PR #24559 完整报告

sgl-project/sglang

Fix weight_checker e2e OOM on 32GB GPU + move to nightly

合并时间：2026-05-07 12:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24559

执行摘要

- 一句话：修复 weight checker 端到端测试 OOM 并移至 nightly
- 推荐动作：无需精读，但值得关注其根因分析模式：_check_tensors 的 CPU→GPU 往返在低显存环境下脆弱。该 PR 展示了如何通过调整 mem-fraction 解决显存瓶颈，对类似问题有参考价值。

功能与动机

test_weight_checker_e2e 在 32GB 5090 的 per-commit runner 上 OOM（见失败 CI 运行），#24553 通过切换到 large (H100 80GB) 规避，但该测试仍占用 per-commit 池。本 PR 根治 OOM 并将测试移出 per-commit 池。

实现拆解

1. 添加 --mem-fraction-static 0.7 参数：在 setUpClass 的 popen_launch_server 中传入 other_args=["--mem-fraction-static", "0.7"]，降低 sglang 引擎的显存占用比例，留出足够空闲显存供 _check_tensors 的 CPU→GPU 张量往返。
2. 修改测试套件注册：将 register_cuda_ci 的 suite 从 stage-b-test-1-gpu-large 改为 nightly-1-gpu，并添加 nightly=True，使测试仅每日运行而非每次提交。
3. 保留 mem-fraction 防御：即使 nightly 运行在 H100 80GB 上，仍保留 0.7 的 mem-fraction 设置，因为 _check_tensors 的显存脆弱性值得在所有环境下防范。

关键文件：

- test/registered/rl/test_weight_checker_e2e.py（模块 权重检查器；类别 test；类型 test-coverage）：唯一变更文件，包含两个关键修改：降低 mem-fraction 修复根本 OOM，以及将测试套件移至 nightly。

关键符号：未识别

关键源码片段

test/registered/rl/test_weight_checker_e2e.py

唯一变更文件，包含两个关键修改：降低 mem-fraction 修复根本 OOM，以及将测试套件移至 nightly。

```
# test/registered/rl/test_weight_checker_e2e.py
```

```
import unittest
from typing import List, Tuple

import requests
import torch

from sglang.srt.utils import MultiprocessingSerializer, kill_process_tree
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.test_utils import (
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

# 将测试从 per-commit large 套件移至 nightly 套件，节省每提交 CI 时间
register_cuda_ci(est_time=150, suite="nightly-1-gpu", nightly=True)

_MODEL_NAME = "Qwen/Qwen3-0.6B"
# We address the up half via the HF-style unfused name "up_proj.weight". sglang's
# stacked_params_mapping rewrites this to "gate_up_proj.weight" with shard_id=1,
# so the upload writes only the up half of the fused tensor. Sending the fused
# name directly hits a name.replace() collision (gate_up_proj contains up_proj),
# producing a malformed key like "gate_gate_up_proj.weight" and crashing load.
_UP_PROJ_SHAPE = (3072, 1024) # intermediate_size, hidden_size for Qwen3-0.6B


class TestWeightCheckerE2E(CustomTestCase):
    """All cases share one launched server (setUpClass).

    The reset case mutates weights to random; it is named to sort last so any
    case that needs intact weights runs first. The server is torn down right
    after, so leaving the engine in a corrupted state is harmless."""

    @classmethod
    def setUpClass(cls):
        cls.url = DEFAULT_URL_FOR_TEST
        # 降低 mem-fraction 至 0.7，确保 _check_tensors 的 CPU->GPU 往返不 OOM
        # 默认 0.88 时 sglang 占用 ~29GB（32GB GPU），仅剩 ~200MB 空闲
        # vocab-embedding 往返需要 ~600MB，导致 snapshot/reset/compare 循环 OOM
        cls.process = popen_launch_server(
            _MODEL_NAME,
            cls.url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=["--mem-fraction-static", "0.7"],
        )

    @classmethod
    def tearDownClass(cls):
```

```
        kill_process_tree(cls.process.pid)

    def _post(self, action: str) -> requests.Response:
        return requests.post(
            f"{self.url}/weights_checker", json={"action": action}, timeout=120
        )

    def _update_weights(
        self, named_tensors: List[Tuple[str, torch.Tensor]]
    ) -> requests.Response:
        return requests.post(
            f"{self.url}/update_weights_from_tensor",
            json={
                "serialized_named_tensors": [
                    MultiprocessingSerializer.serialize(named_tensors, output_str=True)
                ],
                "flush_cache": True,
            },
            timeout=120,
        )

    def test_a_snapshot_then_compare_unchanged_succeeds(self):
        resp = self._post("snapshot")
        self.assertEqual(resp.status_code, 200)
        self.assertTrue(resp.json()["success"])

        resp = self._post("compare")
```

评论区精华

无 review 评论。作者在 PR body 中详细描述了 OOM 根因分析：默认 mem-fraction-static=0.88 时 sglang 占用 ~29.43GB（32GB 总量），仅剩 ~207MB 空闲，vocab-embedding 往返需要 ~600MB 导致 OOM。通过两次 rerun-test 验证：第一次在 32GB 5090 上验证 mem-fraction 0.7 修复 OOM（115s 通过），第二次在 H100 80GB nightly runner 上验证测试通过（190s）。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：降低 mem-fraction-static 可能影响大模型加载，但仅在测试中修改，不影响生产环境。
  2. 测试覆盖风险：移至 nightly 后，per-commit CI 不再运行该测试，可能延迟发现自动化问题，但 unit 测试（test_weight_checker.py）仍在 per-commit 中。
  3. 性能影响：无。

- 影响：影响范围小，仅限于 weight checker 的端到端测试。正面影响： 1) 解除 32GB GPU 上的 OOM 阻塞； 2) 释放 per-commit CI 资源（150s 测试不再每次提交运行）。
  - 风险标记：测试覆盖迁移至 nightly，mem-fraction 仅测试环境调整

关联脉络

- PR #24553 [Misc] Fix breaking weight checker test: 前一个 PR 通过将测试移至 large 套件暂时规避 OOM，本 PR 根治 OOM 并移至 nightly。
- PR #24536 Add unit and end-to-end tests for weight checker: 为 weight checker 添加了端到端测试，本 PR 修复其中 OOM 问题。
- PR #24537 Support getting checksums in weight checker: 为 weight checker 添加 checksum 功能，测试中涉及相关逻辑。
- PR #24538 Refactor buffer patterns in weight checker: 重构了 weight checker 的缓冲区模式，测试覆盖其行为。