

PR #28536 完整报告

sgl-project/sglang

ci: run GB300 nightly suite in the standard Nvidia nightly workflow

合并时间: 2026-06-20 03:24

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28536>

执行摘要

- 一句话: 将 GB300 夜间测试接入标准 Nvidia 工作流
- 推荐动作: 建议批准该 PR, 但需注意:
 1. 首次合并后手动触发 nightly-test-perf-4-gpu-gb300 作业验证全流程。
 2. 关注测试 timeout 风险, 若常超时增加 timeout-minutes 或拆分测试。
 3. 审查测试参数调整的准确性 (尤其是 baseline accuracy 改动)。值得学习的设计点: 针对特定硬件的测试策略从全量回归聚焦到最优量化模式 (NVFP4), 可推广到其他硬件平台。

功能与动机

GB300 夜间测试套件 (nightly-4-gpu-gb300) 自 PR #21487 添加以来, 虽已有 8 个注册测试, 但从未在 CI 工作流中调度执行。此前依赖一个无 GB300 runner 的外部集群运行 cronjob, 导致测试实际处于空转状态。现在仓库具备了 4-gpu-gb300 runner, 需要将该套件接入标准 Nvidia 夜间 CI, 使其按计划自动运行并作为质量门禁的一部分。

实现拆解

1. CI 工作流扩展: 在 .github/workflows/nightly-test-nvidia.yml 中新增 nightly-test-perf-4-gpu-gb300 作业, 使用 4-gpu-gb300 runner, 执行 run_suite.py --suite nightly-4-gpu-gb300, 并注册到 workflow_dispatch.job_filter 和 check-all-jobs.needs。
2. 测试范围精简化: 删除 test_deepseek_v32_nvfp4.py 和 test_deepseek_v32.py (DeepSeek V3.2 NVFP4/BF16), 将 test_qwen35_nvfp4.py 标记为 disabled="not needed"。
3. 剩余测试配置更新: 调整 test_glm5_nvfp4.py、test_qwen35_fp8.py 等文件的变体结构 (移除非 MTP 变体, 区分 TP/DP 的 EAGLE 参数), 更新 baseline accuracy, 并微调 performance_test_runner.py 和 run_suite.py。

关键文件:

- .github/workflows/nightly-test-nvidia.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure): 核心变更文件, 新增 GB300 夜间测试作业, 注册到工作流调度和门禁依赖。

- test/registered/gb300/test_deepseek_v32_nvfp4.py (模块 DeepSeek 测试; 类别 test; 类型 deletion; 符号 TestDeepseekV32Nvfp4, test_deepseek_v32_nvfp4) : 删除的测试文件, 代表 DeepSeek V3.2 NVFP4 模型测试被移除, 体现测试范围调整。
- test/registered/gb300/test_deepseek_v32.py (模块 DeepSeek 测试; 类别 test; 类型 deletion; 符号 TestDeepseekV32, test_deepseek_v32) : 删除的测试文件, 代表 DeepSeek V3.2 BF16 模型测试被移除, 与 NVFP4 版一同删除。
- test/registered/gb300/test_glm5_nvfp4.py (模块 GLM 测试; 类别 test; 类型 test-coverage) : 修改的测试文件, 体现了变体精简和 MTP 参数分离的设计决策。
- test/registered/gb300/test_qwen35_fp8.py (模块 Qwen 测试; 类别 test; 类型 test-coverage) : 修改的测试文件, 调整变体结构并更新 baseline accuracy, 反映测试参数演进。
- test/registered/gb300/test_qwen35_nvfp4.py (模块 Qwen 测试; 类别 test; 类型 test-coverage) : 禁用 Qwen3.5 NVFP4 测试, 体现测试策略调整。

关键符号: TestDeepseekV32Nvfp4.test_deepseek_v32_nvfp4,
TestDeepseekV32.test_deepseek_v32, TestGlm5Nvfp4.test_glm5_nvfp4,
TestQwen35Fp8.test_qwen35_fp8, TestQwen35Nvfp4.test_qwen35_nvfp4

关键源码片段

test/registered/gb300/test_qwen35_fp8.py

修改的测试文件, 调整变体结构并更新 baseline accuracy, 反映测试参数演进。

```
# 关键变更说明
# 1. COMMON_ARGS 增加 "--mamba-scheduler-strategy=extra_buffer"
# 2. 移除非 MTP 变体, 原 MTP_ARGS 拆分为 TP_MTP_ARGS 和 DP_MTP_ARGS
# TP_MTP_ARGS: --speculative-num-steps=3, --speculative-num-draft-tokens=4
# DP_MTP_ARGS: --speculative-num-steps=1, --speculative-num-draft-tokens=2
# 3. baseline_accuracy 从 0.78 降至 0.76
# 完整代码结构与 test_glm5_nvfp4.py 类似, 此处省略重复部分。
```

评论区精华

- 测试范围决策: Fridge003 明确“仅需在 GB300 上测试 NVFP4 模型, DeepSeek V3.2 的 checkpoint 可以从夜间测试中移除”。
- Qwen3.5 NVFP4 禁用确认: mmangkad 发现禁用的是 NVFP4 而非 FP8, Fridge003 解释“NVFP4 推理需求较少, FP8 在训练场景仍常用”。
- Kimi NVFP4 调整: mmangkad 添加了 token speed MLA 优化, 并移除了不支持的 native MTP。
- 测试范围决策 (design): 采纳, 删除了 DeepSeek V3.2 的两个测试文件。
- Qwen3.5 NVFP4 禁用确认 (question): Fridge003 确认主动禁用 NVFP4, 因推理需求少。
- Kimi NVFP4 参数调整 (other): 已应用到 test_kimi_k25_nvfp4.py 的提交中。

风险与影响

- 风险：
 - 测试覆盖漏洞：移除了 DeepSeek V3.2 的 BF16 和 NVFP4 测试，若后续需要回归这些模型在 GB300 上的表现，将无 CI 覆盖。此决策基于当前业务需求，但如果需求变化，需重新添加。
 - 新 CI 作业稳定性：GB300 runner 为 ARM64 架构，依赖 ci_install_dependency.sh 的 aarch64 分支处理。首次运行可能暴露环境兼容问题（如软件包依赖）。
 - 超时风险：单个 job timeout-minutes=600，按 8 个测试顺序执行（每个 est_time=7200 秒），最大耗时 16 小时，已超出 600 分钟（10 小时）。实际可能因 test_time 小于 est_time 而缩短，但仍存在超时风险。建议评估是否需增加 timeout-minutes 或拆分测试。
 - 硬件可用性：4-gpu-gb300 runner 数量有限，可能影响调度频率和可靠性。
- 影响：
 - CI 体系：GB300 夜间测试从外部 cronjob 迁移至仓库 CI，统一入口和监控，提升可观测性和可靠性。
 - 测试团队：维护者可直接在 GitHub Actions 查看 GB300 测试结果，无需访问外部集群日志。
 - 模型验证：GB300 测试范围从全模型（BF16+NVFP4）缩小为仅 NVFP4，与 Grace-Blackwell 架构的 NVFP4 高效推理定位一致。
 - 用户：无直接用户影响，但保证 GB300 平台的质量门禁有效。
 - 风险标记：测试覆盖缩小，新 CI 作业依赖特定硬件，超时风险

关联脉络

- PR #21487 Add GB300 nightly suite (tests only, no workflow): 添加了 GB300 测试套件但未接入 CI 工作流，本 PR 完成其集成。