PR #28458 完整报告

sgl-project/sglang

[AMD] ci: add extra-a 1-gpu-large tier (fp8kv-triton, streaming-session, spec-standalone)

合并时间：2026-06-18 14:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28458

执行摘要

- 一句话：新增 AMD CI extra-a 1-gpu-large 测试层
- 推荐动作：该 PR 展示了良好的 CI 分层设计：将测试按硬件平台和资源需求分层，并通过标签控制触发。其跨平台测试注册方式（register_cuda_ci / register_amd_ci）值得其他后端参考。建议关注后续可能将更多通过验证的测试从 CUDA extra-a 移植到 AMD。

功能与动机

PR body 指出，目的是将 CUDA extra-a 测试层中在 AMD mi325 GPU 上验证通过的模型 e2e 测试纳入 AMD CI，以提前发现回归。之前 AMD extra-a 只包含 mock-model/kv_canary 单元测试，缺乏真实模型测试覆盖。作者通过多轮 discovery run 筛选出可移植的子集。

实现拆解

1. 在 .github/workflows/pr-test-amd-extra.yml 中新增 extra-a-test-1-gpu-large-amd 作业，复用 1-gpu-small 的容器启动脚本，执行 run_suite.py --hw amd --suite extra-a-test-1-gpu-large-amd。
2. 在 test/run_suite.py 的 PER_COMMIT_SUITES[HWBackend.AMD] 列表中注册新 suite extra-a-test-1-gpu-large-amd，并更新注释说明其包含的测试子集。
3. 为三个测试文件添加 register_amd_ci 调用：test_fp8kv_triton.py、test_streaming_session_extra.py、test_spec_standalone_extra.py，将其注册到 AMD CI 的 extra-a 大套件中。
4. 在 test_spec_standalone_extra.py 中，对 FA3 和 FlashInfer 后端的测试类添加 @unittest.skipIf(is_hip(), ...) 跳过逻辑，因为 ROCm 的 sgl_kernel 未构建这些后端，仅保留 Triton 变体在 AMD 上运行。
5. CI 作业的 fan-in 聚合器 pr-test-amd-extra-finish 自动覆盖新 job，无额外配置。

关键文件：

- .github/workflows/pr-test-amd-extra.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：核心 CI 配置，新增 extra-a 1-gpu-large 作业，定义了测试容器启动和执行流程。
- test/registered/spec/test_spec_standalone_extra.py（模块 推测解码测试；类别 test；类型 test-coverage；符号 register_amd_ci, register_cuda_ci,

TestStandaloneSpeculativeDecodingBase, TestStandaloneSpeculativeDecodingTriton）：测试文件，添加 AMD CI 注册并对 ROCm 不可用的后端类应用 skipIf 跳过逻辑，确保测试只在可用后端运行。

- test/run_suite.py（模块 测试注册；类别 test；类型 test-coverage）：测试套件注册文件，将新 suite 'extra-a-test-1-gpu-large-amd' 加入 AMD per-commit suites 列表，并更新注释说明内容。
- test/registered/quant/test_fp8kv_triton.py（模块 FP8KV 测试；类别 test；类型 test-coverage；符号 register_amd_ci, register_cuda_ci, TestFP8KVCacheTritonBackend）：量化测试文件，添加一行 register_amd_ci 注册到 AMD large 套件。
- test/registered/sessions/test_streaming_session_extra.py（模块 流式会话测试；类别 test；类型 test-coverage；符号 register_amd_ci, register_cuda_ci, TestStreamingSessionRetractMixedChunk, TestStreamingSessionRetractLargePage）：流式会话测试文件，添加一行 register_amd_ci 注册到 AMD large 套件。

关键符号：register_amd_ci, register_cuda_ci, TestFP8KVCacheTritonBackend, TestStreamingSessionRetractMixedChunk, TestStreamingSessionRetractLargePage, TestStandaloneSpeculativeDecodingBase, TestStandaloneSpeculativeDecodingTriton, TestStandaloneSpeculativeDecodingFlashinfer

