

PR #37425 完整报告

sgl-project/sglang

HiCache: Add @rank_consensus to various functions

合并时间: 2026-09-14 10:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37425>

执行摘要

- 一句话: 给 HiCache 关键函数加 @rank_consensus 校验并在 CI 启用
- 推荐动作: 值得快速浏览, 重点学习 @rank_consensus 装饰器的参数选择策略: 它没有简单地对整个入参 / 返回值做全量比较, 而是通过表达式精确定位需要达成共识的字段, 兼顾了校验有效性与分布式对象的异构性。若后续排查 HiCache 分布式不一致问题, 可参考此模式继续扩大覆盖率。

功能与动机

PR body 未写明 Motivation 段落, 但从修改内容可推断: HiCache 在多 TP/PP 的分布式场景下, 缓存操作 (前缀匹配、插入、淘汰、存储回写、预取) 需要在所有 rank 上保持一致, 否则会导致请求处理结果分歧。@rank_consensus 是一个可开关的调试校验器, 通过在关键函数上声明 same_params 与 same_results, 可在测试环境中自动比对各 rank 的入参与返回值, 提前暴露缓存状态不一致问题, 而生产环境默认关闭不引入额外开销。

实现拆解

1. 核心源码扩展: 在 python/sglang/srt/mem_cache/unified_radix_cache.py 中为 insert、evict、evict_for_alloc、cache_finished_req、cache_unfinished_req、write_backup_storage、prefetch_from_storage、init_load_back 等函数添加 @rank_consensus 装饰器。装饰器参数通过字符串表达式精确声明需要校验的入参字段 (如 len(params.key)、req.rid) 与结果字段 (如 result.prefix_len), 避免对整对象做全量比较 (不同 rank 的 token 池编号可能不同), 只比较需要达成共识的语义字段。
2. 测试夹具改造: 在 python/sglang/test/server_fixtures/default_fixture.py 的 DefaultServerBase 中新增类属性 server_env: dict, 并在 setUpClass 里将其透传给 popen_launch_server 的 env 参数; python/sglang/test/test_utils.py 也有一行配套调整。这样各个测试类可以按需注入秩一致性检查器环境变量, 而不必修改启动逻辑。
3. 测试覆盖开启: 在 test/registered/hicache、test/registered/radix_cache/unified_radix_tree 等目录下的 20 余个测试文件中, 为 popen_launch_server 的 env 增加 SGLANG_ENABLE_RANK_CONSENSUS_CHECKER: "1", 覆盖面包括 Mamba、HiCache L2/L3、PP+TP、CP/DCP、文件存储、Mooncake 存储、EAGLE 等变体。部分测试通过 HiCacheBaseServer.server_env 类属性统一注入。
4. 性能验证: PR 声称 @rank_consensus 在禁用时无运行时开销, 并给出 DeepSeek-V4-Flash + TP4 场景下的 benchmark 结果确认无回退。

关键文件：

- python/sglang/srt/mem_cache/unified_radix_cache.py（模块 统一缓存；类别 source；类型 core-logic）：核心变更文件：为 8 个 HiCache 缓存操作函数添加 @rank_consensus 装饰器，声明跨 rank 一致性校验点。
- python/sglang/test/server_fixtures/default_fixture.py（模块 测试夹具；类别 test；类型 test-coverage）：测试基建关键改动：新增 server_env 类属性并透传到子进程 env，是全部测试开启 rank consensus 检查器的前提。
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py（模块 测试用例；类别 test；类型 test-coverage）：代表性测试文件：为 Mamba 基础、HiCache L2、L3 三类场景统一注入 rank consensus 环境变量，验证装饰器在 Mamba 混合缓存下的效果。
- test/registered/hicache/test_hicache_variants.py（模块 测试用例；类别 test；类型 test-coverage）：HiCache 变体测试统一启用校验器，覆盖标准、MLA、EAGLE、自定义页面大小四种配置。
- test/registered/hicache/test_pp_with_hicache.py（模块 测试用例；类别 test；类型 test-coverage）：PP+HiCache 集成测试：TP2+PP2 场景下开启校验器，验证流水线并行与层级缓存组合的一致性。

关键符号：insert, evict, evict_for_alloc, cache_finished_req, cache_unfinished_req, write_backup_storage, prefetch_from_storage, init_load_back, DefaultServerBase.setUpClass

关键源码片段

python/sglang/srt/mem_cache/unified_radix_cache.py

核心变更文件：为 8 个 HiCache 缓存操作函数添加 @rank_consensus 装饰器，声明跨 rank 一致性校验点。

```
# rank_consensus 装饰器用于跨 rank 一致性校验：
# 开启 SGLANG_ENABLE_RANK_CONSENSUS_CHECKER 后，会比对各 rank 上同一方法
# 的入参 (same_params) 与返回值 (same_results)，及时发现 TP/PP 间的缓存状态分歧。
# 默认关闭时装饰器透传原函数，不引入运行时开销。
```

```
class UnifiedRadixCache:
```

```
    @rank_consensus(
        same_params=["len(params.key)"],
        same_results=["result.prefix_len"],
    )
    def insert(self, params: InsertParams) -> InsertResult:
        # 只比较 key 长度而非全量 key：各 rank 的 token 池编号可能不同，
        # 但 key 长度与插入命中的 prefix_len 必须是全局一致的共识值。
        ...

    @rank_consensus(same_params=True, same_results=True)
    def evict(self, params: EvictParams) -> EvictResult:
```

```

# 全量比较入参与结果：淘汰逻辑不允许跨 rank 的差别。
return self._evict(params)

@rank_consensus(
    same_params=["req.rid", "is_insert", "kv_len_to_handle"],
)
def cache_finished_req(
    self, req: Req, is_insert: bool = True, *, kv_len_to_handle: int, **kwargs
) -> None:
    # 用请求 ID 标识跨 rank 的同一次调用，只校验关键入参字段；
    # 返回值是 None，因此不声明 same_results。
    ...

@rank_consensus(same_params=["req_id", "len(new_input_tokens)"])
def prefetch_from_storage(self, req_id: str, ...) -> None:
    # 按请求 ID 与输入长度校验存储预取的调用一致性。
    ...

