# PR #2217 完整报告

radixark/miles

test(ci): right-size session model GPU coverage

合并时间：2026-08-07 04:02

原文链接：http://prhub.com.cn/radixark/miles/pull/2217

## 执行摘要

- 一句话：调整 session 模型 CI GPU 规格并启用 EAGLE
- 推荐动作：值得精读，尤其是学习如何根据模型特性（KV heads、模型大小、服务需求）定制 CI 资源分配，以及如何安全地在测试中启用 EAGLE 投机解码。对于负责 CI 基础设施和 session 测试的开发者有参考价值。

## 功能与动机

PR body 明确指出 'Run GLM-4.7, Nemotron-3, and Qwen3 on the 2-GPU H200 lane with per-model TP sizes that match their serving needs. Enable the existing EAGLE preset for the GLM-4.7, Nemotron 3, and Qwen3.5 checkpoints that ship compatible MTP weights, keeping speculative decoding limited to active session models. Remove the disabled Qwen3-Next session verifier because active coverage has moved to Qwen3.5.' 其核心是让 CI 资源分配与模型实际服务需求匹配，避免统一 4-GPU 的浪费，并让投机解码覆盖在具备 MTP 权重的模型上。

## 实现拆解

1. 套件迁移：修改 test_nemotron3.py、test_glm47.py、test_qwen3.py 中的 register_cuda_ci 调用，将 suite 从 stage-c-4-gpu-h200 改为 stage-c-2-gpu-h200；test_qwen35.py 保持 stage-c-4-gpu-h200 不变。
2. 资源与 TP 配置：为每个模型增加 num_gpus=2，并按需调整 tp_size：Nemotron-3 为 TP2（匹配模型两个 KV heads），GLM-4.7 与 Qwen3 为 TP1（单卡即可服务），Qwen3.5 维持 TP2。
3. 启用 EAGLE：在 test_glm47.py、test_nemotron3.py、test_qwen35.py 的 CONFIG 中增加 enable_spec=True，复用现有 EAGLE preset，未引入新 serving flags。
4. 删除 Qwen3-Next 测试：移除 test_qwennext.py 整个文件（此前处于 disabled 状态），覆盖由 Qwen3.5 承接。
5. 验证方式：无新增测试文件，lane 注册即覆盖；通过 pytest --collect-only 确认全部 7 个 lane 均可收集。

关键文件：

- tests/e2e/sglang/test_session_server_multi_role/test_qwennext.py（模块 会话测试；类别 test；类型 deletion；符号 test_qwennext）：删除已禁用的 Qwen3-Next 会话验证器，

覆盖转移到 Qwen3.5

- tests/e2e/sglang/test_session_server_multi_role/test_nemotron3.py（模块 会话测试；类别 test；类型 test-coverage）：迁移到 2-GPU lane，TP2 匹配 KV heads，启用 EAGLE spec
- tests/e2e/sglang/test_session_server_multi_role/test_glm47.py（模块 会话测试；类别 test；类型 test-coverage）：迁移到 2-GPU lane，TP1 服务，启用 EAGLE spec
- tests/e2e/sglang/test_session_server_multi_role/test_qwen3.py（模块 会话测试；类别 test；类型 test-coverage）：迁移到 2-GPU lane，TP1 服务，满足 Qwen3 的 serving 需求
- tests/e2e/sglang/test_session_server_multi_role/test_qwen35.py（模块 会话测试；类别 test；类型 test-coverage）：保持 4-GPU lane，仅启用 EAGLE 投机解码

