

PR #36814 完整报告

sgl-project/sglang

xpu: move prefill-only model tests to the nightly-xpu-1-gpu grid

合并时间: 2026-08-31 14:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36814>

执行摘要

- 一句话: 将四个 XPU prefill-only 测试迁至夜间网格并去掉显存门槛
- 推荐动作: 不推荐精读, 这是一次标准的 CI 网格分层调整, 但有两个值得留意的实践: 一是把 "显存门槛" 与 "运行器规格" 绑定而不是绑定到测试本身, 迁移套件时门槛随之失效; 二是用 `nightly=True` 显式区分触发频率, 避免后续维护者误以为用例仍在 PR gate 上。对负责 XPU/Intel CI 的工程师可快速浏览。

功能与动机

PR body 明确说明: 四个测试原先挂在 `stage-b-test-1-gpu-xpu` PR 套件上, 而 `TestXPUCrossEncoderRerank` 的 `>=20GiB` 门槛只是为了防止 `fp32 + triton` 配置在约 12GiB 的 B580 PR 运行器上 OOM 或长时间挂起; `nightly-xpu-1-gpu` 运行器为 22GiB 的 B60, 因此门槛在夜间网格没有存在意义。测试计划要求夜间网格在 B60 上接住四个测试并保持通过。

实现拆解

这是一个纯测试配置调整, 按以下步骤完成:

1. 套件注册切换 (4 个文件): `test/registered/xpu/test_xpu_classification.py`、`test_xpu_embedding.py`、`test_xpu_rerank.py`、`test_xpu_reward.py` 各自将 `sglang.test.ci.ci_register` 的 `register_xpu_ci(...)` 调用从 `suite="stage-b-test-1-gpu-xpu"` 改为 `suite="nightly-xpu-1-gpu"`, `nightly=True`。这是 CI 注册入口的关键路由参数, `nightly=True` 确保用例仅进入夜间网格、不再参与 PR gate 决策。三个文件各自只有这一行改动, `rerank` 文件除注册行外还有额外清理。
2. 显存门槛删除 (仅 `test_xpu_rerank.py`): 移除模块级符号 `_xpu_total_gib()`、`_LARGE_XPU_VRAM_GIB = 20.0`、`_HAS_LARGE_XPU = _xpu_total_gib() >= _LARGE_XPU_VRAM_GIB`, 以及 `TestXPUCrossEncoderRerank` 上的 `@unittest.skipUnless(_HAS_LARGE_XPU, ...)` 装饰器和相关解释性注释。`_xpu_total_gib()` 原先在导入时通过 `torch.xpu.get_device_properties(0).total_memory` 探测显存, 迁移到固定 22GiB 的 B60 夜间运行器后这段探测逻辑成为死代码。
3. lint 修正 (提交 89f1469): 响应维护者评论, 删除 `test_xpu_rerank.py` 导入区在 `isort` 规则下多余的空行, 属于纯格式提交。

4. CI 触发与验证：第二个提交 `ci: retrigger` 强制重跑 CI；合入前 PR 状态显示 PR Test 与 Extra 两个 CI run 均为失败状态，最终由维护者合入。

整个变更不涉及 schema、部署或运行时配置，属于测试网格资源的重新分配。

关键文件：

- `test/registered/xpu/test_xpu_rerank.py`（模块 XPU 测试；类别 test；类型 test-coverage；符号 `_xpu_total_gib`）：唯一发生实质逻辑变更的文件：删除显存探测辅助函数、20GiB 门槛常量及 `skipUnless` 装饰器，并将套件切换为 `nightly-xpu-1-gpu`。
- `test/registered/xpu/test_xpu_classification.py`（模块 XPU 测试；类别 test；类型 test-coverage）：将 `classification parity` 测试从 PR 冒烟套件切换到夜间网格，仅一行注册参数变更。
- `test/registered/xpu/test_xpu_embedding.py`（模块 XPU 测试；类别 test；类型 test-coverage）：将 `embedding parity` 测试从 PR 冒烟套件切换到夜间网格，仅一行注册参数变更。
- `test/registered/xpu/test_xpu_reward.py`（模块 XPU 测试；类别 test；类型 test-coverage）：将 `reward parity` 测试从 PR 冒烟套件切换到夜间网格，仅一行注册参数变更。

