PR #35502 完整报告

sgl-project/sglang

Add three new test cases

合并时间：2026-08-19 17:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35502

执行摘要

- 一句话：为 NPU 新增 DSV4-Flash 性能与 GLM-5.2 精度测试
- 推荐动作：不值得精读，属于纯测试与 CI 配置变更。但值得保存为模板：在 NPU 上新增模型性能 / 精度用例时，可参考其环境变量组合、多节点拓扑与启动参数配比。若需复用，建议将绝对路径与固定端口抽取为可配置项，降低维护成本。

功能与动机

PR body 明确说明动机：为 DeepSeek-V4-Flash 模型增加两个性能测试用例（16 路并行 1P1D 与 8 路并行），为 GLM-5.2 模型增加一个 GPQA 精度测试用例。这是 NPU 平台对最新模型在昇腾硬件上的回归验证补充，确保 W8A8/W4A8 量化部署与 PD 分离等新特性不出现性能或精度回退。

实现拆解

1. 新增 DeepSeek-V4-Flash W8A8 1P1D 16 卡性能测试：文件 test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms.py（+234 行）。继承 TestNpuPerfMultiNodePdSepTestCaseBase 实现 PD 分离场景，prefill 与 decode 节点分别配置环境变量与启动参数：tp-size 16、dp-size 16 并开启 --enable-dp-attention 与 --enable-dp-lm-head，MoE 通信走 deepep normal 模式，量化用 modelslim，attention-backend 用 dsv4，叠加 ZBAL（TP 内存均衡）、MTP 投机解码（SGLANG_ENABLE_SPEC_V2）与 ascend 传输后端。注册为 nightly、est_time 3600 秒且默认 disabled。
2. 新增 GLM-5.2 W4A8 双节点 GPQA 精度测试：test/registered/npu/accuracy/glm5_2/test_npu_glm_5_2_w4a8_16p_gpqa.py（+106 行）。继承 TestNpuAccuracyMultiNodePdMixTestCaseBase，配置 tp-size 32、2 节点，启用 NEXTN 投机解码（3 步、topk 1、4 个 draft token）与 glm45/glm47 parser，量化 modelslim，精度阈值 0.912，数据集 gpqa_diamond。注意 --dp-size 相关参数被注释，说明该场景走纯 TP 而非 DP attention。
3. 更新既有 8 卡性能用例参数：test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py 从 base-c 套件移除只保留 nightly，est_time 1200→1800；环境变量新增 USE_NPU_MOE_GATING_TOP_K、SGLANG_NPU_USE_MULTI_STREAM、ZBAL 系列、DSV4 fused compressor、MTP 调试与 SGLANG_DISABLE_DRAFT_EXTEND_GRAPH；

启动参数同步调整（mem-fraction-static 0.6→0.7、chunked-prefill-size -1→131072、kv-cache-dtype auto、新增 --skip-server-warmup、--ep-size 16），并固定 dataset_path 指向本地缓存数据集。

4. 夜间 CI 编排：.github/workflows/nightly-test-npu.yml 新增两个多节点任务条目，deepseek_v4_flash 1P1D 以 prefill 1 + decode 1 + router 1 的节点组合运行 perf，glm_5_2 以 2 节点运行 accuracy，均复用 nightly-test-npu-e2e-multi-node.yml。
5. 配套常量与时长调整：python/sglang/test/ascend/e2e/test_npu_performance_utils.py 新增 GLM_5_2_W4A8_MODEL_PATH 指向 ModelScope 缓存的 GLM-5.2-w4a8 权重；in32k 用例 est_time 从 3600 下调到 1800。

