PR #31952 完整报告

sgl-project/sglang

Add pr tests

合并时间：2026-08-01 15:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31952

执行摘要

- 一句话：NPU 测试目录改名并新增 13 个 PR CI 用例
- 推荐动作：建议 NPU 平台与 CI 团队成员精读：本 PR 是理解 SGLang NPU 测试体系（register_npu_ci 注册机制、性能 / 精度用例拆分规范、PR 与 nightly 参数差异化）的最佳入口。普通功能开发只需关注 pr-test-npu.yml 矩阵变化，避免在 test/registered/ascend 遗留引用。值得借鉴的设计：用 GITHUB_EVENT_NAME 区分 CI 场景参数、用 max_attempts 缓解共享集群性能抖动。

功能与动机

PR body 明确说明动机："Supplement test cases covering risk scenarios introduced by community contributions, enabling early interception of issues." 即把社区贡献引入的风险场景固化为 PR 阶段可执行的测试用例，在合入前尽早拦截性能或精度回归。目录重命名（ascend → npu）服务于命名一致性：代码中长期混用 --device npu 与 attention-backend ascend，测试目录统一为 npu 与运行时设备名对齐，降低新贡献者的认知成本。

实现拆解

1. 目录重命名与全仓路径迁移：将 test/registered/ascend 整体重命名为 test/registered/npu，迁动 151 个测试文件；同步更新 .github/workflows/pr-test-npu.yml、nightly 工作流、docs_new 文档与源码内引用，避免路径失效。
2. 混合测试拆分：将 7 个性能与精度混写的文件拆成独立文件——性能文件去掉数据集后缀（_gpqa、_aime25、_aime26），精度文件保留后缀并归入 accuracy 目录；删除已无用的空 setUpClass / tearDownClass 方法。受影响模型包括 qwen3-8b、qwen3_30b_a3b、qwen3_6_35b_a3b、qwen3_32b、qwen3_next_80b、minimax_m2_5、glm5_1、kimi_k2_6。
3. CI 矩阵扩展：pr-test-npu.yml 的 pr-single-node-tests 矩阵新增 13 个用例；nightly 工作流同步区分 perf 与 accuracy 条目归属，防止数据集后缀导致匹配失败。
4. 新增高性能用例：为 kimi_k2_6、qwen3_235b、qwen3_5_397b、qwen3_next_80b、qwen3_30b、qwen3-8b 等模型补充性能回归，每个用例携带专用 ENVS（FIA、ZBAL、DeepEP、HCCL 等）与 OTHER_ARGS（TP / DP、quant、EAGLE3 / NEXTN、PD 分离等）。

5. 场景化细节调整：qwen3_vl_8b_thinking 的 MMMU 用例通过 GITHUB_EVENT_NAME 区分 PR（limit=5、accuracy=0.64）与 nightly（limit=100000、accuracy=0.7011）参数；autoround_moe 超时提升至 3600 秒；拆出的 6 个性能与 12 个精度用例补齐 disabled 标记，避免误入常规 PR CI。