关键符号：`_xpu_total_gib`

关键源码片段

`test/registered/xpu/test_xpu_rerank.py`

唯一发生实质逻辑变更的文件：删除显存探测辅助函数、20GiB 门槛常量及 `skipUnless` 装饰器，并将套件切换为 `nightly-xpu-1-gpu`。

```
# test_xpu_rerank.py —— 套件注册与 cross-encoder 测试的变更部分
```

```
from sglang.test.ci.ci_register import register_xpu_ci
```

```
# 从 stage-b-test-1-gpu-xpu (B580 PR 运行器，约 12GiB) 迁到 nightly-xpu-1-gpu
```

```
# (B60 夜间运行器，22GiB)，并显式标记 nightly=True 只进夜间网格。
```

```
register_xpu_ci(est_time=180, suite="nightly-xpu-1-gpu", nightly=True)
```

```
# cross-encoder 测试仍搭配 fp32 + triton，因为 bf16 + intel_xpu 后端在该模型上
```

```
# 与 HF 分数不完全对齐（已知偏差，单独跟踪）。
```

```
CROSS_ENCODER_MODEL_PATH = "BAAI/bge-reranker-v2-m3"
```

```
CROSS_ENCODER_TP_SIZE = 1
```

```
CROSS_ENCODER_SCORE_TOLERANCE = 1e-2
```

```
CROSS_ENCODER_TORCH_DTYPE = torch.float32
```

```
CROSS_ENCODER_ATTENTION_BACKEND = "triton"
```

```
CROSS_ENCODER_MEM_FRACTION_STATIC = 0.65
```

```
# 原 @unittest.skipUnless(_HAS_LARGE_XPU, ...) 已删除：夜间运行器为 22GiB B60，
```

```
# 不再需要 >=20GiB 门槛；类在夜间网格中无条件执行。
```

```
class TestXPUCrossEncoderRerank(CustomTestCase):
```

```
@classmethod
def setUpClass(cls):
    mp.set_start_method("spawn", force=True)
```

评论区精华

两条 issue 评论均来自合入者 mingfeima:

1. lint 修复要求: 评论指出 "need to fix lint, other than that LGTM.", 作者随后以独立提交删除 test_xpu_rerank.py 导入区多余空行, 满足 isort 检查后合入。
 2. CI 标签精简: 评论 "no need to add label run-ci-extra", 说明对这类纯测试路由调整不需要额外触发 run-ci-extra 流水线, 最终合入时未携带该标签。
- 暂无高价值评论线程

风险与影响

- 风险:
 - 回归发现延迟: 四个用例从每次 PR 必跑的冒烟套件移到每日夜间网格, 相关模型 (Qwen3-Reranker、bge-reranker-v2-m3、apeach 分类、gte-Qwen2 embedding、Skywork-Reward) 的回归只有夜间才能被发现, PR 合入门不再兜底。
 - 硬件显存假设: test_xpu_rerank.py 的 cross-encoder 测试现在无条件在夜间网格执行, 依赖夜间运行器始终保持 22GiB B60; 若日后夜间网格被换到小显存设备, fp32 + triton 配置会直接 OOM 或超时 (est_time 为 180 秒)。
 - 网格配置不一致风险: 四个文件只改了字符串与标志位, 若 nightly-xpu-1-gpu 网格上的用例收集逻辑或 suite 命名存在偏差, 测试会静默丢失; 该风险依赖 PR body 提到的 B60 实测通过来缓解。
 - 无源码路径、无性能与安全影响。
- 影响:
 - 对 CI 运行者: stage-b-test-1-gpu-xpu 冒烟套件减少 4 个用例 (合计约 540 秒估算时间), XPU PR 的反馈时间缩短; nightly-xpu-1-gpu 网格新增 4 个 prefill-only 覆盖, 且 cross-encoder rerank 从跳过变为真正执行。
 - 对 Intel XPU 团队: 测试分层逻辑更清晰——常规 PR 门禁覆盖轻量场景, 重显存场景固定由夜间 B60 承接。
 - 对最终用户: 无任何运行时功能影响, 属于内部 CI 资源调配。
 - 风险标记: PR 门禁覆盖移除, 硬件显存假设, 夜间回归延迟

关联脉络

- 暂无明显关联 PR