关键文件：

- test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms.py（模块 性能测试；类别 test；类型 test-coverage；符号 TestNPUDeepSeekV4FlashW8A81P1D16PIn8kOut1k50ms, test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms）：新增的 DeepSeek-V4-Flash W8A8 1P1D 16 卡 PD 分离性能测试，是本 PR 最核心的用例，覆盖 prefill/decode 两套完整部署参数。
- test/registered/npu/accuracy/glm5_2/test_npu_glm_5_2_w4a8_16p_gpqa.py（模块 精度测试；类别 test；类型 test-coverage；符号 TestNPUGLM_5_2_W4A8_16P_GPQA, test_npu_glm_5_2_w4a8_16p_gpqa）：新增 GLM-5.2 W4A8 双节点 GPQA 精度测试，精度阈值 0.912，验证 NEXTN 投机解码下的模型准确性。
- 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）：既有 8 卡性能用例参数全面刷新：移除 base-c 注册、叠加 ZBAL 与 DSV4 fused compressor 环境变量、调整 mem-fraction 与 chunked-prefill 等启动参数。
- .github/workflows/nightly-test-npu.yml（模块 CI 编排；类别 infra；类型 infrastructure）：把新增用例接入夜间多节点 CI 编排，分别是 1P1D 分离部署的 deepseek_v4_flash perf 任务与 2 节点 glm_5_2 accuracy 任务。
- python/sglang/test/ascend/e2e/test_npu_performance_utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 GLM_5_2_W4A8_MODEL_PATH）：为 GLM-5.2 精度测试提供模型路径常量 GLM_5_2_W4A8_MODEL_PATH，是新增用例的依赖项。
- 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）：将 est_time 从 3600 秒下调到 1800 秒，配合夜间资源调度调整。

关键符号：test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms, TestNPUDeepSeekV4FlashW8A81P1D16PIn8kOut1k50ms, test_npu_glm_5_2_w4a8_16p_gpqa, TestNPUGLM_5_2_W4A8_16P_GPQA, run_accuracy, register_npu_ci

关键源码片段

test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms.py

新增的 DeepSeek-V4-Flash W8A8 1P1D 16 卡 PD 分离性能测试，是本 PR 最核心的用例，覆盖 prefill/decode 两套完整部署参数。

```
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,
    TestNpuPerfMultiNodePdSepTestCaseBase,
)
from sglang.test.ci.ci_register import register_npu_ci

# 注册为 nightly 性能测试，预计耗时 3600 秒，当前默认禁用
register_npu_ci(
    est_time=3600,
    suite="",
    nightly=True,
    disabled="performance testcase",
)

# Prefill 节点环境变量：开启 ZBAL 均衡、跳过 GPU 分支优化、启用 PD 分离与 MTP
DEEPSEEK_V4_FLASH_W8A8_1P1D_PREFILL_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "INF_NAN_MODE_FORCE_DISABLE": "1",
    # 本机回环通信，避免多网卡干扰
    "HCCL_SOCKET_IFNAME": "lo",
    "GLOO_SOCKET_IFNAME": "lo",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    # 跳过 GPU 分支，统一走 NPU 实现
    "SGLANG_OPT_FP8_WO_A_GEMM": "0",
    "SGLANG_OPT_USE_OVERLAP_STORE_CACHE": "False",
    "FORCE_DRAFT_MODEL_NON_QUANT": "1",
    "SGLANG_DSV4_FP4_EXPERTS": "False",
    # ZBAL 防止 TP 内存失衡，固定本地内存预算
    "HCCL_BUFFSIZE": "8",
    "SGLANG_ZBAL_LOCAL_MEM_SIZE": "62084",
    "SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK": "0",
    "ZBAL_NPU_ALLOC_CONF": "use_vmm_for_static_memory:True",
    "SGLANG_ZBAL_BOOTSTRAP_URL": "tcp://127.0.0.1:24669",
    "ZBAL_ENABLE_GRAPH": "1",
    # PD 分离：等待 decode 节点 bootstrap 的超时时间
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "60",
    # MTP 投机解码
```

```
    "SGLANG_ENABLE_SPEC_V2": "1",
    "SGLANG_ENABLE_OVERLAP_PLAN_STREAM": "1",
}

# Prefill 节点启动参数：16 卡 TP + 16 路 DP attention，deepep MoE 通信
DEEPSEEK_V4_FLASH_W8A8_1P1D_PREFILL_ARGS = [
    "--page-size", 128,
    "--tp-size", 16,
    "--trust-remote-code",
    "--device", "npu",
    "--attention-backend", "dsv4",
    "--watchdog-timeout", 9000,
    "--disaggregation-mode", "prefill",
    "--disaggregation-transfer-backend", "ascend",
    "--mem-fraction-static", 0.62,
    "--dp-size", 16,
    "--enable-dp-attention",
    "--moe-a2a-backend", "deepep",
    "--deepep-mode", "normal",
    "--quantization", "modelslim",
    "--enable-dp-lm-head",
    "--kv-cache-dtype", "bfloat16",
    "--disable-cuda-graph",
    "--load-balance-method", "round_robin",
    "--ep-dispatch-algorithm", "static",
]
```

