PR #22815 完整报告

sgl-project/sglang

Add page_size and SWA coverage to unified radix cache bench test

合并时间：2026-04-14 23:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22815

执行摘要

- 一句话：为统一 radix 缓存基准测试添加页面大小和滑动窗口注意力覆盖。
- 推荐动作：建议关注新增的 _alloc 函数和参数化测试设计，这对于理解缓存分配在 SWA 和不同页面大小下的行为有价值。如果是测试或缓存模块的开发者，值得精读以了解测试扩展方法。

功能与动机

PR 标题表明动机是添加 page_size 和 SWA 覆盖到基准测试中，以测试更多缓存配置。虽然没有明确的 issue 或 body 描述，但从变更内容推断，目的是提升测试的完整性和准确性，确保缓存系统在不同参数下的正确性和性能。

实现拆解

实现集中在 test/registered/unit/mem_cache/test_unified_radix_cache_bench.py 文件中：1) 更新 CI 注册时间从 60 秒增加到 120 秒，以适应更长的测试运行；2) 修改 create_bench_cache 函数，添加 sliding_window_size 参数，并调整逻辑以支持 SWA 和页面大小；3) 扩展 _make_env 函数，增加 page_size 参数；4) 新增 _alloc 函数，处理 SWA 和页面大小大于 1 时的对齐分配；5) 更新 _alloc_with_evict 函数使用新分配逻辑。

关键文件：

- test/registered/unit/mem_cache/test_unified_radix_cache_bench.py（模块 测试 / 缓存基准测试）：这是统一 radix 缓存基准测试的主文件，所有修改都在此文件中，添加了 page_size 和 SWA 支持，是 PR 的核心变更。

关键符号：create_bench_cache, _make_env, _alloc, _alloc_with_evict

评论区精华

review 过程中没有具体讨论，只有 reviewer hzh0425 的批准，表明变更被认为直接且无争议，已通过审查。

- 批准变更 (other): 变更被接受。

风险与影响

- 风险：风险较低：1) 新添加的 _alloc 函数逻辑可能出错，影响基准测试结果的准确性；2) CI 时间加倍可能延长测试流水线效率；3) 如果 page_size 或 SWA 处理不当，可能导致测试覆盖不准确。但由于是测试代码，不影响生产环境。
- 影响：影响范围限于测试套件，特别是统一 radix 缓存的基准测试模块。影响程度低：对最终用户无直接影响，但有助于开发者更全面地测试缓存性能，可能间接提升系统可靠性；CI 时间增加可能轻微影响开发效率。
- 风险标记：测试逻辑变更，CI 时间增加

关联脉络

- PR #22812 Refactor unified radix cache UT into parameterized test suite: 都修改了统一 radix 缓存的测试代码，本 PR 扩展基准测试，PR 22812 重构单元测试，关联紧密。