PR #35220 完整报告

sgl-project/sglang

[CI] Move DSA PD+MTP+CP Layersplit test to basic B300 test suite

合并时间：2026-08-18 08:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35220

执行摘要

- 一句话：DSA 层切分端到端测试迁入基础 B300 CI，回归覆盖提前
- 推荐动作：该 PR 不需要精读源码，但它反映了 DSA 层切分回归策略的升级信号：组合路径已稳定到可以承担基础 CI 的日常开销。建议关注两点：一是 base-c 的 B300 runner 配额与模型本地缓存管理是否就绪；二是重命名后是否有旧文件名残留引用。如果想了解 DSA cache layer split 的实现细节，可以去读同仓库的 pool configurator 与 DSA 缓存层分配相关代码。

功能与动机

PR body 只保留了模板，没有填写具体说明；但从标题、diff 和同仓库近期改动可以推断：团队希望把 DSA 层切分这条关键路径的回归保护从 nightly 提前到 base-c 基础 CI，更早、更频繁地拦截 GLM-5.2 上 PD + MTP + CP + layer split 组合的精度退化。配套将模型源改为本地缓存快照，也是为了减少高频回归拉取外网模型的抖动。

实现拆解

变更只涉及一个测试文件，主要包含 3 步调整：

1. 测试文件重命名：test/registered/models_e2e/test_dsa_glm52_cache_layer_split.py 重命名为 test/registered/models_e2e/test_dsa_glm52_pd_mtp_cp_layersplit.py，文件名更精确地表达了测试覆盖的 PD 切分、MTP 投机、prefill-CP 和 DSA cache layer split 组合。
2. CI 注册参数调整：register_cuda_ci 的 stage 从 nightly 改为 base-c，runner_config 从 8-gpu-b200 改为 8-gpu-b300，意味着该用例从每晚一次的回归升级为基础提交即跑的常驻回归；est_time 仍为 750 秒，测试集保持完整 GSM8K（1319 题、准确率下限 0.935）。
3. 模型路径调整：类属性 model 从 HF Hub 的 nvidia/GLM-5.2-NVFP4 改为 radixark 本地缓存快照路径，依赖预置的模型缓存，减少 base-c 高频运行的网络拉取和启动抖动。

该文件是 end-to-end 精度测试，通过 PDDisaggregationServerBase 拉起 PD 分离部署：prefill worker 使用 TP=4、attn-cp=4、interleave CP 策略并开启 --enable-dsa-cache-layer-split，decode worker 通过 PD transfer 接收完整 cache shard。测试配套没有其他配置或 workflow 文件改动。

关键文件：

- test/registered/models_e2e/test_dsa_glm52_pd_mtp_cp_layersplit.py（模块 回归测试；类别 test；类型 rename-or-move；符号 register_cuda_ci, TestGLM52DSACacheLayerSplit, model）：本 PR 唯一变更文件，核心动作是重命名测试文件并将 CI 注册从 nightly/B200 迁移到 base-c/B300，同时切换模型到本地缓存快照，直接决定了回归覆盖频率和运行环境。