关键文件：

- .github/workflows/pr-test-npu.yml（模块 CI 配置；类别 infra；类型 configuration；符号 pr-single-node-tests）：PR CI 矩阵核心入口，新增 13 个测试用例注册，直接影响 CI 资源消耗与回归覆盖范围。
- test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_8p_in3k5_out1k5_20ms.py（模块 性能测试；类别 test；类型 test-coverage；符号 TestKimiK25W4A8, test_kimi_k2_6_w4a8）：新增 Kimi K2.6 W4A8 性能回归用例，覆盖 EAGLE3 投机解码、多模态与 DP attention 组合场景，是本 PR 三个指定新增性能模型之一。
- test/registered/npu/performance/qwen3_5_397b/test_npu_qwen3_5_397b_w4a8_8p_in3k5_out1k5_50ms.py（模块 性能测试；类别 test；类型 test-coverage；符号 TestNPUQwen3_5_397B_A17B_3K5_1K5_50ms, test_npu_qwen3_5_397b_a17b_3k5_1k5）：新增 Qwen3.5-397B（当前最大规模 W4A8 + 多模态 + Mamba 混合架构）性能用例，配置 ZBAL 与 FIA 等集群级通信优化，是本 PR 指定新增的三个性能模型之一。
- test/registered/npu/performance/glm5_1/test_npu_glm5_1_w4a8_1p1d_32p_in64k_out1k_50ms.py（模块 性能测试；类别 test；类型 test-coverage；符号 TestNPUGLM5_1_W4A8_PD_SEP_In3k5_Out1k5, test_npu_glm5_1_w4a8_pd_sep_in3k5_out1k5）：新增 GLM-5.1 W4A8 多节点 PD 分离长上下文性能用例，覆盖 32 卡 prefill/decode 分离部署场景，是本 PR 新增用例中部署形态最复杂的一个。
- test/registered/npu/performance/qwen3_6_35b_a3b/test_npu_qwen3_6_35b_a3b_1p_in64k_out1k_prefix90_50ms.py（模块 性能测试；类别 test；类型 rename-or-move；符号 TestNPUQwen3_6_35BA3B_1P_In64k_Out1k_Prefix90_50ms, test_npu_qwen3_6_35b_a3b_1p_in64k_out1k_prefix90_50ms）：混合测试文件拆分的典型样例：从原 AIME2026 精度 + 性能混写文件中剥离精度类，仅保留性能类，并删除惰性 setUpClass/tearDownClass。
- .github/CODEOWNERS（模块 仓库配置；类别 infra；类型 configuration）：目录重命名后 owner 规则需要同步，reviewer cherryblo 明确指出 Ascend 需批量替换为 NPU，否则路径匹配失效。

关键符号：run_throughput, run_accuracy, test_npu_glm5_1_w4a8_pd_sep_in3k5_out1k5, test_npu_qwen3_5_397b_a17b_3k5_1k5, test_kimi_k2_6_w4a8, test_qwen3_235b, test_npu_qwen3_6_35b_a3b_1p_in64k_out1k_prefix90_50ms, TestNpuPerformanceTestCaseBase, TestNpuAccuracyTestCaseBase, register_npu_ci

关键源码片段

test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_8p_in3k5_out1k5_20ms.py

新增 Kimi K2.6 W4A8 性能回归用例，覆盖 EAGLE3 投机解码、多模态与 DP attention 组合场景，是本 PR 三个指定新增性能模型之一。

```
import unittest

from sglang.test.ascend.e2e.test_npu_multi_node_utils import NIC_NAME
from sglang.test.ascend.e2e.test_npu_performance_utils import (
    AISBENCHMARK_DATASET_DEFAULT,
    BENCHMARK_TOOL_DEFAULT,
    KIMI_K2_6_EAGLE3_MODEL_PATH,
    KIMI_K2_6_W4A8_MODEL_PATH,
    TestNpuPerformanceTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

# register_npu_ci 把用例登记到 NPU CI 体系：est_time 给调度器预估时长，
# suite 决定归入哪套集群矩阵，nightly=True 表示不阻塞 PR 合并，
# disabled 标记说明实际由 NPU 性能专用工作流触发执行
register_npu_ci(
    est_time=1800,
    suite="full-16-npu-a3",
    nightly=True,
    disabled="Currently it is executed by the npu performance workflow.",
)

# 大模型环境变量组合：MLAPO 注意力优化 + 多流并行 + DeepEP INT8 量化
# 通信，EAGLE3 投机解码依赖 SGLANG_ENABLE_SPEC_V2
KIMI_K2_6_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "SGLANG_SET_CPU_AFFINITY": "1",
    "HCCL_SOCKET_IFNAME": NIC_NAME,
    "GLOO_SOCKET_IFNAME": NIC_NAME,
    "STREAMS_PER_DEVICE": "32",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "600",
    "SGLANG_ENABLE_SPEC_V2": "1",
    "SGLANG_ENABLE_OVERLAP_PLAN_STREAM": "1",
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "96",
    "DEEPEP_HCCL_BUFFSIZE": "1200",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "SGLANG_NPU_USE_MLAPO": "1",
    "SGLANG_NPU_USE_MULTI_STREAM": "1",
}


class TestKimiK25W4A8(TestNpuPerformanceTestCaseBase):
    """Kimi K2.6 W4A8 在 16 卡 NPU 上的吞吐回归测试，目标 TPOT 20 ms"""

    benchmark_tool = BENCHMARK_TOOL_DEFAULT
```

```
    dataset_type = AISBENCHMARK_DATASET_DEFAULT
    # 共享集群上的性能波动允许重试 5 次，减少误报
    max_attempts = 5
    model = KIMI_K2_6_W4A8_MODEL_PATH
    other_args = KIMI_K2_6_OTHER_ARGS
    envs = KIMI_K2_6_ENVS
    dataset_name = "random"
    max_concurrency = 64
    num_prompts = 256
    input_len = 3500
    output_len = 1500
    tpot = 20
    output_token_throughput = 1900

    def test_kimi_k2_6_w4a8(self):
        # 基类 run_throughput 内部执行 benchmark 并断言
        # tpot 与 output_token_throughput 两个吞吐指标
        self.run_throughput()


if __name__ == "__main__":
    unittest.main()
```

