

PR #37214 完整报告

sgl-project/sglang

test: re-enable DSV4-Flash W8A8 8p nightly perf cases

合并时间: 2026-08-31 17:20

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37214>

执行摘要

- 一句话: 重新启用 DSV4-Flash W8A8 8p 两个夜间性能测试用例
- 推荐动作: 值得快速浏览, 但无需精读。该 PR 展示了 SGLang 使用 `register_npu_ci` 的 `disabled` 标志作为轻量 CI 开关的模式: 临时禁用用例、问题修复后移除标志即可恢复注册。对于需要管理长时间夜间性能测试的团队, 这是一个值得复用的开关设计。

功能与动机

PR body 中明确说明: 两个 DeepSeek-V4-Flash W8A8 8p 夜间性能测试用例 (in8k/in32k) 此前因测试用例 bug 被临时禁用, 问题已解决, 应在夜间流水线中恢复运行。

实现拆解

1. 定位变更入口: 两个测试文件顶部的 `register_npu_ci` 调用, 位于 `test/registered/npu/performance/deepseek_v4_flash/` 目录下。
2. 核心变更: 删除 `register_npu_ci` 调用中的 `disabled="Testcase bug, temporarily disabled"` 参数, 并将原本多行的调用压缩为单行, 使用例恢复注册。
3. 保留既有配置: `est_time=1800` (单用例预估耗时 30 分钟)、`suite="nightly-perf-16-npu-a3"`、`nightly=True` 均保持不变; 两个文件中的 `DEEPSEEK_V4_FLASH_W8A8_8P_ENVS` 环境变量字典 (deeppep、zbal、DSV4 融合压缩器、PD-mix 部署相关) 也未改动。
4. 配套说明: 无模型 / 内核代码变更, PR body 标注精度与速度测试均为 N/A; 除两个测试文件外无其他配套改动。

关键文件:

- `test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in32k_out1k_50ms.py` (模块 性能测试; 类别 test; 类型 test-coverage): DSV4-Flash W8A8 8p in32k 夜间性能用例, 本 PR 移除其 `disabled` 标志使其恢复注册, 是变更的两个目标文件之一。
- `test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py` (模块 性能测试; 类别 test; 类型 test-coverage): DSV4-Flash W8A8 8p in8k 夜间性能用例, 与 in32k 用例同步恢复注册, 是变更的两个目标文件之一。

关键符号：未识别

关键源码片段

[test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in32k_out1k_50ms.py](#)

DSV4-Flash W8A8 8p in32k 夜间性能用例，本 PR 移除其 disabled 标志使其恢复注册，是变更的两个目标文件之一。

```
import unittest

from sglang.test.ascend.e2e.test_npu_performance_utils import (
    AISBENCHMARK_DATASET_DEFAULT,
    BENCHMARK_TOOL_DEFAULT,
    DEEPSEEK_V4_FLASH_W8A8_MTP_MODEL_PATH,
    TestNpuPerformanceTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

# 注册 Ascend NPU 夜间性能测试用例。
# 本 PR 移除了 disabled 参数（此前值为 "Testcase bug, temporarily disabled"），
# 使 DSV4-Flash W8A8 8p 的 in32k 用例重新纳入 nightly-perf-16-npu-a3 套件。
# est_time 1800 表示该用例预估耗时 30 分钟，nightly=True 表示仅在夜间流水线执行。
register_npu_ci(est_time=1800, suite="nightly-perf-16-npu-a3", nightly=True)

# DSV4-Flash 单节点 PD-mix 部署所需的环境变量，保持原有配置不变。
DEEPSEEK_V4_FLASH_W8A8_8P_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "USE_NPU_MOE_GATING_TOP_K": "1",
    # deeppep: 启用 int8 量化，并限制每 rank 最大 dispatch token 数
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "64",
    # dsV4: 启用 NPU 融合压缩器，prefill 阶段在该用例中关闭
    "SGLANG_DSV4_NPU_FUSED_COMPRESSOR": "1",
    "SGLANG_DSV4_NPU_FUSED_COMPRESSOR_PREFILL": "0",
    # 跳过 GPU 分支，强制走 NPU 侧实现
    "SGLANG_OPT_FP8_WO_A_GEMM": "0",
    "SGLANG_OPT_USE_OVERLAP_STORE_CACHE": "False",
    "FORCE_DRAFT_MODEL_NON_QUANT": "1",
}
```

[test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py](#)

DSV4-Flash W8A8 8p in8k 夜间性能用例，与 in32k 用例同步恢复注册，是变更的两个目标文件之一。

```
import unittest
```

```

from sglang.test.ascend.e2e.test_npu_performance_utils import (
    AISBENCHMARK_DATASET_DEFAULT,
    BENCHMARK_TOOL_DEFAULT,
    DEEPSEEK_V4_FLASH_W8A8_MTP_MODEL_PATH,
    TestNpuPerformanceTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

# 注册 Ascend NPU 夜间性能测试用例。
# 本 PR 移除了 disabled 参数（此前值为 "Testcase bug, temporarily disabled"），
# 使 DSV4-Flash W8A8 8p 的 in8k 用例重新纳入 nightly-perf-16-npu-a3 套件。
# 与 in32k 用例相比，本用例在 deepseek dispatch 上限与 prefill 压缩器开关上取值不同。
register_npu_ci(est_time=1800, suite="nightly-perf-16-npu-a3", nightly=True)

# DSV4-Flash 单节点 PD-mix 部署所需的环境变量，保持原有配置不变。
DEEPSEEK_V4_FLASH_W8A8_8P_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "USE_NPU_MOE_GATING_TOP_K": "1",
    # deepseek: 启用 int8 量化，每 rank 最大 dispatch token 数放宽到 128
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "128",
    # dsy4: 启用 NPU 融合压缩器，prefill 阶段在该用例中同样开启
    "SGLANG_DSV4_NPU_FUSED_COMPRESSOR": "1",
    "SGLANG_DSV4_NPU_FUSED_COMPRESSOR_PREFILL": "1",
    # 跳过 GPU 分支，强制走 NPU 侧实现
    "SGLANG_OPT_FP8_WO_A_GEMM": "0",
    "SGLANG_OPT_USE_OVERLAP_STORE_CACHE": "False",
    "FORCE_DRAFT_MODEL_NON_QUANT": "1",
}

```

评论区精华

本 PR 没有任何 review 评论，也没有实质讨论线程。仅由 sglang-npu-bot 自动批准（APPROVED），提交记录为单 commit、无返工。整体属于机械式的测试标志恢复操作。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 夜间流水线资源恢复：两个用例各自 est_time=1800 秒，重新启用后会恢复占用 nightly-perf-16-npu-a3 套件的运行时长，需要确保夜间资源容量足够。
2. 回归验证证据不足：PR body 声称用例 bug 已解决，但未附上修复 commit 链接或复跑通过的结果；若问题未彻底修复，重新启用会导致夜间流水线再次失败。
3. 无运行时风险：变更只涉及测试注册参数，不影响推理路径、性能和精度。- 影响：对 CI 体系：恢复两个 DSV4-Flash W8A8 8p 在 NPU 单机 8 卡场景下的夜间性能回归覆盖（in8k/in32k 输入长度各一），有助于及时发现该模型组合的性能退化。对用户与系统

：无任何运行时影响。对团队：需要确认夜间环境的稳定性；由于 labels 含 deepseek和 npu，相关维护者需留意后续 nightly 报告。 - 风险标记：夜间流水线资源占用，缺少复跑验证

关联脉络

- 暂无明显关联 PR