PR #26141 完整报告

sgl-project/sglang

Add non-MTP DSV4 test coverage

合并时间：2026-05-23 10:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26141

执行摘要

- 一句话：为 DeepSeek V4 增加非 MTP 模式测试覆盖
- 推荐动作：值得查阅，了解 DeepSeek V4 在 B200 和 H200 上的非 MTP 测试配置差异，为后续类似测试添加提供模板。

功能与动机

PR body 明确指出 "Cherrypicked from #25801"，目的是为非 MTP 模式的 DeepSeek V4 配置增加端到端测试覆盖，确保在没有推测解码的配置下模型推理的正确性和稳定性。

实现拆解

1. 在 B200 测试文件中新增测试类：在 test/registered/models_e2e/test_deepseek_v4_flash_fp4_b200.py 中新增 TestDSV4FlashFP4NonMTPB200 类，继承 BasicDecodeCorrectnessMixin、GSM8KMixin 和 CustomTestCase，验证在 TP=4、DP=4、DeepEP 后端、不使用推测解码（无 --speculative-algorithm 等参数）时 GSM8K 准确率 ≥ 0.93。
2. 在 H200 测试文件中新增测试类：在 test/registered/models_e2e/test_deepseek_v4_flash_fp4_h200.py 中新增 TestDSV4FlashFP4NonMTPH200 类，使用 TP=4、Marlin FP4 后端、--watchdog-timeout 900，同样无推测解码参数，GSM8K 准确率阈值 0.93。
3. 测试配置差异：B200 使用 DeepEP 后端和 DP 注意力，H200 使用 Marlin FP4 后端，两者均未启用 EAGLE 推测解码，与已有的 MTP 版本形成对照。
4. CI 验证：PR 的 CI 测试在 B200（4-gpu-b200）和 H200（8-gpu-h200）上均通过（绿色勾号）。

关键文件：

- test/registered/models_e2e/test_deepseek_v4_flash_fp4_b200.py（模块 测试脚本；类别 test；类型 test-coverage；符号 TestDSV4FlashFP4NonMTPB200, setUpClass, tearDownClass）：新增非 MTP 测试类，验证 B200 上 FP4 无推测解码的正确性。
- test/registered/models_e2e/test_deepseek_v4_flash_fp4_h200.py（模块 测试脚本；类别 test；类型 test-coverage；符号 TestDSV4FlashFP4NonMTPH200, setUpClass, tearDownClass）：新增非 MTP 测试类，验证 H200 上 FP4 无推测解码的正确性。

关键符号：setUpClass, tearDownClass

关键源码片段

test/registered/models_e2e/test_deepseek_v4_flash_fp4_b200.py

新增非 MTP 测试类，验证 B200 上 FP4 无推测解码的正确性。

```
class TestDSV4FlashFP4NonMTPB200(
    BasicDecodeCorrectnessMixin, GSM8KMixin, CustomTestCase
):
    """Non-MTP recipe: TP=4, DP=4, DeepEP, no speculative decoding."""

    # GSM8K 准确率阈值，与 MTP 版本保持一致
    gsm8k_accuracy_thres = 0.93

    @classmethod
    def setUpClass(cls):
        cls.model = try_cached_model(MODEL)
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=SERVER_LAUNCH_TIMEOUT,
            other_args=[
                "--trust-remote-code",
                "--tp",
                "4", # Tensor 并行度 4
                "--dp",
                "4", # Data 并行度 4
                "--enable-dp-attention",
                "--moe-a2a-backend",
                "deepep", # DeepEP MoE All-to-All 后端
                "--deepep-config",
                DEEPEP_CONFIG,
            ],
            env=_DEEPEP_ENV,
        )

    @classmethod
    def tearDownClass(cls):
        if hasattr(cls, "process") and cls.process:
            kill_process_tree(cls.process.pid)
```

test/registered/models_e2e/test_deepseek_v4_flash_fp4_h200.py

新增非 MTP 测试类，验证 H200 上 FP4 无推测解码的正确性。

```
class TestDSV4FlashFP4NonMTPH200(
    BasicDecodeCorrectnessMixin, GSM8KMixin, CustomTestCase
):
    """LowLatency recipe without MTP: TP=4, Marlin FP4, no speculative decoding."""
```

```
    gsm8k_accuracy_thres = 0.93

    @classmethod
    def setUpClass(cls):
        cls.model = try_cached_model(MODEL)
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=SERVER_LAUNCH_TIMEOUT,
            other_args=[
                "--trust-remote-code",
                "--tp",
                "4", # Tensor 并行度 4
                "--moe-runner-backend",
                "marlin", # Marlin FP4 MoE 后端
                "--watchdog-timeout",
                "900", # 超时保护
            ],
        )

    @classmethod
    def tearDownClass(cls):
        if hasattr(cls, "process") and cls.process:
            kill_process_tree(cls.process.pid)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅添加测试用例，不修改任何生产代码或源码逻辑。测试覆盖的配置（无 MTP）可能在特定条件下暴露潜在缺陷，但测试本身不会引入回归风险。
- 影响：对用户和系统无直接影响。对团队影响：增强了对 DeepSeek V4 非推测解码模式的 CI 验证覆盖，有助于在后续重构中快速捕获回归。影响范围仅限于两个测试文件。
- 风险标记：暂无

关联脉络

- PR #25801 (original PR): 本 PR 从中 cherry-pick 了非 MTP 测试用例。