PR #24973 完整报告

sgl-project/sglang

[CI] Add DSV4 Flash disaggregation test

合并时间：2026-05-13 18:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24973

执行摘要

- 一句话：新增 DSV4 Flash 分离式推理 CI 测试
- 推荐动作：本 PR 结构清晰、注释简洁，适合作为编写分离式端到端测试的范例。建议关注 GSM8K 阈值在长期运行中的稳定性，若出现 flaky 可调整为 0.90。另外，TP/DP 配比的设计选择值得注意：在 8 卡环境下用 4+4 分离而非 8+0 或 1+1，体现了对 DP 注意力测试覆盖的刻意倾斜。

功能与动机

确保 DeepSeek-V4 Flash 模型在分离式 prefill/decode 部署下的推理质量和功能正确性，防止回归。PR body 未详细说明动机，但从 CI 测试性质推断，旨在填补该模型在分离式场景下的测试盲区。

实现拆解

1. 继承 PDDisaggregationServerBase 编写测试类 TestDisaggregationDSV4，复写 setUpClass 编排整个流程。
2. 在 setUpClass 中依次调用 start_prefill 和 start_decode 启动两个独立服务进程：
   - prefill 服务配置 --tp 4 --dp 4 --enable-dp-attention --moe-a2a-backend deepep，并传入 EAGLE 投机参数（_EAGLE_SPEC_ARGS）。
   - decode 服务同样配置 TP/DP=4，但指定 --base-gpu-id 4 将 GPU 分配偏移 4 张卡，与 prefill 共用 8 卡。 服务启动后等待 health 端点就绪，最后调用 launch_lb 负载均衡。
3. 实现 test_gsm8k 方法，使用 sglang.test.run_eval 加载 GSM8K 评估（200 样本、128 线程），断言准确率大于 0.95。
4. 文件顶部通过 register_cuda_ci(est_time=250, suite="stage-c-test-dsv4-8-gpu-h200") 注册到指定 CI 套件。
5. 定义环境变量 DSV4_FLASH_ENV 禁用 FP4 专家、限制 DeepEP 最大调度 token 数。

关键文件：

- test/registered/distributed/test_disaggregation_dsv4.py（模块 分离式推理；类别 test；类型 test-coverage；符号 TestDisaggregationDSV4, setUpClass, start_prefill, start_decode）：唯一变更文件，新增 DSV4 Flash 分离式推理端到端测试，覆盖 DP 注意力、EAGLE 投机、DeepEP 等关键特性，为 CI 提供回归保护。

关键符号：setUpClass, start_prefill, start_decode, test_gsm8k

关键源码片段

test/registered/distributed/test_disaggregation_dsv4.py

唯一变更文件，新增 DSV4 Flash 分离式推理端到端测试，覆盖 DP 注意力、EAGLE 投机、DeepEP 等关键特性，为 CI 提供回归保护。

```
# 测试类继承自 PDDisaggregationServerBase，自动获得 prefill/decode 服务管理能力
class TestDisaggregationDSV4(PDDisaggregationServerBase):
    @classmethod
    def setUpClass(cls):
        super().setUpClass()
        cls.model = try_cached_model(DSV4_FLASH_MODEL) # 从缓存获取模型
        cls.start_prefill() # 启动 prefill 服务（4 卡）
        cls.start_decode() # 启动 decode 服务（另 4 卡）
        cls.wait_server_ready(cls.prefill_url + "/health", process=cls.process_prefill)
        cls.wait_server_ready(cls.decode_url + "/health", process=cls.process_decode)
        cls.launch_lb() # 启动负载均衡器

    @classmethod
    def start_prefill(cls):
        # prefill 参数：DP 注意力、DeepEP、EAGLE 投机
        prefill_args = [
            "--trust-remote-code",
            "--disaggregation-mode", "prefill",
            "--disaggregation-bootstrap-port", cls.bootstrap_port,
            "--tp", 4, "--dp", 4, # 4 卡 TP，4 路 DP → 共 4 卡
            "--enable-dp-attention",
            "--moe-a2a-backend", "deepep", # 使用 DeepEP 通信
            "--deepep-config", DEEPEP_CONFIG,
            "--cuda-graph-max-bs", "128",
            "--max-running-requests", "256",
            "--mem-fraction-static", "0.7",
            *_EAGLE_SPEC_ARGS, # 追加投机解码参数
        ] + cls.transfer_backend + cls.rdma_devices
        cls.process_prefill = popen_launch_pd_server(
            cls.model, cls.prefill_url, timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=prefill_args, env=DSV4_FLASH_ENV)

    # test_gsm8k 使用统一评估框架，200 样本，128 线程，设定 0.95 阈值
    def test_gsm8k(self):
        args = SimpleNamespace(
            base_url=self.base_url, model=self.model,
            eval_name="gsm8k", api="completion",
            max_tokens=512, num_examples=200, num_threads=128,
        )
        metrics = run_eval(args)
        print(f"Evaluation metrics: {metrics}")
```

```
        self.assertGreater(metrics["score"], 0.95)
```