test/registered/npu/performance/qwen3_5_397b/test_npu_qwen3_5_397b_w4a8_8p_in3k5_out1k5_50ms.py

新增 Qwen3.5-397B（当前最大规模 W4A8 + 多模态 + Mamba 混合架构）性能用例，配置 ZBAL 与 FIA 等集群级通信优化，是本 PR 指定新增的三个性能模型之一。

```
import unittest

from sglang.test.ascend.e2e.test_npu_performance_utils import (
    AISBENCHMARK_DATASET_DEFAULT,
    BENCHMARK_TOOL_DEFAULT,
    QWEN3_5_397B_W4A8_MODEL_PATH,
    TestNpuPerformanceTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

# 归入 nightly-16-npu-a3 套件：16 卡 NPU A3 集群的夜间回归矩阵
register_npu_ci(
    est_time=3600,
    suite="nightly-16-npu-a3",
    nightly=True,
    disabled="performance testcase",
)

# Qwen3.5-397B 是当前最大规模的 W4A8 + 多模态 + Mamba 混合架构模型，
# 这套环境变量针对 16 卡集群调优：
# - ZBAL 系列：Zero-Bubble AllReduce 的本地显存与图模式配置
```

```
# - ASCEND_USE_FIA 与 HCCL_OP_EXPANSION_MODE=AIV：集合通信加速
# - DEEPEP_NORMAL_LONG_SEQ_PER_ROUND_TOKENS：长序列按 3584 token 分片
QWEN3_5_397B_A17B_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "SGLANG_SET_CPU_AFFINITY": "1",
    "STREAMS_PER_DEVICE": "32",
    "ASCEND_USE_FIA": "1",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "128",
    "HCCL_BUFFSIZE": "0",
    "DEEPEP_NORMAL_LONG_SEQ_ROUND": "6",
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "GDN_ATTN_BACKEND_TRITON": "1",
    "DEEPEP_NORMAL_LONG_SEQ_PER_ROUND_TOKENS": "3584",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "HCCL_SOCKET_IFNAME": "lo",
    "GLOO_SOCKET_IFNAME": "lo",
    "SGLANG_ENABLE_SPEC_V2": "1",
    "SGLANG_ENABLE_OVERLAP_PLAN_STREAM": "1",
    "SGLANG_ZBAL_LOCAL_MEM_SIZE": "59648",
    "SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK": "0",
    "SGLANG_ZBAL_BOOTSTRAP_URL": "tcp://127.0.0.1:24669",
    "ZBAL_NPU_ALLOC_CONF": "use_vmm_for_static_memory:True",
    "ZBAL_ENABLE_GRAPH": "1",
}


class TestNPUQwen3_5_397B_A17B_3K5_1K5_50ms(TestNpuPerformanceTestCaseBase):
    """Qwen3.5-397B-A17B 在 16 卡 NPU 上的吞吐回归，目标 TPOT 50 ms"""

    benchmark_tool = BENCHMARK_TOOL_DEFAULT
    dataset_type = AISBENCHMARK_DATASET_DEFAULT
    model = QWEN3_5_397B_W4A8_MODEL_PATH
    # 启动参数配合上面的 ENVS：TP 16 + DP 8 + DP attention，
    # NEXTN 投机解码 + DeepEP MoE 通信 + 多模态注意力后端
    other_args = QWEN3_5_397B_A17B_3K5_1K5_OTHER_ARGS
    envs = QWEN3_5_397B_A17B_ENVS
    dataset_name = "random"
    warmup_requests = 16
    max_concurrency = 432
    num_prompts = 432
    input_len = 3500
    output_len = 1500
    tpot = 50
    output_token_throughput = 5415
    request_rate = float("inf")
    temperature = 0.6
    top_p = 0.95

    def test_npu_qwen3_5_397b_a17b_3k5_1k5(self):
```

```
        # 执行吞吐基准，基类断言实际吞吐不低于 5415 tokens/s
        self.run_throughput()


if __name__ == "__main__":
    unittest.main()
```

