PR #27335 完整报告

sgl-project/sglang

6-5 nightly failed test case fix

合并时间：2026-06-05 11:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27335

执行摘要

- 一句话：修复 Ascend NPU 夜间测试超时和 warning 问题
- 推荐动作：该 PR 为常规维护性修复，无深度技术洞察，可快速合并。

功能与动机

夜间测试（nightly）中，test_npu_deepep_auto_deepseek_v3_2_w8a8.py 因权重加载超时而失败，需延长 watchdog 超时时间；test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py 和 test_npu_deepep_low_latency_qwen3_480b.py 因 transformers 版本升级输出大量兼容性警告，需压制日志以避免干扰测试输出。

实现拆解

1. 在 test_npu_deepep_auto_deepseek_v3_2_w8a8.py 的 other_args 列表末尾添加 --watchdog-timeout 和 900 两个元素，使服务端在权重加载时有更长的等待时间。
2. 在 test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py 的 env 字典中添加 "TRANSFORMERS_VERBOSITY": "error"，将 transformers 日志级别设为 error，抑制兼容性警告。
3. 在 test_npu_deepep_low_latency_qwen3_480b.py 的 env 字典中添加相同的环境变量，达到同样效果。

关键文件：

- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py（模块 测试；类别 test；类型 test-coverage）：添加 TRANSFORMERS_VERBOSITY=error 环境变量以屏蔽 transformers 兼容性警告。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_qwen3_480b.py（模块 测试；类别 test；类型 test-coverage）：添加 TRANSFORMERS_VERBOSITY=error 环境变量以屏蔽 transformers 兼容性警告。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_deepseek_v3_2_w8a8.py（模块 测试；类别 test；类型 test-coverage）：添加 --watchdog-timeout 900 参数以解决权重加载超时问题。

关键符号：未识别

关键源码片段

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py

添加 TRANSFORMERS_VERBOSITY=error 环境变量以屏蔽 transformers 兼容性警告。

```
# 修改前：env 字典未设置 TRANSFORMERS_VERBOSITY，导致 transformers
升级后输出大量兼容性警告
# 修改后：env 字典新增以下条目
env={
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "128",
    "HCCL_BUFFSIZE": "2048",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "TASK_QUEUE_ENABLE": "0",
    "TRANSFORMERS_VERBOSITY": "error", # 新增：将 transformers 日志级别设为
    error，抑制兼容性警告
    **os.environ,
}
```

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_qwen3_480b.py

添加 TRANSFORMERS_VERBOSITY=error 环境变量以屏蔽 transformers 兼容性警告。

```
# 修改前：env 字典未设置 TRANSFORMERS_VERBOSITY，导致 transformers
升级后输出大量兼容性警告
# 修改后：env 字典新增以下条目
env={
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "600",
    "HCCL_BUFFSIZE": "2100",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "TRANSFORMERS_VERBOSITY": "error", # 新增：抑制 transformers 兼容性警告
    **os.environ,
}
```

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_deepseek_v3_2_w8a8.py

添加 --watchdog-timeout 900 参数以解决权重加载超时问题。

```
# 修改前：other_args 列表未设置 --watchdog-timeout，服务端可能因权重加载超时被 watchdog
杀死
# 修改后：在列表末尾添加 --watchdog-timeout 和 900 两个元素
other_args = [
    "--trust-remote-code",
    "--tp-size", "16",
    "--quantization", "modelslim",
```

```
        "--moe-a2a-backend", "deepep",
        "--deepep-mode", "auto",
        "--mem-fraction-static", 0.82,
        "--disable-cuda-graph",
        "--disable-radix-cache",
        "--context-length", 40960,
        "--max-prefill-tokens", 40960,
        "--max-total-tokens", 40960,
        "--watchdog-timeout", 900, # 新增：将 watchdog 超时设为 900
        秒，避免权重加载耗时过长被杀死
    ]
```

评论区精华

该 PR 无 review 评论，变更简单明确。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改测试配置，不涉及生产代码。--watchdog-timeout 900 延长了超时时间，不会影响功能正确性；TRANSFORMERS_VERBOSITY=error 屏蔽的只是兼容性警告，不影响 transformers 实际行为。
- 影响：影响范围仅限于三个 Ascend NPU 测试用例，使它们能稳定通过夜间测试。对系统其他部分无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR