PR #21487 完整报告

sgl-project/sglang

feat(ci): add GB300 nightly benchmark test suites

合并时间：2026-03-29 12:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21487

执行摘要

此 PR 为 sglang 仓库的 GB300（4x B200 NVL4）硬件新增了 nightly 基准测试套件，包含 8 个测试文件覆盖 Qwen3.5-397B、GLM-5、Kimi-K2.5 和 DeepSeek-V3.2 的 NVFP4 和 FP8 变体。关键变更为集成 NeMo Skills 进行 MMMU-Pro 评估，并优化测试结果显示。尽管存在硬编码和维护风险，但显著扩展了 CI 测试覆盖，支持团队自动化性能监控。

功能与动机

PR 的主要功能是添加 GB300 硬件的 nightly benchmark 测试套件，旨在自动化 CI 任务和分摊测试负载。根据 PR body，动机源于 ' 为 GB300（4x B200 NVL4, arm64）添加 nightly benchmark 测试套件 '，并设计为 K8s CronJob 编排，以支持顺序执行和清理。Issue 评论中 Fridge003 提到 'we can split some of our CI tasks here'，表明团队希望分散 CI 压力并增强硬件特定测试。

实现拆解

实现按以下模块拆解：

1. 新增测试文件：在 test/registered/gb300/ 目录下创建 8 个文件，如 test_deepseek_v32.py，每个文件使用 ModelLaunchSettings 定义模型变体，并通过 run_combined_tests 运行精度和性能测试。代码示例：
2. 测试运行器增强：修改 accuracy_test_runner.py，新增 _run_nemo_skills_eval 函数，使用隔离 uv venv 安装 NeMo Skills，缓存数据，并处理 MMMU-Pro 评估。关键函数包括 _get_nemo_venv（创建 venv）和 _ensure_nemo_data_prepared（准备数据集）。
3. 测试结果优化：修改 run_combined_tests.py，在结果字典中添加 variant 字段，并在失败消息中显示变体名称，例如 f" Model {i + 1} ({r['model']}{variant_str}): {failed_test_str} - {error_str}"。
4. CI 套件集成：修改 run_suite.py，将 'nightly-4-gpu-gb300' 套件添加到 CUDA CI 列表中，确保 CI 系统能发现和调度这些测试。

评论区精华

review 中仅有 gemini-code-assist[bot] 的评论，聚焦于设计改进：

> "The profile_dir string is hardcoded. Consider deriving it dynamically from the test_name to reduce repetition and potential for inconsistencies."

评论在多个测试文件中重复提出，建议动态派生 profile_dir 路径，但无后续讨论或采纳，显示为一个未解决的设计权衡点。

风险与影响

风险分析：

- 硬编码 profile_dir 路径在 8 个测试文件中，若 test_name 变更或套件扩展，易导致不一致和维护负担。
- NeMo Skills 集成依赖外部库安装和网络，可能引入安装失败、版本冲突或评估超时风险。
- 测试时间设计为 7200 秒，可能影响 CI 流水线效率，尤其在并发测试时增加资源压力。
- 依赖 GB300 特定硬件，降低了测试可移植性，需额外配置以适应其他环境。

影响评估：

- 对系统：增强 CI 测试覆盖，特别是大型模型和量化变体，有助于早期发现性能回归，但增加运行开销。
- 对团队：提供标准化 GB300 基准测试框架，支持持续监控，但需管理外部依赖和长测试时间。

关联脉络

从历史 PR 看，此 PR 与多个 CI 和测试相关 PR 关联：

- PR 21534（添加 AMD MI35x 测试）：类似硬件特定 CI 扩展，反映团队跨平台测试策略。
- PR 21482（跳过 CI 非代码文件）：CI 优化举措，与本 PR 共同体现 CI 流水线的持续改进趋势。 整体上，这些 PR 显示团队在扩展测试覆盖的同时，注重 CI 效率和维护性，GB300 套件是硬件多元化测试的重要一步。