test/registered/npu/performance/glm5_1/test_npu_glm5_1_w4a8_1p1d_32p_in64k_out1k_50ms.py

新增 GLM-5.1 W4A8 多节点 PD 分离长上下文性能用例，覆盖 32 卡 prefill/decode 分离部署场景，是本 PR 新增用例中部署形态最复杂的一个。

```
import unittest

from sglang.test.ascend.e2e.test_npu_multi_node_utils import NIC_NAME
from sglang.test.ascend.e2e.test_npu_performance_utils import (
    AISBENCHMARK_DATASET_DEFAULT,
    BENCHMARK_TOOL_DEFAULT,
    GLM_5_1_W4A8_MODEL_PATH,
    TestNpuPerfMultiNodePdSepTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

register_npu_ci(
    est_time=3600,
    suite="",
    nightly=True,
    disabled="performance testcase",
)

# PD 分离部署下 prefill 与 decode 节点使用不同环境：
# prefill 侧重长序列分片与 INT8 量化传输，
# decode 侧重 overlap reflow 与多流 plan stream 并行
GLM_5_1_PD_SEP_PREFILL_ENVS = {
    "SGLANG_SET_CPU_AFFINITY": "1",
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "1200",
    "SGLANG_DISAGGREGATION_WAITING_TIMEOUT": "1200",
    "DEEPEP_HCCL_BUFFSIZE": "1200",
    "DEEPEP_NORMAL_LONG_SEQ_ROUND": "72",
    "DEEPEP_NORMAL_LONG_SEQ_PER_ROUND_TOKENS": "1024",
    "DEEPEP_NORMAL_COMBINE_ENABLE_LONG_SEQ": "1",
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "TASK_QUEUE_ENABLE": "2",
    "ENABLE_PROFILING": "0",
    "HCCL_SOCKET_IFNAME": NIC_NAME,
    "GLOO_SOCKET_IFNAME": NIC_NAME,
}
```

```
GLM_5_1_PD_SEP_DECODE_ENVS = {
    "SGLANG_SET_CPU_AFFINITY": "1",
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "1200",
    "SGLANG_DISAGGREGATION_WAITING_TIMEOUT": "1200",
    "SGLANG_SPEC_ENABLE_OVERLAP_REFLOW": "1",
    "SGLANG_ENABLE_OVERLAP_PLAN_STREAM": "1",
    "HCCL_BUFFSIZE": "200",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "16",
    "TASK_QUEUE_ENABLE": "0",
    "HCCL_SOCKET_IFNAME": NIC_NAME,
    "GLOO_SOCKET_IFNAME": NIC_NAME,
}


class TestNPUGLM5_1_W4A8_PD_SEP_In3k5_
Out1k5(TestNpuPerfMultiNodePdSepTestCaseBase):
    """GLM-5.1 W4A8 PD 分离（2 节点 32 卡）长上下文吞吐回归测试"""

    # 通过 model_config 一次描述 prefill / decode / router 三类进程的
    # 启动参数与环境，基类负责多节点拉起与端到端基准
    model_config = {
        "model_path": GLM_5_1_W4A8_MODEL_PATH,
        "prefill_args": GLM_5_1_PD_SEP_PREFILL_ARGS,
        "decode_args": GLM_5_1_PD_SEP_DECODE_ARGS,
        "prefill_envs": GLM_5_1_PD_SEP_PREFILL_ENVS,
        "decode_envs": GLM_5_1_PD_SEP_DECODE_ENVS,
        "router_args": ["--policy", "round_robin"],
        "router_envs": {},
    }
    benchmark_tool = BENCHMARK_TOOL_DEFAULT
    dataset_type = AISBENCHMARK_DATASET_DEFAULT
    max_concurrency = 1
    num_prompts = 1
    input_len = 65536
    output_len = 1024
    tpot = 50
    output_token_throughput = 160

    def test_npu_glm5_1_w4a8_pd_sep_in3k5_out1k5(self):
        # 64k 输入单请求长上下文场景，断言端到端吞吐不低于 160 tokens/s
        self.run_throughput()


if __name__ == "__main__":
    unittest.main()
```

