PR #50766 完整报告

vllm-project/vllm

[Bugfix] serving_llama70B_tp4 benchmark was silently running at tensor_parallel_size=1

合并时间：2026-08-03 10:26

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50766

执行摘要

PR #50766 为 serving_llama70B_tp4_random_128_128 基准条目显式补上 tensor_parallel_size: 4，纠正了该测试因 merge_serving_tests_stream 的 defaults 合并机制而长期静默以 TP1 运行的问题。改动仅 1 行、无生产代码变化，但直接决定了 perf.vllm.ai 上该基准历史数据的可信度与未来口径。

功能与动机

PR body 说明：构建 AMD 特定变体配置（serving-tests-rocm.json，镜像自 serving-tests.json）时，作者通过 run-performance-benchmarks.sh 中真实的 merge_serving_tests_stream 合并逻辑验证复制过来的条目，发现 serving_llama70B_tp4_random_128_128 的 server_parameters 从未设置 tensor_parallel_size，静默继承了文件 defaults 块中的 tensor_parallel_size: 1。该测试名以及固定 8192 max_num_batched_tokens、async_scheduling 参数都是为 4 GPU 运行准备的，实际 TP 度却默认为 1，导致它一直在持续基准上以 tp1 而非 tp4 运行。

实现拆解

1. 问题定位：在 AMD 变体配置验证中，用 merge_serving_tests_stream 合并默认值后逐个检查条目，发现 serving_llama70B_tp4_random_128_128 的合并结果里 tensor_parallel_size 为 1，而非其名称暗示的 4。
2. 根因确认：合并逻辑按 defaults → 每测试覆盖补全字段，缺省字段静默继承 defaults；该条目的 8192 max_num_batched_tokens 与 async_scheduling 均为 4 卡设计，却从未声明 TP，属于配置遗漏而非有意为之。
3. 修复变更：在 .buildkite/performance-benchmarks/tests/serving-tests.json 的该条目 server_parameters 中加一行 tensor_parallel_size: 4，共 +1/-0，无其他文件改动。
4. 验证方式：用真实合并逻辑跑 before/after 对比（jq 输出 1 → 4）；确认 JSON 语法合法；7 个测试名合并无回归；pre-commit 对文件的 JSON 检查通过。提交历史中有一次 merge main，保证分支同步、无冲突。
5. 配套情况：无代码、单测、schema 或部署配套改动，属于纯基准配置修复。

.buildkite/performance-benchmarks/tests/serving-tests.json

唯一变更文件。为 serving_llama70B_tp4_random_128_128 的 server_parameters 显式补上 tensor_parallel_size: 4，修复其静默继承 defaults 的 TP1 而长期以错误并行度运行的问题，直接影响 perf.vllm.ai 基准数据正确性。

```
{
  "test_name": "serving_llama70B_tp4_random_128_128",
  "server_parameters": {
    "model": "meta-llama/Llama-3.3-70B-Instruct",
    // 本次修复：显式声明 TP 为 4，避免 merge_serving_tests_stream
    // 合并 defaults 时静默继承 tensor_parallel_size: 1
    "tensor_parallel_size": 4,
    "async_scheduling": "",
    "no_enable_prefix_caching": "",
    // 8192 的 batch 上限与 async_scheduling 均为 4 卡设计，
    // TP1 运行时既不匹配也无代表性，数据无可比性
    "max_num_batched_tokens": 8192
  },
  "client_parameters": {
    "model": "meta-llama/Llama-3.3-70B-Instruct",
    "dataset_name": "random",
    "random-input-len": 128,
    "random-output-len": 128
  }
}
```

评论区精华

- 作者 wjabbour 在行内评论说明缺陷来源："This fixes a missing TP arg for this Llama test, originally introduced in this PR（#43262），same issue that Gemini caught"。即该遗漏由 #43262 引入，Gemini 也独立发现过同一问题，本次是在 AMD 配置验证中经真实合并逻辑暴露。
- bot 自动评论提示 fork 来源的自动化审查被禁用，需维护者手动触发；人工 reviewer bigPYJ1151 直接批准，无实质争议，PR 已合并。

风险与影响

- 历史数据污染：修复前该测试在 perf.vllm.ai 上长期以 TP1 运行，过去积累的 serving_llama70B TP4 数据不能当作 4 卡基线使用，需要标注或废弃。
- 同类遗漏的系统性风险：defaults 合并语义意味着同文件其他条目也可能存在静默缺省字段，本次只修了一个条目，建议对 serving-tests.json 做一次合并后字段完整性审计，并在 run-performance-benchmarks.sh 侧增加断言。
- 资源要求：显式 TP4 后该基准必须由 4 GPU 环境承接，资源不足时会直接失败而不是静默降级；这提升了数据可信度，但需要确认基准集群满足资源。

关联脉络

- 缺陷源头 #43262：作者评论确认该测试条目及缺失的 tensor_parallel_size 由 #43262 引入，本 PR 是同一功能线上的收尾修复。
- 同文件 #46870：也修改 serving-tests.json，但只是删除相邻条目的重复 JSON 片段；PR body 明确排查后确认无内容重叠，且其新增的 check-json hook 只能校验语法，无法发现此类语义缺省——这恰好说明“校验工具覆盖不到语义正确性”才是本 PR 的深层教训。

- 从更大脉络看，该修复是作者构建 AMD 变体基准配置（serving-tests-rocm.json）时的副产物，说明“用真实合并逻辑做配置验证”的方法值得推广到所有基准条目。