

PR #37194 完整报告

sgl-project/sglang

[Fix] Shut hicache test servers down gracefully before SIGKILL

合并时间: 2026-08-31 13:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37194>

执行摘要

- 一句话: 测试服务器关机改为优雅关闭, 释放钉住主机 KV 池
- 推荐动作: 值得快速浏览而非精读。作为“测试进程清理标准化”的示例, `test_hicache_storage.py` 中清晰的动机注释和统一 helper 的使用方式是主要看点; 不需要深入理解 HiCache 内部实现。

功能与动机

PR body 明确说明了动机: Replaces bare `kill_process_tree` with the existing `terminate_and_kill_process_tree` helper in tests that enable hierarchical cache, so the server releases its pinned host KV pool in userspace instead of leaving the kernel to unpin it during reclaim. 即裸 SIGKILL 会让内核延迟解钉, 显存残留时间过长, 导致下一个测试因显存不足而失败。

实现拆解

1. 锁定目标范围: 覆盖所有启用 `--enable-hierarchical-cache` 的 registered 测试, 包括 HiCache 存储、变体、PP、Qwen3.5、NPU、AMD MI35X、Inkling-Small-NVFP4、Kimi-K3-B300 以及 PD 解耦卸载测试, 共 13 个文件。
2. 统一导入来源: 从 `sglang.test.test_utils` 导入 `terminate_and_kill_process_tree`, 同时移除 `sglang.srt.utils` 中不再使用的 `kill_process_tree` 引用, 将测试进程清理逻辑收敛到同一个工具函数。
3. 替换手写关闭序列: `test_hicache_storage.py` 中原来的 `terminate() + wait(timeout=60) + kill_process_tree(pid)` 三行被 helper 单行替代; `test_disaggregation_decode_offload.py` 中 `prefill/decode/load balancer` 三个进程的 `kill_process_tree(pid) + wait()` 也改为三次 helper 调用, 语义等价 (helper 内部负责 `TERM` → 等待 → `KILL` 的完整序列)。
4. 保留必要的收尾等待: 各测试继续保留 `time.sleep(5)` 或 `_wait_for_gpu_idle_in_ci(timeout=...)`, 给驱动与内存回收留出余量。
5. CI 配套验证: 作者通过 `/rerun-test` 在 1-gpu-h100、1-gpu-5090、2-gpu-h100、4-gpu-h100 等 runner 上重跑了 7 个目标测试, 全部通过; 无新增配置或部署改动。

关键文件:

- test/registered/hicache/test_hicache_storage.py (模块 缓存存储; 类别 test; 类型 test-coverage; 符号 TestHiCache.tearDownClass) : 改动最核心的 HiCache 存储测试, 原手写 terminate+wait+kill 序列被 helper 单行替代, 保留了对钉住主机 KV 池泄漏问题的解释性注释。
- test/registered/models_e2e/test_inkling_small_nvfp4.py (模块 模型精度; 类别 test; 类型 test-coverage) : 覆盖 B200 4 卡 NVFP4 精度测试, 三处 tearDownClass 的 kill_process_tree 全部替换为 helper, 涉及 Mamba/HiCache 组合路径。
- test/registered/disaggregation/test_disaggregation_decode_offload.py (模块 PD 解耦; 类别 test; 类型 test-coverage; 符号 TestDisaggregationDecodeOffload.test_mmlu_double_eval) : PD 解耦卸载测试中三个进程 (prefill/decode/load balancer) 的 kill+wait 被 helper 替代, 删除显式 wait, 是行为变化最明显的一处。
- test/registered/models_e2e/test_kimi_k3_b300.py (模块 K3 机型; 类别 test; 类型 test-coverage; 符号 _stop_server) : B300 8 卡 Kimi-K3 测试, 通过 _stop_server 辅助函数统一替换, 涉及代号 kimi_k3 解析器与 HiCache。
- test/registered/hicache/test_hicache_storage_runtime_attach_detach.py (模块 运行时挂载; 类别 test; 类型 test-coverage) : HiCache 存储后端运行时 attach/detach 生命周期测试, 两处 finally 清理路径替换为 helper。
- test/registered/hicache/test_hicache_variants.py (模块 缓存变体; 类别 test; 类型 test-coverage; 符号 HiCacheBaseServer.tearDownClass) : HiCache 多配置变体测试 (standard/MLA/EAGLE/page size), 基类 tearDownClass 统一替换, 影响多个子类。
- test/registered/amd/test_deepseek_r1_hicache_mi35x.py (模块 DeepSeek 缓存; 类别 test; 类型 test-coverage) : AMD MI35X 上的 DeepSeek-R1 + HiCache 测试, 收尾清理同时处理进程与 L3 存储目录。
- test/registered/hicache/test_hicache_storage_file_backend.py (模块 文件后端; 类别 test; 类型 test-coverage) : HiCache file 后端存储测试, tearDownClass 中进程清理替换为 helper, 并保留 shutil 清理。
- test/registered/hicache/test_pp_with_hicache.py (模块 管道并行; 类别 test; 类型 test-coverage) : PP + HiCache 组合测试, tearDownClass 中替换进程清理, 并保留 MoonCake 服务停止逻辑。
- test/registered/hicache/test_qwen35_hicache.py (模块 Qwen 缓存; 类别 test; 类型 test-coverage) : Qwen3.5 + HiCache 测试, tearDownClass 中替换进程清理, 同时保留 storage_dir 清理。
- test/registered/npu/basic_function/HiCache/test_npu_hicache_mha.py (模块 NPU 缓存; 类别 test; 类型 test-coverage) : NPU 上的 HiCache MHA 测试, 将进程清理替换为优雅关闭, 覆盖 NPU 平台。
- test/registered/npu/basic_function/HiCache/test_npu_hicache_mla.py (模块 NPU 缓存; 类别 test; 类型 test-coverage) : NPU 上的 HiCache MLA 测试, 清理逻辑同步更新。
- test/registered/npu/basic_function/HiCache/test_npu_hierarchical_cache.py (模块 NPU 缓存; 类别 test; 类型 test-coverage) : NPU 上的通用层次化缓存测试, 清理逻辑同步更新。

