

PR #2169 完整报告

THUDM/slime

Merging profiling info into router

合并时间: 2026-07-02 21:18

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2169>

执行摘要

- 一句话: 将 Profiling 信息合并到 Router, 重构 sglang 补丁
- 推荐动作: 建议仔细 review router 侧对应的 profiling 实现是否已合入, 并充分测试 PD 分离场景下的 timeout 和 memory 释放逻辑。

功能与动机

旨在将 profiling 信息的收集和聚合从 sglang 后端转移到 router 组件, 降低对 sglang 版本的侵入性, 并集中管理性能数据。从变更可见, 原本在 decode 中通过 `apply_prefill_timing_payload`、`is_slime_profiling_enabled` 等函数执行的 profiling 逻辑被移除, 转而依赖 router 侧的能力。

实现拆解

1. 重构 sglang.patch: 在 `docker/patch/latest/sglang.patch` 中移除了与 profiling 相关的导入和函数调用 (如 `apply_prefill_timing_payload`、`is_slime_profiling_enabled`), 并添加了 `release_memory_occupation` 和 `resume_memory_occupation` 方法以支持显存释放。同时引入了 bootstrapping 超时处理机制。
2. 增强 sglang-top_p.patch: 在 `MetadataBuffers` 类中添加 `output_top_p_token_ids_len` 和 `output_top_p_token_ids` 缓冲区, 并在 `DecodeTransferQueue` 中传输这些数据, 以支持 top_p 采样的 token IDs 回传。
3. 移除 rollout 中的 profiling 标志: 在 `slime/ray/rollout.py` 中去除 `SLIME_ENABLE_PROFILING="true"` 环境变量, 因为 profiling 已由 router 接管。
4. 更新测试用例: 修改 `tests/utils/test_trace_utils.py`, 验证 `build_sglang_meta_trace_attrs` 返回的 trace 属性包含 `sglang_pd_prefill` 和 `sglang_pd_decode` 子 span。
5. 配套基础设施: 更新 `docker/Dockerfile` 中 sglang_router 包的版本和 `docker/version.txt`。

关键文件:

- `docker/patch/latest/sglang.patch` (模块 sglang 补丁; 类别 test; 类型 test-coverage; 符号 `DisaggregationMode`, `get_buf_infos`, `get_buf`, `set_buf`): 核心变更, 移除 profiling 代码, 添加超时和内存管理方法, 重构最大
- `docker/patch/latest/sglang-top_p.patch` (模块 top-p 补丁; 类别 test; 类型 test-coverage; 符号 `MAX_PD_TOP_P_TOKEN_IDS`, `MetadataBuffers`, `get_buf_infos`):

新增 top_p_token_ids 缓冲区支持, 增强 PD 传输能力

- tests/utils/test_trace_utils.py (模块 trace 工具; 类别 test; 类型 test-coverage; 符号 build_sglang_meta_trace_attrs, TRACE_CHILDREN_KEY) : 验证新的 trace 子 span 结构, 确保 build_sglang_meta_trace_attrs 行为正确
- slime/ray/rollout.py (模块 rollout 引擎; 类别 source; 类型 core-logic; 符号 start_engines) : 移除 profiling 环境变量, 标志着 profiling 逻辑迁移完成
- docker/Dockerfile (模块 Docker 镜像; 类别 infra; 类型 infrastructure) : 更新 sglang_router 包版本以匹配 profiling 合并
- docker/version.txt (模块 版本文件; 类别 docs; 类型 documentation) : 更新版本号, 标记镜像版本

关键符号: build_sglang_meta_trace_attrs, start_engines, release_memory_occupation, resume_memory_occupation, get_buf, get_buf_infos

关键源码片段

`docker/patch/latest/sglang-top_p.patch`

新增 top_p_token_ids 缓冲区支持, 增强 PD 传输能力

docker/patch/latest/sglang-top_p.patch 修改后的 MetadataBuffers 新增部分
MAX_PD_TOP_P_TOKEN_IDS = 4096

```
class MetadataBuffers:
    def __init__(self, size, max_top_logprobs_num, device):
        # ... 原有初始化 ...
        # 新增: 为 top_p token ids 分配固定大小缓冲区
        self.output_top_p_token_ids_len = torch.zeros(
            (size, 16), dtype=torch.int32, device=device
        )
        self.output_top_p_token_ids = torch.zeros(
            (size, MAX_PD_TOP_P_TOKEN_IDS), dtype=torch.int32, device=device
        )

    def get_buf(self, idx):
        # 返回缓冲区切片时包含新增字段
        return (
            # ... 原有返回值 ...
            self.output_top_p_token_ids_len[idx].clone(),
            self.output_top_p_token_ids[idx].clone(),
        )
```

`tests/utils/test_trace_utils.py`

验证新的 trace 子 span 结构, 确保 build_sglang_meta_trace_attrs 行为正确

```
# tests/utils/test_trace_utils.py 修改后的测试函数
@pytest.mark.unit
def test_build_sglang_meta_trace_attrs_keeps_standard_and_pd_fields():
```

```

meta = {
    "prompt_tokens": 12,
    "completion_tokens": 7,
    "cached_tokens": 3,
    "pd_prefill_forward_duration": 0.125,
    "pd_decode_transfer_duration": 0.05,
    "finish_reason": {"type": "stop"},
    "unused_field": "ignored",
}

attrs = build_sglang_meta_trace_attrs(meta)
# 弹出子 span 字典
trace_children = attrs.pop	TRACE_CHILDREN_KEY)

# 验证标准字段被保留, finish_reason 被展平
assert attrs == {
    "prompt_tokens": 12,
    "completion_tokens": 7,
    "cached_tokens": 3,
    "finish_reason": "stop",
}
# 验证 trace 子 span 包含 prefill 和 decode 信息
assert trace_children[0]["name"] == "sglang_pd_prefill"
assert trace_children[0]["children"][0]["attrs"] == {
    "pd_prefill_forward_duration": 0.125,
}
assert trace_children[1]["name"] == "sglang_pd_decode"
assert trace_children[1]["children"][0]["attrs"] == {
    "pd_decode_transfer_duration": 0.05,
}

```

评论区精华

由于该 PR 无 review 评论，无讨论值得提炼。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于 `sglang.patch` 的大范围重构（+126/-610），可能影响 PD 分离系统的稳定性。新引入的 bootstrapping 超时逻辑可能在某些场景下过早中断请求。移除 `SLIME_ENABLE_PROFILING` 环境变量后，若 router 尚未完整实现 profiling 聚合，会导致数据丢失。
- 影响：对开发者而言，sglang 补丁维护复杂度降低；对系统而言，profiling 数据流发生变化，需要确保 router 版本兼容。对用户透明，但可能影响性能分析工具的输出。
- 风险标记：补丁大范围重构，依赖 router 侧能力，环境变量移除影响 profiling

关联脉络

- PR #2145 [docker] fix top_p mask speed issue: 同样修改了 `sglang-top_p.patch` 和 `version.txt`, 与本 PR 的 `top_p` 增强相关
- PR #2167 Always requires `rollout_top_p_token_ids` when `rollout_top_p` is not 1.0: 修改了 `rollout.py`, 涉及 `rollout` 配置, 与本 PR 的环境变量移除有间接关联