```

python/sglang/test/server_fixtures/default_fixture.py

测试基建关键改动：新增 `server_env` 类属性并透传到子进程 `env`，是全部测试开启 `rank consensus` 检查器的前提。

```

class DefaultServerBase(CustomTestCase):
    model = None
    base_url = DEFAULT_URL_FOR_TEST
    timeout = DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH
    other_args: list[str] = []
    # Extra env vars passed to the launched server subprocess.
    server_env: dict = None

    @classmethod
    def setUpClass(cls):
        assert cls.model is not None, "Please set cls.model in subclass"

        # 仅当子类显式设置 server_env 时才透传，避免覆盖 popen_launch_server 的默认行为。
        kwargs = {}
        if cls.server_env:
            kwargs["env"] = cls.server_env

        # 设置 OpenAI API 环境变量 (Imm-evals 需要)。
        with openai_api_env(cls.api_key):
            cls.process = popen_launch_server(
                cls.model,
                cls.base_url,
                timeout=cls.timeout,
                other_args=cls.other_args,
                **kwargs,
            )

```

test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_mamba.py

代表性测试文件：为 Mamba 基础、HiCache L2、L3 三类场景统一注入 rank consensus 环境变量，验证装饰器在 Mamba 混合缓存下的效果。

```
class TestUnifiedMamba(UnifiedRadixTreeTestMixin, CustomTestCase):
    """纯 UnifiedRadixCache + Mamba，不带 HiCache L2/L3。"""

    @classmethod
    def setUpClass(cls):
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[...],
            env={
                "SGLANG_ENABLE_RANK_CONSENSUS_CHECKER": "1",
                "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1",
            },
        )
        # 开启 rank 一致性检查后，Mamba 缓存命中前缀在每个 rank 上必须一致。
        cls.input_ids = get_input_ids(cls.model, num_samples=18)

class TestUnifiedMambaHiCache(UnifiedRadixTreeTestMixin, CustomTestCase):
    """Mamba hybrid + HiCache L2 + UnifiedRadixCache。"""

    @classmethod
    def setUpClass(cls):
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--enable-hierarchical-cache",
                "--hcache-ratio", "4",
                # ... 其余 HiCache L2 参数
            ],
            env={
                "SGLANG_ENABLE_RANK_CONSENSUS_CHECKER": "1",
                "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1",
            },
        )
        cls.input_ids = get_input_ids(cls.model, num_samples=18)
```

评论区精华

本 PR 无代码审查评论。PR 评论区仅作者两次评论 [/rerun-failed-ci](#) 用于触发失败 CI 重跑（AMD ROCm 10 作业曾失败一次）。没有关于装饰器参数选择或测试覆盖范围的讨论记录。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 误报风险：@rank_consensus 在测试中开启后，若 same_params/same_results 选择的字段在不同 rank 间存在合法差异（例如请求对象属性顺序、非确定性字典序），可能导致 CI 误报失败。
2. 漏检风险：cache_finished_req 只校验 req.rid、is_insert、kv_len_to_handle，未覆盖 prev_prefix_len 等字段；若实际分歧发生在未声明字段上，校验器无法发现。
3. 测试基建耦合：DefaultServerBase.server_env 为 None 时不会传 env，但若子类误设为空 dict 会覆盖默认环境；popen_launch_server 的具体行为依赖测试工具，存在隐式耦合。
4. 运行时回归：装饰器默认关闭，理论上零开销；但若未来默认打开或误用，write_backup_storage 等无参数声明的全量校验可能因对象不可哈希或含大张量而引入额外通信 / 序列化成本。- 影响：对生产用户无行为影响（校验器默认关闭）。对开发团队，这是一次调试基建增强：多 rank 场景下可自动发现 HiCache 缓存状态分歧，显著降低 TP/PP 下缓存一致性问题的排查成本。对测试套件，新增环境变量只在指定测试类生效，不影响其他测试；CI 多跑一个下游检查，但带来的主要是测试稳定性收益。- 风险标记：装饰器默认关闭，运行时零开销，same_params 选择面窄可能漏检，测试开启校验可能引入 CI 误报，server_env 透传存在隐式耦合

关联脉络

- PR #38483 [HiCache] Release buffer prefetch anchor locks during storage cleanup: 同一 HiCache 存储生命周期模块的 bug 修复，rank 一致性校验可帮助发现此类跨 rank 状态泄漏问题。
- PR #38486 [HiCache] Publish a host store event for storage-prefetch refills: HiCache 存储预取事件发布机制，本 PR 对 prefetch_from_storage 的校验与此类预取一致性直接相关。
- PR #37914 [Unified Cache][7/N] Support MTP, EAGLE, and DSpark draft KV caches in the external linker: UnifiedRadixCache 外部链接器演进，本 PR 为这些新路径的缓存操作补充了 rank 一致性检查。
- PR #39162 [HiCache][LoRA] Simplify decode offload hash inputs: HiCache 解码卸载哈希入参重构，涉及多 rank 哈希一致性，与本 PR 的共识校验目标互补。