

PR #39110 完整报告

sgl-project/sglang

[Router] Real-GPU e2e coverage for storage-tier-aware cache routing (3/4)

合并时间: 2026-09-16 06:48

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39110>

执行摘要

- 一句话: 真实 GPU e2e 测试覆盖 HiCache 分层缓存路由
- 推荐动作: 值得精读。虽然纯测试变更, 但测试设计方法论有很高参考价值: 用引擎自身记账而非树状态做 payoff 断言以避免 vacuous pass; 按 owner 份额而非集群总量计算驱逐预算; 驱动式而非固定请求数; 验证前提 (双 worker 注册) 而非假设; 参数化覆盖面相反事件顺序。建议结合系列前序 PR (tier-aware tree 与 metrics 支撑) 一起阅读, 并关注后续第 4 部分。

功能与动机

PR body 明确说明: 此前所有 tier 逻辑的测试都是合成事件序列, 无法覆盖真实引擎的事件产生通路——这是 "唯一能抓到 **medium** 标签在这条链路上任何一环丢失的方式", 也是唯一能断言 payoff (被驱逐的前缀仍由持有它的层服务) 而非仅断言树状态的方式。同时指出双写策略事件顺序相反 (**write_through** 先发布 CPU_PINNED store 再发布 GPU removal, **write_back** 反之), "一个只处理了 write_through 顺序的树会在所有稳态检查中看起来正确, 却在 write_back 集群上丢掉前缀"。

实现拆解

1. 新增真实 GPU e2e 测试文件 `experimental/sgl-router/tests/e2e/chat_completions/test_hicache_storage_tiers.py` (482 行): 通过 `spawn_worker` 拉起两个真实引擎 worker, `worker_args()` 统一封装 `--enable-hierarchical-cache --hicache-ratio 4 --max-total-tokens 8192 --page-size 64`, 刻意缩小设备 KV 池使少数长 prompt 即可在数秒内翻转设备层, 同时 host 层按 4 倍比例保留被驱逐块; 经 `--kv-events-config` 将引擎事件经 ZMQ 发布给路由器。
2. 以引擎自身记账作为 payoff 断言: `_chat()` 通过 `return_cached_tokens_details` 请求参数让响应携带 `sglext.cached_tokens_details`, `_cached_tiers()` 提取每层缓存 token 拆分并丢弃非整数字段, 断言 `host > 0` 是测试名 `test_repeat_is_served_from_host` 的核心主张——tier-blind 树在前缀仍在设备上时也会正确路由, 因此仅路由断言无法区分修复是否生效。
3. 驱动式反脆弱循环: `_drive_until_host_served` 以观察效果而非固定请求数驱动, 每轮先发 `FILLERS_PER_PROBE = 16` 条 filler 请求 (按 owner 份额而非集群总量计算, 约 2 倍翻面 owner 的设备池), 再发探测请求; 探针本身会重新预热前缀, 故每轮之间必须用 filler 冲刷。`_wait_until` 轮询容忍事件异步到达的延迟, `_chat_and_attribute` 等计数器入账后再

返回。

4. 双写策略参数化与前提验证: `WRITE_POLICIES = ["write_through", "write_back"]` 覆盖相反事件顺序; `_registered_workers` 先从 `sgl_router_worker_health` 确认两个 worker 均已注册, 防止单 worker 注册失败导致所有请求落在幸存者上造成 " 隐性通过 "。
5. 测试基建修复 `experimental/sgl-router/tests/e2e/infra/model_pool.py`: `spawn_worker` 把 `stdout=subprocess.PIPE` 改为行缓冲日志文件 `/tmp/sglang-worker-{port}.log`, 避免父进程不排空管道导致 64 KB 缓冲填满后引擎阻塞写、请求挂到超时且无日志; 新增 `ModelInstance.log_tail()` 并在启动失败、健康检查超时错误中附带日志尾部。
6. CI 配套: 无 workflow 变更, 由现有 `pr-test-sgl-router.yml` 的 `pytest tests/e2e/` 步骤收集; 在 2-gpu-h100 runner 上 10 个用例 (双参数化) 6 分 25 秒全部通过。

关键文件:

- `experimental/sgl-router/tests/e2e/chat_completions/test_hicache_storage_tiers.py` (模块 层级缓存; 类别 `test`; 类型 `test-coverage`; 符号 `worker_args`, `_scrape`, `_samples`, `_sum_where`): 新增 482 行真实 GPU e2e 测试, 是存储层级感知缓存路由修复的端到端验证核心: 驱动整条事件链路并断言被驱逐前级仍由 host 层服务, 参数化两种写策略。
- `experimental/sgl-router/tests/e2e/infra/model_pool.py` (模块 测试基建; 类别 `test`; 类型 `test-coverage`; 符号 `log_tail`, `spawn_worker`): 修复 `spawn_worker` 将引擎输出写入无人读取的 `subprocess.PIPE` 导致 64 KB 缓冲填满后引擎阻塞写、请求挂起无日志的隐患, 新增 `log_tail()` 诊断能力, 惠及所有使用该基建的 e2e 测试。

关键符号: `worker_args`, `_drive_until_host_served`, `_chat_and_attribute`, `_cached_tiers`, `_wait_until`, `_registered_workers`, `_tier_summary`, `log_tail`