评论区精华

评审由 cherryblo 提出 2 个线程、sglang-npu-bot 最终 APPROVED。核心讨论包括：1）CI 并行度资源消耗——新增性能用例后 parallelism=6 的资源压力；2）CODEOWNERS 术语一致性——目录改名后 Ascend 需批量替换为 NPU。两道讨论均未在评论区出现后续回复确认，但 PR 已合并。

- 性能用例增多后 CI 并行度资源消耗 (performance): 评论区未见明确调整结论，PR 最终获批合并；后续需在实际 CI 运行中观察资源水位并动态调整并行度。
- CODEOWNERS 中 Ascend 术语批量替换 (style): reviewer 要求明确，评论中未见修改确认回复，但 PR 最终由 sglang-npu-bot 批准合并，推测已随重命名一并处理。

风险与影响

- 风险：
  1. CI 资源压力：13 个新用例单个 est_time 在 1800–3600 秒之间，叠加后 parallelism=6 可能导致 NPU CI 集群资源竞争与排队，cherryblo 已明确提出此风险，需在后续运行中动态调整并行度。
  2. 重命名遗漏风险：151 个文件迁移若在源码、文档或 CI 配置中残留 ascend 路径（如 CODEOWNERS 的 owner 规则、docs_new 内的引用），会导致路径匹配失效或 CI 无法正确触发。
  3. 性能阈值稳定性：tpot、output_token_throughput 为硬性断言，共享集群负载抖动易造成误报；部分用例已用 max_attempts=5 缓解，但未覆盖全部用例。
  4. 环境耦合与可移植性：部分用例硬编码 /root/.cache/modelscope/... 等绝对路径与集群私有环境变量（如 ZBAL_BOOTSTRAP_URL、ZBCCL 系列），换集群或改目录需同步调整。 - 影响：对用户无运行时影响。对团队而言，NPU 回归防线从 nightly 扩展到 PR 阶段，可更早暴露社区贡献引入的性能 / 精度回归；CI 运维需要持续观察资源水位并动态调整并行度；测试目录命名统一降低了新贡献者维护测试的认知成本。 - 风险标记：CI 资源消耗风险，大规模重命名遗漏风险，性能阈值稳定性风险，环境路径硬编码

关联脉络

- PR #32649 add NPU GSM8K accuracy tests for 7 models: 同属 NPU 测试体系扩展，同样修改 pr-test-npu.yml 的注册矩阵，说明 NPU 测试覆盖正按模型与任务逐批补齐。
- PR #32862 [AMD] Pin mem_fraction_static for the piecewise CUDA graph 1-GPU test on MI300: 同为异构硬件平台 CI 稳定性调优，体现各平台测试参数需要在真实集群上反复校准。
- PR #33179 [CI] Fix runtime context setup in flat logprob tests: 属于 CI 测试修复，与本 PR 共同构成 SGLang 测试基础设施持续维护的脉络。