test/registered/npu/accuracy/glm5_2/test_npu_glm_5_2_w4a8_16p_gpqa.py

新增 GLM-5.2 W4A8 双节点 GPQA 精度测试，精度阈值 0.912，验证 NEXTN 投机解码下的模型准确性。

```
import unittest

from sglang.test.ascend.e2e.test_npu_accuracy_utils import (
    BENCHMARK_TOOL_DEFAULT,
    TestNpuAccuracyMultiNodePdMixTestCaseBase,
)
from sglang.test.ascend.e2e.test_npu_multi_node_utils import NIC_NAME
from sglang.test.ascend.e2e.test_npu_performance_utils import (
    GLM_5_2_W4A8_MODEL_PATH,
)
from sglang.test.ci.ci_register import register_npu_ci

# 注册为 nightly 精度测试，预计耗时 3600 秒，当前默认禁用
register_npu_ci(
    est_time=3600,
    suite="",
    nightly=True,
    disabled="accuracy testcase",
```

```
)

# 双节点环境变量：用 NIC_NAME 保证 HCCL / GLOO 走实际网卡而非回环
GLM_5_2_W4A8_16P_TWO_NODE_ENVS = {
    "SGLANG_SET_CPU_AFFINITY": "1",
    "STREAMS_PER_DEVICE": "32",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "600",
    "SGLANG_ENABLE_SPEC_V2": "1",
    "SGLANG_ENABLE_OVERLAP_PLAN_STREAM": "1",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "32",
    "TRANSFORMERS_VERBOSITY": "error",
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "DEEPEP_HCCL_BUFFSIZE": "2500",
    "HCCL_SOCKET_IFNAME": NIC_NAME,
    "GLOO_SOCKET_IFNAME": NIC_NAME,
}

# 启动参数：2 节点 tp-size 32、NEXTN 投机解码、modelslim 量化、GLM 专用 parser
GLM_5_2_W4A8_16P_TWO_NODE_OTHER_ARGS = [
    "--attention-backend", "ascend",
    "--device", "npu",
    "--tp-size", 32,
    "--nnodes", 2,
    "--chunked-prefill-size", 65536,
    "--max-prefill-tokens", 280000,
    "--trust-remote-code",
    "--mem-fraction-static", 0.70,
    "--served-model-name", "glm-5",
    "--cuda-graph-max-bs", 32,
    "--max-running-requests", 32,
    "--quantization", "modelslim",
    "--moe-a2a-backend", "deepep",
    "--deepep-mode", "auto",
    "--load-balance-method", "round_robin",
    "--speculative-algorithm", "NEXTN",
    "--speculative-num-steps", 3,
    "--speculative-eagle-topk", 1,
    "--speculative-num-draft-tokens", 4,
    "--reasoning-parser", "glm45",
    "--tool-call-parser", "glm47",
]