评论区精华

Review 主要围绕三个问题展开：

- TP/DP 配比与 GPU 限制（正确性）：机器人评论指出 --tp 4 --dp 4 在 prefill 和 decode 各自需要 16 张 GPU，超出 CI 环境的 8 卡限制，建议将 tp 降为 1。作者最终选择将测试从 stage-c-test-8-gpu-h20 迁移到 stage-c-test-dsv4-8-gpu-h200，利用 H200 的 8 卡独立运行，prefill 和 decode 各用 4 张卡，实际总消耗 8 卡并未超限，因此未修改 tp/dp 参数。
- 评估模块迁移（风格）：机器人建议弃用已废弃的 sglang.test.few_shot_gsm8k，改用 sglang.test.run_eval。作者在后续提交中采纳并修正。
- 准确率阈值（测试）：机器人质疑 0.95 阈值对 Flash 模型可能过高，容易导致测试不稳定。作者最终保留 0.95（CI 通过），但 flaky 风险仍未完全解决，可视为待观察的公开问题。
- TP/DP 配置超出 8 卡 CI 环境限制 (correctness): 作者将测试注册目标从 8-H20 改为 8-H200，实际使用 4+4 共 8 卡，符合限制，未修改 tp/dp。
- 评估模块弃用警告 (style): 作者在后续提交中替换为 run_eval，遵循建议。
- GSM8K 准确率阈值过高 (testing): 作者保留 0.95 阈值（最终代码未改），CI 通过但 flaky 风险仍存在，可视为待观察。

风险与影响

- 风险：
  1. 硬件环境依赖：测试依赖 8×H200 GPU、DeepEP 通信库及 flash_mla，无法在低配环境运行，限制了 CI 复用的灵活性。
  2. 准确率阈值风险：GSM8K 阈值 0.95 可能对 Flash 模型偏高，模型量化或批次差异可能导致偶发下降，引发 CI 误报。
  3. CI 时间开销：预估 250 秒，服务启动、评估耗时较长，可能延长流水线周期。
  4. 无源码变更：本 PR 只添加测试，不引入业务逻辑风险。
- 影响：
  - 用户/模型质量保护：覆盖DSV4Flash分离式部署关键路径（DP注意力、EAGLE投机、DeepEP 通信），显著降低回归漏测风险。
  - CI 资源占用：新增一个 8-GPU H200 测试，对 H200 队列造成额外压力，需确保机器可用。
  - 团队工作流程：为后续同类测试提供参考模板，降低编写分离式测试的门槛。
  - 风险标记：硬件环境依赖，准确率阈值偏高，CI 时间较长

关联脉络

- PR #25039 [AMD] Disable unittest fail-fast for deepseekv4 perf test: 同为 DeepSeek-V4 系列测试配置调整，涉及 CI 测试稳定性，可相互参考。