关键符号：TestHiCache.tearDownClass, HiCacheBaseServer.tearDownClass, _stop_server, TestDisaggregationDecodeOffload.test_mmlu_double_eval

关键源码片段

test/registered/hicache/test_hicache_storage.py

改动最核心的 HiCache 存储测试，原手写 terminate+wait+kill 序列被 helper 单行替代，保留了对钉住主机 KV 池泄漏问题的解释性注释。

```
class TestHiCache(CustomTestCase, MMLUMixin):
    @classmethod
    def setUpClass(cls):
        # 启动带分层缓存的服务器：file 后端、大页，hicache 容量按平台调整
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--enable-hierarchical-cache",
                "--mem-fraction-static", 0.7,
                "--hicache-size", 100 if not _is_hip else 200,
                "--page-size", "64",
                "--hicache-storage-backend", "file",
            ],
        )

    @classmethod
    def tearDownClass(cls):
        # 先发 SIGTERM 让服务器在用户态注销钉住的主机 KV 池，
        # 超时后再 SIGKILL；裸 kill 会让内核在页回收时才解钉，
        # 显存迟迟不释放，会拖垮下一个测试用例。
        terminate_and_kill_process_tree(cls.process)
        time.sleep(5) # 给驱动与内存回收留出余量
```

test/registered/disaggregation/test_disaggregation_decode_offload.py

PD 解耦卸载测试中三个进程（prefill/decode/load balancer）的 kill+wait 被 helper 替代，删除显式 wait，是行为变化最明显的一处。

```
def test_mmlu_double_eval(self):
    """两轮 MMLU：先 offload 到磁盘，重启节点后再加载，验证分数一致。"""
    args = SimpleNamespace(
        base_url=f"http://{self.base_host}:{self.lb_port}",
        model=self.model,
        eval_name="mmlu",
        num_examples=256,
        num_threads=32,
    )
    metrics1 = run_eval(args)
    time.sleep(10) # 确保所有 offload 都提交到磁盘
```

```
# 依次优雅关闭 prefill / decode / load balancer 三个进程。
# 相比裸 kill_process_tree: 先 SIGTERM 让服务器完成 KV 池注销,
# 超时后再 SIGKILL; helper 内部负责等待退出, 无需显式 wait()。
terminate_and_kill_process_tree(self.process_prefill)
terminate_and_kill_process_tree(self.process_decode)
terminate_and_kill_process_tree(self.process_lb)

self.start_prefill()
self.start_decode()
self.launch_lb()
self.wait_server_ready(self.prefill_url + "/health")
self.wait_server_ready(self.decode_url + "/health")
metrics2 = run_eval(args)
```

评论区精华

本 PR 无实质 review 讨论 (review_comments_count 为 0) , Issue 评论仅为流程性操作: 作者发出 `/rerun-test test_hicache_variants.py ...` 和 `/tag-and-rerun-ci`, github-actions bot 汇总了各 runner 的重跑结果, 全部为 🟢。

- 暂无高价值评论线程

风险与影响

- 风险: 风险面仅限于测试基础设施, 无产品代码变更。关键点如下:
 - `terminate_and_kill_process_tree` 依赖服务器对 SIGTERM 的响应; 若服务器无法在内部超时内退出, helper 会回退到 SIGKILL, 行为与旧的裸 kill 相同, 因此最坏情况不会比之前更差。
 - `test_disaggregation_decode_offload.py` 删除了显式 `wait()`, 改为依赖 helper 内部等待, 语义等价, 但未来若修改 helper 的等待行为, 需回归该测试。
 - 涉及 13 个文件、多平台 (AMD、NPU、Blackwell B200/B300、H100), 但全部是同一模式的机械替换, 批量替换引入不一致的风险较低。
 - 改动集中在 `tearDownClass/_stop_server` 等收尾路径, 不影响任何测试断言逻辑。
 - 影响: 对用户无影响; 对 CI 稳定性有明显正面影响。HiCache 测试常启用大容量的钉住主机内存池 (如 `--hicache-size 100/200`、file 后端、direct I/O), 连续跑多个用例时, 裸 SIGKILL 留下的内核解钉延迟会占用 GPU 显存并导致后续用例失败。改为优雅关闭后, 服务器能在用户态完成 KV 池注销, 显存及时释放, 预计可减少 HiCache/PD 测试队列的偶发失败。对测试维护者而言, 本 PR 确立了统一的进程清理惯用法, 后续新测试可直接复用 `terminate_and_kill_process_tree`。
 - 风险标记: 测试基础设施变更, 多文件机械替换, 依赖信号处理助手, 无源码主路径变更

关联脉络

- PR #36798 [HiCache] Align chunked CUDA host registrations: 与钉住主机内存池的注册 / 解钉机制直接相关, 本 PR 的优雅关闭正是为了在用户态完成该池的注销。

- PR #36834 [HiCache] buffer mode: decide staged-fetch fate against the live tree: 同一 HiCache 测试与调度链路的近期改动，涉及缓存树与预取行为，本 PR 是其测试稳定性的配套收尾。
- PR #37151 [Unified Cache Linker][3/N]: Add backend-independent linker core: HiCache/ 统一外部缓存链路的演进主线，本 PR 确保相关测试服务器能干净退出，为后续链路测试铺路。