PR #21767 完整报告

sgl-project/sglang

[CI] add nvfp4 ci test for b200;

合并时间：2026-04-02 11:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21767

执行摘要

本 PR 为 B200 GPU 添加了专用的 NVFP4 量化扩散模型 CI 测试路径，通过更新 CI 工作流、测试套件和工具脚本，实现了硬件特定测试的自动化。影响范围限于 CI 基础设施，无用户端直接变更，但提升了团队对 B200 硬件的测试覆盖。

功能与动机

动机源于需要为 B200 类 runner 建立独立的多模态生成扩散 CI 路径，以补充现有 H100 作业。PR body 中明确表述：“添加专用的多模态生成扩散 CI 路径在 B200 类 runner 上，与现有的 1-GPU/2-GPU H100 作业分离”，目标是使新测试阶段可发现和可重跑，类似其他硬件特定 CI。

实现拆解

- CI 工作流：在 .github/workflows/pr-test-multimodal-gen.yml 中新增 job multimodal-gen-test-1-b200，指定 B200 runner 并运行测试套件。
- 测试套件：python/sglang/multimodal_gen/test/run_suite.py 添加“1-gpu-b200”套件，指向新文件 test_server_c.py。
- 测试文件：新增 test_server_c.py 定义 TestDiffusionServerOneGpuB200 类，使用配置驱动测试。
- 性能基线：大幅更新 perf_baselines.json，移除旧数据，添加“flux_2_nvfp4_t2i”等新基线。
- 工具脚本：更新 gen_diffusion_ci_outputs.py 和 slash_command_handler.py 以支持新 suite 和 stage。
- 配置清理：修改 testcase_configs.py，移除过时 TODO 并启用 B200 测试。

评论区精华

review 讨论聚焦于代码维护性和测试启用：

- gemini-code-assist[bot] 建议：“For better maintainability... use list(SUITES.keys())”，已被采纳，提升脚本可维护性。
- mickqian 与 Prozac614 交互：mickqian 评论“we should enable this”，Prozac614 回复“OK, I'll enable this”，确保测试配置正确激活。

风险与影响

- 技术风险：CI 配置变更可能导致 runner 依赖问题或测试失败；性能基线更新可能引入数据偏差，影响回归检测；新测试套件覆盖可能不足。
- 影响分析：对用户无感知；系统层面扩展了 CI 硬件支持，团队需维护新 job，但增强了测试自动化能力。

关联脉络

从历史 PR 看，本 PR 与扩散模型硬件支持（如 PR 18648）、量化集成（如 PR 21576）和硬件特定 CI 测试（如 PR 20717）密切相关，共同推动多模态生成和量化测试的演进。