关键符号：register_cuda_ci, TestGLM52DSACacheLayerSplit

关键源码片段

test/registered/models_e2e/test_dsa_glm52_pd_mtp_cp_layersplit.py

本 PR 唯一变更文件，核心动作是重命名测试文件并将 CI 注册从 nightly/B200 迁移到 base-c/B300，同时切换模型到本地缓存快照，直接决定了回归覆盖频率和运行环境。

```
"""End-to-end GSM8K 精度测试：DSA cache layer split（GLM-5.2）。

本用例覆盖 PD 分离部署下的 prefill-CP + DSA 层切分 + MTP 投机组合；
本次迁移后从 nightly 阶段升级到 base-c 基础 CI，runner 由 8 卡 B200 改为 8 卡 B300。
"""

from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.kits.eval_accuracy_kit import GSM8KMixin
from sglang.test.server_fixtures.disaggregation_fixture import (
    PDDisaggregationServerBase,
)

# 注册为 base-c 阶段用例，限定在 8 卡 B300 runner 上执行；
# est_time = 750 表示预计耗时 750 秒，供 CI 调度器分配超时预算。
register_cuda_ci(est_time=750, stage="base-c", runner_config="8-gpu-b300")


class TestGLM52DSACacheLayerSplit(PDDisaggregationServerBase, GSM8KMixin):
    # 模型改为 radixark 本地缓存快照，避免在 base-c 高频运行时反复从 HF Hub 拉取，
    # 从而降低启动抖动与网络依赖；代价是本地缓存必须预先就绪。
    model = (
        "/data/radixark/model-cache/hub/models--nvidia--GLM-5.2-NVFP4/"
        "snapshots/aec724e8c7b8ee9db3b48c01c320f63f9cdaf8aa"
    )

    # 完整 GSM8K 测试集（1319 题），准确率下限收紧到 0.935，
    # 用于严格拦截 DSA 层切分导致的精度退化。
    gsm8k_accuracy_thres = 0.935
    gsm8k_num_questions = 1319
    gsm8k_num_threads = 200
    gsm8k_num_shots = 20

    # Prefill worker：TP = 4 + attn_cp_size = 4，interleave CP 策略，
    # 并开启 --enable-dsa-cache-layer-split，在 4 个 CP rank 间切分 KV/indexer 缓存层。
    extra_prefill_args = [
```

```
        "--tp", "4",
        "--attn-cp-size", "4",
        "--dsa-prefill-backend", "trtllm",
        "--kv-cache-dtype", "fp8_e4m3",
        "--enable-dsa-cache-layer-split",
        "--enable-prefill-cp",
        "--cp-strategy", "interleave",
        "--mem-fraction-static", "0.85",
        "--chunked-prefill-size", "4096",
        "--max-prefill-tokens", "4096",
        "--speculative-algorithm", "EAGLE",
        "--speculative-num-steps", "...", # MTP 步数由后续参数配置
    ]
```

评论区精华

PR 没有 review 评论，评论区仅有一次测试重跑交互：

- Fridge003 发出 /rerun-test test/registered/models_e2e/test_dsa_glm52_pd_mtp_cp_layersplit.py，请求在该用例改名后重新验证。
- github-actions bot 回复：在 8-gpu-b300 runner 上 1 个测试通过（workflow run #32084320617），并贴出了执行命令 cd test/ && python3 registered/models_e2e/test_dsa_glm52_pd_mtp_cp_layersplit.py。

可见迁移后的测试在 B300 上已验证通过，没有引发正确性或资源层面的争议。

- 迁移后的测试重跑验证 (test): 重命名并迁移到 B300 后，端到端测试一次通过，确认迁移本身没有引入正确性或环境问题。

风险与影响

- 风险：具体风险如下：

1. CI 负载变化：用例从 nightly 迁入 base-c，会在每个基础提交上执行，est_time=750 的长时间用例可能拉长 base-c 队列；若 B300 runner 资源有限，可能造成排队等待。
2. 模型路径依赖：model 指向 /data/radixark/model-cache/hub/models--nvidia--GLM-5.2-NVFP4/snapshots/aec724e8c7b8ee9db3b48c01c320f63f9cdaf8aa 本地快照，一旦缓存被清理或快照哈希变更，测试会立即失败，且不再像 HF 路径那样自动下载兜底。
3. runner 硬件差异：B300 与 B200 在驱动、拓扑、NCCL 行为上存在差异，虽然本次重跑已通过，但长期稳定性仍需观察，尤其涉及 PD 传输与 DSA layer split 的远程 scratch buffer 行为。
4. 引用同步风险：文件重命名后，若 CI 配置、文档或其他脚本仍引用旧文件名 test_dsa_glm52_cache_layer_split.py，可能导致测试找不到或重复执行。 - 影响：对终端用户无功能影响，仅影响 CI 流水线和回归覆盖策略。对团队而言，DSA GLM-5.2 PD + MTP + CP + Layersplit 组合的精度回归从 nightly 升级为 base-c 常驻回归，显著提高了对此路径的回归敏感度，任何相关改动都会更快暴露精度下降；同时 B300 逐步成为标准回归平台，与近期多个 B300 测试基建 PR 方向一致。整体影响范围限定在测试与 CI 基础设施，程度中等偏低。 - 风险标记：CI 阶段升级（nightly→base-c），本地模型缓存依赖，

B300 runner 硬件差异，测试重命名引用同步

关联脉络

- PR #35044 Stabilize GB300 nightly tests: 同一 B300 测试基础设施演进线，该 PR 固定 GB300 NCCL 端口并拆分测试文件，本 PR 进一步把 DSA 组合测试迁入 B300 基础 CI，延续对 B300 平台的回归加固。
- PR #30531 [DSA] Skip indexer KV cache for skip-topk layers: 同属 DSA 缓存层功能线，涉及 DSA indexer 缓存分配与 layer split 相关实现，本 PR 所迁移的精度测试正是为了守护该功能线的回归安全。