# PR #24599 完整报告

sgl-project/sglang

[codex] Split diffusion quant CI coverage

合并时间：2026-05-16 22:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24599

## 执行摘要

- 一句话：拆分 diffusion 量化 CI 为 FP8 和 B200 两套测试
- 推荐动作：值得阅读，尤其是了解如何通过简单的配置拆分优化 CI 硬件利用率。建议关注 gpu_cases.py 中列表定义的模式，以及 diffusion_case_parser.py 中分区映射的写法，这种模式可以在其他需要硬件隔离的测试场景中复用。

## 功能与动机

随着 diffusion 模型量化覆盖的增加，FP8 和 NVFP4 测试对硬件要求不同——FP8 可在 H100 运行，NVFP4 需要 B200。原设计将所有 ModelOpt 量化测试混在一个列表，导致每次 CI 都需要 B200 资源。拆分后，FP8 测试可在常规 GPU CI 上执行，NVFP4 测试仅在 B200 上执行，从而降低 CI 整体等待时间并提高硬件利用率。

## 实现拆解

1. 拆分测试用例列表：在 python/sglang/multimodal_gen/test/server/gpu_cases.py 中，将原有的 ONE_GPU_MODELOPT_CASES 拆分为 ONE_GPU_MODELOPT_FP8_CASES（6 个 FP8 测试）和 ONE_GPU_MODELOPT_NVFP4_CASES（3 个 NVFP4 测试）。新增别名 ONE_GPU_B200_CASES = ONE_GPU_MODELOPT_NVFP4_CASES，并修改 ONE_GPU_CASES 的拼接来源为 ONE_GPU_MODELOPT_FP8_CASES，确保 FP8 测试加入常规 1-GPU suite。
2. 更新 B200 测试入口：修改 python/sglang/multimodal_gen/test/server/test_server_b200.py，将 fixture 参数从 ONE_GPU_MODELOPT_CASES 改为 ONE_GPU_B200_CASES，类文档字符串也随之更新。
3. 同步文档注释：修改 python/sglang/multimodal_gen/test/server/testcase_configs.py 中的用例添加指引文档，反映新的列表名称。
4. 更新 CI 分区映射：在 scripts/ci/utils/diffusion/diffusion_case_parser.py 中添加 "ONE_GPU_MODELOPT_FP8_CASES": "1-gpu" 和 "ONE_GPU_B200_CASES": "1-gpu-b200" 条目，保证分区脚本能正确将 NVFP4 测试分配到 B200 分区。

关键文件：

- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 测试用例；类别 test；类型 test-coverage；符号 ONE_GPU_MODELOPT_FP8_CASES, ONE_GPU_MODELOPT_NVFP4_CASES, ONE_GPU_B200_CASES）：核心变更：拆分原

有的 ONE_GPU_MODELOPT_CASES 为 FP8 和 NVFP4 两组列表，并新增 ONE_GPU_B200_CASES 别名。

- python/sglang/multimodal_gen/test/server/test_server_b200.py（模块 B200 测试；类别 test；类型 test-coverage）：B200 测试入口更新，导入列表从 ONE_GPU_MODELOPT_CASES 改为 ONE_GPU_B200_CASES。
- python/sglang/multimodal_gen/test/server/testcase_configs.py（模块 测试配置；类别 test；类型 documentation）：更新文档注释，引用新的列表名称。
- scripts/ci/utils/diffusion/diffusion_case_parser.py（模块 CI 配置；类别 infra；类型 infrastructure）：更新 CI 分区映射，添加新列表的分区标识。

关键符号：未识别

## 关键源码片段

### python/sglang/multimodal_gen/test/server/gpu_cases.py

核心变更：拆分原有的 ONE_GPU_MODELOPT_CASES 为 FP8 和 NVFP4 两组列表，并新增 ONE_GPU_B200_CASES 别名。

```
# 在 AMD 上跳过全部 ModelOpt 测试
if current_platform.is_hip():
    ONE_GPU_MODELOPT_FP8_CASES = []
    ONE_GPU_MODELOPT_NVFP4_CASES = []
else:
    # FP8 测试：可运行在 H100 等常规 GPU 上
    ONE_GPU_MODELOPT_FP8_CASES = [
        _make_modelopt_ci_case("flux1_modelopt_fp8_t2i", ...),
        _make_modelopt_ci_case("flux2_modelopt_fp8_t2i", ...),
        _make_modelopt_ci_case("wan22_modelopt_fp8_t2v", ...),
        _make_modelopt_ci_case("hunyuanvideo_modelopt_fp8_t2v", ...),
        _make_modelopt_ci_case("qwen_image_modelopt_fp8_t2i", ...),
        _make_modelopt_ci_case("qwen_image_edit_modelopt_fp8_ti2i", ...),
    ]
    # NVFP4 测试：仅在 Blackwell (B200) 上运行
    ONE_GPU_MODELOPT_NVFP4_CASES = [
        _make_modelopt_ci_case("flux1_modelopt_nvfp4_t2i", ...),
        _make_modelopt_ci_case("flux2_modelopt_nvfp4_t2i", ...),
        _make_modelopt_ci_case("wan22_modelopt_nvfp4_t2v", ...),
    ]

# B200 suite 直接引用 NVFP4 列表
ONE_GPU_B200_CASES = ONE_GPU_MODELOPT_NVFP4_CASES

# 常规 1-GPU suite 只包含 FP8 测试
ONE_GPU_CASES += ONE_GPU_MODELOPT_FP8_CASES
```

## 评论区精华

Review 中 gemini-code-assist[bot] 在 transformer_load_utils.py 上对 nunchaku 路径选择逻辑提出改进建议，建议使用 safetensors_list[0] 而非 nunchaku_config.transformer_weights_path 以避免潜在错误。但该 PR 最终移除了全部 nunchaku 相关变更，此评论已不再适用。除此之外无其他实质讨论。

- Nunchaku 文件路径选择建议 (design): 该评论对应的 nunchaku 相关变更在后续提交中被完全移除，因此建议未采纳。

## 风险与影响

- 风险：本次变更仅涉及测试配置和 CI 映射，核心逻辑无变化。主要风险在于分区配置错误导致 FP8 测试被错误调度到 B200 分区（反之亦然），但作者已在 H200 环境下通过 compute_diffusion_partitions.py 验证分区结果符合预期。另一风险是未来新增测试用例时忘记更新分区映射，但已有文档说明。
- 影响：对用户无影响。对 CI 系统：1-GPU 常规 CI 将承担更多 FP8 量化测试，B200 CI 则专注于 NVFP4 和未来可能的 Nunchaku 测试，从而减少 B200 资源占用，提高整体 CI 吞吐。对开发流程：测试用例列表拆分为更细粒度的职责，方便按需调整覆盖。
- 风险标记：配置一致性，测试覆盖拆分遗漏，硬件依赖

## 关联脉络

- 暂无明显关联 PR