

PR #28103 完整报告

sgl-project/sglang

Add DeepSeek V4 Pro GB300 nightly and expand Kimi K25 nightly test

合并时间: 2026-06-26 03:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28103>

执行摘要

- 一句话: 新增 DeepSeek V4 Pro GB300 夜间测试并扩展 Kimi K25
- 推荐动作: 该 PR 是测试基础设施的重要补充, 展示了如何为新硬件 (GB300) 系统化地添加夜间测试, 并通过拆分子套件提高 CI 效率。建议团队关注其套件拆分模式, 在后续添加更多模型时复用。

功能与动机

PR 描述中明确提到: Add a GB300 nightly registered test for deepseek-ai/DeepSeek-V4-Pro FP4, cover the cookbook low-latency, balanced, and high-throughput single-node strategies on nightly-4-gpu-gb300, run combined GSM8K accuracy and performance profiling for each strategy。同时扩展 Kimi K2.5 测试以利用 EAGLE3 加速, 并细化 GB300 测试套件结构以便后续维护。

实现拆解

1. 新建 test/registered/gb300/test_deepseek_v4_pro_fp4.py: 定义 TestDeepSeekV4ProFp4 测试类, 包含低延迟、均衡、高吞吐三种部署策略的配置和 GSM8K 精度测试, 并注册到 nightly-4-gpu-gb300-deepseek-v4-pro-fp4 套件。
2. 修改 test/registered/gb300/test_kimi_k25_nvfp4.py: 添加 EAGLE3 投机解码参数, 将原有 TP4 和 TP4+DP4+DPA 变体升级为 TP4+EAGLE3 和 TP4+DP4+DPA+EAGLE3, 并更新 suite 注册名称为 nightly-4-gpu-gb300-kimi-k25-nvfp4。
3. 更新 test/run_suite.py: 在 NIGHTLY_SUITES 的 CUDA 条目下添加 7 个 GB300 模型专属 suite 名称 (如 nightly-4-gpu-gb300-deepseek-v4-pro-fp4 等), 从通用套件中拆分出来。
4. 调整其他 GB300 测试文件 (test_glm5_fp8.py, test_kimi_k25.py, test_qwen35_nvfp4.py, test_glm5_nvfp4.py, test_qwen35_fp8.py): 将其 register_cuda_ci 中的 suite 参数从公用 nightly-4-gpu-gb300 改为各自专属名称, 并保持原有 disabled 状态。
5. 修改 .github/workflows/nightly-test-nvidia.yml: 添加新的 GB300 nightly job, 使用专属 runner 标签和矩阵策略以匹配新增套件。

关键文件:

- test/registered/gb300/test_deepseek_v4_pro_fp4.py (模块 DeepSeek V4; 类别 test; 类型 test-coverage; 符号 TestDeepSeekV4ProFp4, test_deepseek_v4_pro_fp4) : 新增文件, 定义了 DeepSeek V4 Pro FP4 在 GB300 上的夜间测试, 包含三种策略的配置和 GSM8K 精度验证, 是整个 PR 的核心。
- test/registered/gb300/test_kimi_k25_nvfp4.py (模块 Kimi K25; 类别 test; 类型 test-coverage; 符号 TestKimiK25Nvfp4) : 修改文件, 为 Kimi K2.5 NVFP4 测试添加 EAGLE3 支持, 变更了测试变体和配置, 是 PR 的另一核心内容。
- test/run_suite.py (模块 测试套件; 类别 test; 类型 test-coverage) : 修改测试套件注册表, 将 GB300 测试拆分为多个模型专属子套件, 影响所有 GB300 夜间测试的调度。
- test/registered/gb300/test_glm5_fp8.py (模块 GLM5; 类别 test; 类型 test-coverage) : 修改 suite 注册, 将通用套件改为专属套件, 并保持 disabled 状态。
- test/registered/gb300/test_kimi_k25.py (模块 Kimi K2.5; 类别 test; 类型 test-coverage) : 修改 suite 注册, 将通用套件改为专属套件, 并保持 disabled 状态。
- test/registered/gb300/test_qwen35_nvfp4.py (模块 Qwen3.5; 类别 test; 类型 test-coverage) : 修改 suite 注册, 将通用套件改为专属套件, 并保持 disabled 状态。
- .github/workflows/nightly-test-nvidia.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 修改 CI 工作流, 添加新的 GB300 nightly job 以支持新增的测试套件。
- test/registered/gb300/test_glm5_nvfp4.py (模块 GLM5; 类别 test; 类型 test-coverage) : 修改 suite 注册, 适配新的套件命名 (变更很小)。
- test/registered/gb300/test_qwen35_fp8.py (模块 Qwen3.5; 类别 test; 类型 test-coverage) : 修改 suite 注册, 适配新的套件命名 (变更很小)。

关键符号: TestDeepSeekV4ProFp4.test_deepseek_v4_pro_fp4,
TestKimiK25Nvfp4.test_kimi_k25_nvfp4

评论区精华

该 PR 没有正式的 review 评论。仅有的 5 条 comment 均为作者 Fridge003 的 CI 运行状态更新 (如尝试不同 runner 和确认运行链接), 最终表示“先合并这个, 后续再修复 V4 Pro 测试”, 未涉及技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 主要风险: 测试依赖外部模型权重 (DeepSeek-V4-Pro 和 Kimi-K2.5-NVFP4), 下载可能超时或失败; GB300 硬件 (4x GB300 NVL4) 在 CI 中的可用性和稳定性尚不确定; 新增的 suite 注册可能导致 CI 工作流触发错误, 但测试均标记为 nightly, 不影响 PR 流程。整体风险较低。
- 影响: 对用户无直接影响。对系统: 细化了 GB300 夜间测试结构, 将原本单一的 nightly-4-gpu-gb300 拆分为 7 个模型专属套件, 提升了可维护性和错误隔离性; 为 DeepSeek V4 Pro 和 Kimi K2.5 提供了精度和性能基线。对团队: 便于单独对某个模型触发测试和调试, 降低全套件重跑成本。

- 风险标记：外部模型依赖，CI 资源消耗增加，GB300 硬件可用性

关联脉络

- 暂无明显关联 PR