiedRadixCache 功能线, 本 PR 测试的正是该组件在 GLM-5.2 上的精度表现, 两个 PR 相互支撑。