关键源码片段

test/registered/spec/test_spec_standalone_extra.py

测试文件，添加 AMD CI 注册并对 ROCm 不可用的后端类应用 skipIf 跳过逻辑，确保测试只在可用后端运行。

```
import unittest

from sglang.srt.utils import is_hip
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.server_fixtures.standalone_fixture import StandaloneServerBase
from sglang.test.test_utils import CustomTestCase

# 注册到 CUDA extra-a 套件（不变）
register_cuda_ci(est_time=406, stage="extra-a", runner_config="1-gpu-large")
# 注册到 AMD extra-a large 套件（新增）
register_amd_ci(est_time=103, suite="extra-a-test-1-gpu-large-amd")

# 定义跳过原因：fa3 / flashinfer 后端未在 ROCm sgl_kernel 中构建
_AMD_SKIP_BACKEND = (
    "fa3 / flashinfer attention backends are CUDA-only "
    "(not in the ROCm sgl_kernel build)"
)


@unittest.skipIf(is_hip(), _AMD_SKIP_BACKEND)
```

```
class TestStandaloneSpeculativeDecodingBase(StandaloneServerBase, CustomTestCase):
    """FA3 后端测试类，在 ROCm 上跳过"""
    attention_backend = "fa3"
    speculative_eagle_topk = 2
    speculative_num_draft_tokens = 7
    disable_overlap = True


class TestStandaloneSpeculativeDecodingTriton(StandaloneServerBase, CustomTestCase):
    """Triton 后端测试类，所有平台均运行（包括 ROCm）"""
    attention_backend = "triton"
    speculative_eagle_topk = 2
    speculative_num_draft_tokens = 7
    disable_overlap = True
    enable_deterministic_inference = True


@unittest.skipIf(is_hip(), _AMD_SKIP_BACKEND)
class TestStandaloneSpeculativeDecodingFlashinfer(StandaloneServerBase, CustomTestCase):
    """FlashInfer 后端测试类，在 ROCm 上跳过"""
    attention_backend = "flashinfer"
    speculative_eagle_topk = 2
    speculative_num_draft_tokens = 7
    disable_overlap = True


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

评论区精华

PR 中无审查评论，但作者在 PR body 和后续评论中详细展示了通过 continue_on_error 发现模式对多个候选测试进行筛选的过程，最终仅保留 3 个测试。作者也列出了因 ROCm gaps（如缺失 flash_attn.cute、精度不达标等）而被排除的测试及其失败原因（如 ngram tree-mask op 缺失、lora 精度失败、chunked_prefill 超时等）。决策结论是只移植确认通过的子集，其余保持 CUDA-only。

- AMD extra-a 测试子集选择 (other): 仅移植确认通过的三个测试，其余保持 CUDA-only，并更新注册注释说明原因。

风险与影响

- 风险：风险较低，因为新增 CI 仅在标签 run-ci-extra 触发时执行，不影响默认 PR 流程。主要风险有： 1) CI 资源消耗增加，三个测试在 mi325 上总耗时约 735 秒； 2) spec_standalone 测试中跳过了 fa3 和 flashinfer 后端，导致 AMD 平台上这部分覆盖缺失，若 Triton 后端的测试通过但其他后端有回归则无法捕获； 3) 测试依赖外部模型（如 neuralmagic/Meta-Llama-3-8B-Instruct-FP8-KV、EAGLE3 模型），若模型不可用可能导致 CI 失败。

- 影响：影响范围局限于 AMD CI 的 *extra-a* 层。对 CUDA CI 无影响。对 AMD 开发者意味着需要时可通过添加 run-ci-*extra* 标签触发更多测试。对下游用户无直接影响。整体影响程度中等，因为增加了 ROCm 平台的真实模型测试覆盖，有助于提升 AMD 后端的稳定性。
- 风险标记：CI 时间增加，测试覆盖盲区 (fa3/flashinfer 未测试)

关联脉络

- PR #28378 [AMD] Fix Always mask padded topk_ids on HIP to prevent garbage MoE routing (DeepSeek-R1-MXFP4 accuracy regression): 同为 AMD 平台的 bugfix，提升了 ROCm MoE 精度，本 PR 增加的测试有助于验证此类修复不会引入回归。