关键符号：test_qwennext, test_nemotron3, test_glm47, test_qwen3, test_qwen35

## 关键源码片段

### tests/e2e/sglang/test_session_server_multi_role/test_nemotron3.py

迁移到 2-GPU lane，TP2 匹配 KV heads，启用 EAGLE spec

```
from tests.ci.ci_register import register_cuda_ci
from tests.e2e.sglang.test_session_server_multi_role._common import ModelConfig, run_both_
versions

# 注册 CI lane：Nemotron-3 使用 2-GPU H200 套件，替代原 4-GPU 统一分配
register_cuda_ci(est_time=800, suite="stage-c-2-gpu-h200", labels=["sglang"])

# Nemotron-3-Super-120B-A12B-FP8 约 120GB，TP2 跨越两块 H200 并匹配模型两个 KV heads
CONFIG = ModelConfig(
    model_name="nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-FP8",
    reasoning_parser="nemotron_3",
    tool_call_parser="qwen3_coder", # 工具调用使用与 Qwen3.5 相同的 XML 包裹
    tito_model="nemotron3",
    num_gpus=2,
    tp_size=2,
    enable_spec=True, # 启用 EAGLE 投机解码，复用已有 preset
    cycles=2,
    assistant_text_threshold=1.0, # nemotron_3 parser 尾部换行导致 roundtrip 漂移，放宽阈值
    tool_call_failure_mode="append_tool",
)


def test_nemotron3():
    run_both_versions(CONFIG)


if __name__ == "__main__":
    test_nemotron3()
```

### tests/e2e/sglang/test_session_server_multi_role/test_glm47.py

迁移到 2-GPU lane，TP1 服务，启用 EAGLE spec

```
from tests.ci.ci_register import register_cuda_ci
from tests.e2e.sglang.test_session_server_multi_role._common import ModelConfig, run_both_
versions

# 迁移到 2-GPU H200 lane，GLM-4.7 以 TP1 方式服务（单 GPU 即可容纳）
register_cuda_ci(est_time=500, suite="stage-c-2-gpu-h200", labels=["sglang"])

CONFIG = ModelConfig(
    model_name="zai-org/GLM-4.7-Flash",
    reasoning_parser="glm45",
    tool_call_parser="glm47",
    tito_model="glm47",
    num_gpus=2,
    tp_size=1,
    enable_spec=True, # 启用 EAGLE 投机解码
    # Lenient template：tool message 渲染不校验前序 assistant 的 tool_call.id，
    # 使 APPEND_TOOL 哨兵（"tool_call_id": "none"）可干净回环
    tool_call_failure_mode="append_tool",
)


def test_glm47():
    run_both_versions(CONFIG)


if __name__ == "__main__":
    test_glm47()
```

## 评论区精华

PR 无 review 评论。作者在 body 中列出三个 Review Focus：关注 Nemotron-3 TP2 与模型两个 KV heads 的匹配、GLM-4.7 TP1 在 2-GPU lane 上的 serving 可行性，以及 Qwen3-Next 删除的覆盖转移合理性。

- 暂无高价值评论线程

## 风险与影响

- 风险：
  1. 资源调整风险：若模型实际推理需要更高 TP，TP1/TP2 可能导致显存不足（OOM）或性能下降，尤其 GLM-4.7 在 TP1 下需验证显存占用是否可接受。
  2. EAGLE 稳定性风险：enable_spec=True 会在 session 场景下启用投机解码，可能引入额外的超时或不稳定因素，需持续观察 CI 结果。

3. 覆盖移除风险：删除 Qwen3-Next 测试后该模型将失去 CI 覆盖，若后续重新启用需恢复测试文件。
4. 依赖风险：本 PR 依赖 #2130，若父 PR 未合入，当前分支可能无法独立在 CI 中运行。 - 影响：影响范围集中在 CI 测试配置，不涉及生产代码。对团队而言，可更高效利用 H200 资源，减少闲置 GPU 浪费；同时 session 模型的验证与线上服务配置更一致，确保投机解码只在支持的模型上启用。对开发流程的影响是 CI 耗时可能缩短，但需关注新的 lane 配置是否稳定。 - 风险标记：依赖 #2130, EAGLE 启用 , QwenNext 覆盖移除

## 关联脉络

- PR #2202 fix(tito): prevent DeepSeek V4 system-tail mismatch: 同一测试目录（ tests/e2e/sglang/test_session_server_multi_role/）下的 session 覆盖调整
- PR #2075 session: apply the trained LoRA adapter to session-server rollouts: 同属 session 功能线，涉及 session server 行为，本 PR 对其 CI 验证资源进行右规模调整