

PR #22053 完整报告

sgl-project/sglang

[HiCache & Bench] add cache hit breakdown in bench_serving

合并时间: 2026-06-18 07:02

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/22053>

执行摘要

- 一句话: bench_serving 新增 --cache-report 展示缓存命中细分
- 推荐动作: 该 PR 代码清晰、测试充分 (通过现有 CI), 适合合并。建议后续为该功能添加单元测试, 确保 _extract_cache_from_sglex 在不同响应格式下的稳定性。审核时已注意兼容性, 值得参考其安全处理 null 的方式。

功能与动机

在测试 HiCache 时, 通常需要查看缓存命中情况、哪些缓存层被使用以及哪些层贡献了命中。SGLang 已支持每个请求返回 cache details (PR #17648 & #21764), 但 bench_serving 缺乏聚合展示。本 PR 通过 --cache-report 标志启用该功能。

实现拆解

1. 扩展输出数据结构: 在 RequestFuncOutput dataclass 中新增 cached_tokens (int) 和 cached_tokens_details (Optional[Dict]) 字段, 用于存储每个请求的缓存命中计数和详细来源。
2. 新增解析函数: 定义 _extract_cache_from_sglex(data, output) 函数, 从 OpenAI 兼容的响应体中的 sglex 字段提取 cached_tokens_details, 并安全地累加 device、host、storage 三级缓存到 cached_tokens。
3. 注入到请求处理流程: 在 async_request_openai_completions (流式响应时)、async_request_openai_chat_completions (流式和非流式) 以及 async_request_sglang_generate (从 meta_info 直接读取) 中, 当 --cache-report 开启时调用 _extract_cache_from_sglex 或直接赋值。
4. 聚合与报表输出: 在结果汇总部分, 通过单次遍历 outputs 累加 total_prompt_tokens、total_cached、以及各层级计数, 最后计算缓存命中率并打印层级占比表格。

关键文件:

- python/sglang/bench_serving.py (模块 工具脚本; 类别 source; 类型 core-logic; 符号 _extract_cache_from_sglex): 该文件是 bench_serving 工具的唯一实现, 本次新增了 --cache-report 标志的解析、缓存命中数据的提取逻辑以及聚合输出功能。所有变更集中于此文件。

关键符号: _extract_cache_from_sglex

评论区精华

Review 中 bot 提出两点改进：

- 在 `_extract_cache_from_sglex` 中，使用 `(details.get("field") or 0)` 替代 `details.get("field", 0)` 以正确处理服务器返回的显式 `null` 值，避免 `TypeError`。
- 在聚合打印部分，建议合并多次循环为单次遍历，提升性能。开发者采纳了这些建议（最终代码中已体现）。此外，审核人 `xiezhq-hermann` 批准了 PR，并提醒注意不影响其他基准测试。
- `null` 值安全性：在 `_extract_cache_from_sglex` 中使用 `(details.get("field") or 0)` 替代 `details.get("field", 0)` (correctness)：作者采纳了建议，在最终代码中使用了 `or 0` 模式。
- 聚合指标循环优化 (performance)：作者采纳建议，最终代码使用单次循环计算所有聚合指标。

风险与影响

- 风险：主要风险是新增的 `--cache-report` 标志和解析逻辑可能对现有基准测试流程造成轻微影响：
 - 如果服务器返回的 `sglex` 格式不兼容，`_extract_cache_from_sglex` 可能跳过或抛出异常。当前实现通过 `or 0` 安全处理了 `null`，且只在标志开启时才触发，默认不影响原有行为。
 - 聚合循环增加了一次对 `outputs` 的遍历，但规模通常不大，性能影响可忽略。
 - 没有新增测试文件，但该功能是对已有工具的扩展，风险较低。
- 影响：对用户：为用户（尤其是 `HiCache` 开发者）提供了可视化的缓存命中明细，帮助定位缓存效率问题。对系统：修改仅限基准测试工具，不影响推理服务核心逻辑。对团队：增强了 `SGLang benchmark` 的可观测性，有助于缓存层优化。影响度较小，仅限工具层面。
- 风险标记：新增命令行参数可能影响现有测试，依赖服务端 `sglex` 格式

关联脉络

- PR #17648 Add per-request cache details to SGLang responses: 此 PR 在服务端添加了每个请求的缓存细节，本 PR 依赖该功能在 `bench_serving` 中聚合展示。
- PR #21764 Enhance cache details reporting: 进一步增强了缓存细节报告，本 PR 使用的 `cached_tokens_details` 格式来源于此。