

PR #29332 完整报告

sgl-project/sglang

[test] Report token tps in fwd occupancy kit and force ignore_eos

合并时间: 2026-06-26 06:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29332>

执行摘要

- 一句话: 前向占用率测试工具增加 token TPS 报告
- 推荐动作: 建议合并。该 PR 小而精, 增强了测试工具的可观测性, 代码变更清晰无副作用。
值得关注的设计决策: 将 TPS 计时放在请求内部而不是采样循环, 避免了尾部采样间隔的干扰, 使 TPS 更准确。

功能与动机

原始 `fwd_occupancy` 工具只报告 GPU 占用率, 但缺少实际的单流解码吞吐量。添加 `token tps` 行使得报告将占用率与 tokens/s 配对, 便于回归分析。此外, 贪婪解码下代码生成可能早于 `max_new_tokens` 发出 EOS, 这会缩短占用率采样窗口 (可能低于 `min_samples`) 并稀释测得的 TPS。

实现拆解

1. `_fwd_occupancy_fire` 返回值扩展: 在 `python/sglang/test/kits/fwd_occupancy_kit.py` 中, 该方法现在返回 `(completion_tokens, wall_time)` 元组。新增 `ignore_eos: True` 参数确保解码窗口始终运行到 `max_new_tokens`。通过 `time.perf_counter` 精确计时。
2. `_fwd_occupancy_measure` 返回 TPS: 改为返回 `(samples, token_tps)`, 其中 `token_tps = completion_tokens / elapsed`。使用线程安全字典传递结果。
3. `test_fwd_occupancy` 打印 TPS: 解包元组并打印 `token_tps` 行, 仅提供信息, 不添加断言。
4. 断言风格调整: `_assert_metrics_device_timer_enabled` 中的 `assert` 改为显式 `raise AssertionError`, 避免被 `-O` 选项剥离。

关键文件:

- `python/sglang/test/kits/fwd_occupancy_kit.py` (模块 测试工具; 类别 `test`; 类型 `test-coverage`; 符号 `_fwd_occupancy_fire`, `_fwd_occupancy_measure`, `test_fwd_occupancy`, `_assert_metrics_device_timer_enabled`): 该文件是唯一的变更文件, 实现了所有功能增强: token TPS 报告、`ignore_eos`、断言风格调整。

关键符号: `_fwd_occupancy_fire`, `_fwd_occupancy_measure`, `test_fwd_occupancy`, `_assert_metrics_device_timer_enabled`

评论区精华

无

- 暂无高价值评论线程

风险与影响

- 风险：风险极低： 1) `ignore_eos=True` 仅影响测试请求，不会触及生产代码。 2) TPS 仅打印报告，不改变任何断言逻辑。 3) 返回元组变化会破坏下游，但该模块仅用于内部测试，不公开 API。 4) `assert` 改为 `raise AssertionError` 不会影响行为，但更可靠。
- 影响：影响范围限于 `fwd_occupancy_kit.py` 这一测试工具文件。用户可以通过报表直接看到 token TPS 和占用率的对应关系，便于性能回归分析。对系统性能无影响。
- 风险标记：仅测试文件变更

关联脉络

- 暂无明显关联 PR