GLM_5_2_W4A8_16P_TWO_NODE_MODEL_CONFIG = {
    "model_path": GLM_5_2_W4A8_MODEL_PATH,
    "other_args": GLM_5_2_W4A8_16P_TWO_NODE_OTHER_ARGS,
    "node_envs": GLM_5_2_W4A8_16P_TWO_NODE_ENVS,
}

class TestNPUGLM_5_2_W4A8_16P_GPQA(TestNpuAccuracyMultiNodePdMixTestCaseBase):
```

```
    """NPU 上 GLM-5.2-w4a8 双节点 16p 在 gpqa_diamond 上的精度测试"""

    benchmark_tool = BENCHMARK_TOOL_DEFAULT
    model_config = GLM_5_2_W4A8_16P_TWO_NODE_MODEL_CONFIG
    # 精度阈值 0.912，低于该值即判定回归
    accuracy = 0.912
    datasets = ["gpqa_diamond"]
    eval_batch_size = 32
    generation_config = {"max_tokens": 65536, "temperature": 1.0}

    def test_npu_glm_5_2_w4a8_16p_gpqa(self):
        """运行 GPQA 精度测试"""
        self.run_accuracy()


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

评论区精华

该 PR 没有任何 review 评论或讨论线程（comments_count=0、review_comments_count=0），最终由 sglang-npu-bot 直接合入。隐含共识是：性能 / 精度用例默认标记 nightly 且 disabled，避免常驻 CI 资源消耗；多节点部署参数（网卡选择 NIC_NAME、ZBAL bootstrap URL、HCCL buffer size）集中维护在测试基类与用例级常量中，便于后续复制为模板。

- 暂无高价值评论线程

风险与影响

- 风险：模型与数据集路径全部指向 /root/.cache/modelscope/hub/... 的绝对路径，一旦缓存不存在或路径变动用例即失败，可维护性差；ZBAL bootstrap URL 固定为 tcp://127.0.0.1:24669，多节点并发跑不同用例时存在端口冲突风险；SGLANG_ZBAL_LOCAL_MEM_SIZE 等数值依赖具体机型。新增 3 个 nightly 用例每个 est_time 达 1800-3600 秒，会显著增加 NPU 夜间 runner 占用。8P in8k 用例从 base-c 日常 CI 移除，日常 PR 覆盖该性能场景的能力被削弱。不涉及推理路径，源码回归风险低。
- 影响：对 NPU 平台 nightly 验证矩阵是正向扩展：覆盖 DeepSeek-V4-Flash W8A8 在 1P1D PD 分离与 8 卡 PD-mix 下的吞吐，以及 GLM-5.2 W4A8 的 GPQA 精度。用户侧无感知；NPU 团队新增维护负担与夜间资源开销；仓库其他模块不受影响。
- 风险标记：纯测试变更，无源码风险，依赖 ModelScope 绝对路径缓存，ZBAL 端口硬编码，新增夜间 CI 资源开销

关联脉络

- PR #33313 [AMD] DeepSeek-V4: route decode wo_a bf16 batched matmul to aiter batched_gemm_bf16: 同属 DeepSeek-V4 在 NPU/AMD 上的性能验证线，本 PR 的 DSV4-Flash 性能用例正是这类 kernel 路由优化的端到端回归保障。

- PR #33165 [AMD] DeepSeek-V4 MI355X: eliminate bpreshuffle fp8-scale relayout copy in dense w8a8 linear: 同为 DeepSeek-V4 W8A8 在 NPU/AMD 上的性能测试体系，本 PR 覆盖同模型族在昇腾的 W8A8 吞吐场景。
- PR #35467 [AMD][DI][CI] Run MI355X disagg nightly at 7AM UTC: 同为 nightly CI 编排调整，说明 NPU/AMD 夜间测试矩阵正在持续扩展，本 PR 新增的多节点任务与其互补。