关键源码片段

`experimental/sgl-router/tests/e2e/infra/model_pool.py`

修复 `spawn_worker` 将引擎输出写入无人读取的 `subprocess.PIPE` 导致 64 KB 缓冲填满后引擎阻塞写、请求挂起无日志的隐患, 新增 `log_tail()` 诊断能力, 惠及所有使用该基建的 e2e 测试。

```
# 把引擎输出写入日志文件, 而不是无人读取的 PIPE。父进程从不排空
# 该管道, ~64 KB 的 OS 缓冲区一旦填满, 引擎阻塞在 write 上、停止
# 服务——请求挂到客户端超时, 且没有任何日志能解释原因。仅启动阶段
# (权重加载、内存池、CUDA graph 捕获) 就可能逼近这个量, 长测试的
# 每请求日志则会直接超过它。conftest.py 的 session 级 fixture 早已
# 踩过这个坑并修复; 这里对 per-test worker 应用同样的修复。
log_path = Path(tempfile.gettempdir()) / f"sglang-worker-{port}.log"
log_handle = open(log_path, "w", buffering=1) # 行缓冲, 便于实时查看
try:
    proc = subprocess.Popen(
        cmd,
        env=env,
        stdout=log_handle,
        stderr=subprocess.STDOUT,
        start_new_session=True,
```

```
)  
finally:  
    # 子进程持有自己的文件描述符，父进程的这份可以立刻关闭。  
    # 一直持有着，每个 worker 泄漏一个 fd，且文件在有写方之前  
    # 就被占住。失败诊断从 log_path 读回，而不是从这个句柄读。  
    log_handle.close()
```

评论区精华

sherlockwu 提出唯一一条实质讨论：priming 阶段背靠背发送的两次请求存在竞态——第一次 HTTP 响应可能先于缓存事件到达路由器，第二次请求落在间隙内会被负载均衡送到另一 worker，导致两个 worker 都持有前缀，`_sole_device_owner` 检查超时，而路由器行为完全正确。建议把 owner-wait 移到两次请求之间并确认第二次请求落到同一 owner。作者以 commit 738266b 修复并回复 "fixed"，随后 sherlockwu APPROVED。

- priming 阶段请求竞态：事件未达路由器前二次请求可能落错 worker (correctness)：作者回复 fixed，commit 738266b 将 owner 等待移动到两次请求之间：prime 后先等树在唯一 worker 上索引前缀，再发第二次请求并确认落到同一 owner。

风险与影响

- 风险：
 1. 依赖引擎私有扩展字段：sglctx.cached_tokens_details 是引擎返回体中的非标准扩展，若该字段被改名或删除，测试会整体失效（呈失败态而非误通过，可接受但需注意）。
 2. 参数依赖 tokenizer 实测：PROMPT_WORDS = 300 与 FILLERS_PER_PROBE = 16 基于 Qwen3 tokenizer 实测（primed ~1.1k tokens、filler ~2.3k tokens），更换模型规格需重新标定，否则翻面预算可能不足导致测试超时。
 3. CI 资源依赖：需 2-gpu-h100 双卡 runner，GPU 不可用时会跳过；测试耗时 6 分半，负载高时有超时风险。PR Test (Extra) CI 显示失败（Run #35024684491），正文未说明具体原因，需关注是否为环境问题。
 4. 日志文件残留：/tmp/sglang-worker-{port}.log 测试结束后不会主动清理，长时间运行会在 runner 上积累；端口随机分配避免冲突，但磁盘占用需留意。
 5. 无生产代码回归面：model_pool.py 修改影响所有使用 spawn_worker 的 e2e 测试，改为日志文件后父进程不再直接持有 stdout，需确认无其他代码依赖 proc.stdout（当前实现已读回日志文件，逻辑等价）。- 影响：用户与生产系统无直接影响（无 src/ 变更）。对团队而言，这是 HiCache tier-aware 缓存路由修复（系列 #39108/#39109）唯一的真实 GPU 回归防线，能捕捉事件链路上任意一环丢失 medium 标签的缺陷；PR body 报告 tier-blind 对照（H200 上两种参数化均失败）证明测试有判别力。model_pool.py 的 PIPE 修复惠及所有使用 spawn_worker 的 sgl-router e2e 测试，消除一类 " 请求挂起且无日志 " 的难排查故障。- 风险标记：依赖引擎私有扩展字段，需 GPU 双卡 CI 环境，长耗时有例约 6 分半，filler 参数依赖 tokenizer 实测，extra CI 状态失败待查

关联脉络

- PR #39283 [HiCache] Rework the buffer-mode storage prefetch pipeline and retry bookkeeping: 同一 HiCache 功能线，涉及 kv_events 事件流、树记账与存储预取；本测试正是对事件流全链路端到端的验证，与该预取重构形成验证闭环。
- PR #39480 [HiCache] Optimize buffer-mode storage existence bookkeeping: 同为 HiCache 存储层优化，本测试 host 层保留断言依赖 buffer-mode 记账的正确性。
- PR #39516 [Fix] HiCache startup ImportError on the pinned kernel wheel: HiCache 相关修复的回归防线，与本测试同属 hicache 标签下的验证工作。