PR #36360 完整报告

sgl-project/sglang

[Intel XPU] Fix cross-encoder rerank hang on B580 runners

合并时间：2026-08-27 09:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36360

执行摘要

- 一句话：修复 B580 上 cross-encoder rerank 挂起，按显存跳过测试。
- 推荐动作：该 PR 值得一读，特别是对于处理跨平台 CI 中硬件差异问题的工程师。核心设计决策——按显存大小动态选择测试配置，并在不稳定配置下跳过而非降级——具有参考价值。但注意该 PR 最终采用了跳过策略，而非最初设想的 bf16 后备，因此阅读时需关注提交历史中的演进。

功能与动机

PR 描述指出 stage-b-test-1-gpu-xpu 在 B580 runner 上失败，TestXPUCrossEncoderRerank.test_prefill_logits 运行 BAAI/bge-reranker-v2-m3 于 fp32 + Triton 后端，在 SRT 引擎启动时超出内存预算，父进程等待 ready 信号永不到达，1800s 看门狗超时，重试落在同一 runner，导致整个 stage-b 作业耗时约 1h50m 后退出。B60 上同一测试通过，说明测试本身没问题，是配置不适合小显存。

实现拆解

1. 引入显存检测函数：在 test/registered/xpu/test_xpu_rerank.py 顶部新增 _xpu_total_gib() 和 _xpu_free_cache()，前者通过 torch.xpu.get_device_properties(0).total_memory 获取显存大小，后者执行 gc.collect()、torch.xpu.empty_cache() 和 torch.xpu.synchronize() 以释放 HFRunner 泄漏的 ZMQ 上下文占用的显存。
2. 按显存决定测试是否运行：定义 _LARGE_XPU_VRAM_GIB = 20.0 和 _HAS_LARGE_XPU，为 TestXPUCrossEncoderRerank 添加 @unittest.skipUnless(_HAS_LARGE_XPU, ...) 装饰器，即仅在显存 >= 20GiB 时运行 cross-encoder 测试，否则跳过。
3. 显式传递内存比例并清理显存：在 _assert_close_scores 中新增 mem_fraction_static 参数，并传递给 SRTRunner；同时在 HFRunner 关闭后、SRTRunner 启动前调用 _xpu_free_cache() 释放显存。
4. 配置常量调整：新增 CROSS_ENCODER_MEM_FRACTION_STATIC = 0.65 常量，并在调用处传入。原先的 CROSS_ENCODER_ATTENTION_BACKEND = "triton" 和 CROSS_ENCODER_TORCH_DTYPE = torch.float32 保持不变，仅用于大显存分支。
5. 更新注释说明：修改 cross-encoder 配置上方的注释，说明 fp32+triton 在 B60 上可行但 B580 OOM，且 bf16+intel_xpu 对该模型不匹配 HF，因此选择跳过而非使用错误配置。

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, _xpu_free_cache）：唯一变更文件，承载全部修复逻辑。

关键符号：_xpu_total_gib, _xpu_free_cache

关键源码片段

test/registered/xpu/test_xpu_rerank.py

唯一变更文件，承载全部修复逻辑。

```
import gc
import torch

# 读取当前 XPU 设备的总显存（单位 GiB），用于决定测试配置。
# 注意：在非 XPU 环境下返回 0.0，保证模块可导入。
def _xpu_total_gib() -> float:
    if not torch.xpu.is_available():
        return 0.0
    # 注意：total_memory 单位为字节，除以 1024**3 转换为 GiB。
    return torch.xpu.get_device_properties(0).total_memory / (1024**3)

# 在切换 HF Runner 与 SRT Runner 之间主动回收显存。
# 日志显示 HFRunner 退出时会泄漏 ZMQ 上下文，先回收再启动第二个引擎可避免叠加内存压力。
def _xpu_free_cache() -> None:
    gc.collect()
    if torch.xpu.is_available():
        torch.xpu.empty_cache()
        torch.xpu.synchronize()

# 配置阈值：大于等于 20GiB 视为大显存（如 B60 的 22GiB），否则视为小显存（如 B580 的
12GiB）。
_LARGE_XPU_VRAM_GIB = 20.0
_HAS_LARGE_XPU = _xpu_total_gib() >= _LARGE_XPU_VRAM_GIB

# 跨编码器测试：fp32 + Triton 在 B60 上通过，但 B580 上会 OOM 导致挂起；
# 而 bf16 + intel_xpu 对该模型会产生错误分数（delta ~0.84），因此直接跳过而非使用错误配置。
@unittest.skipUnless(
    _HAS_LARGE_XPU,
    "bge-reranker-v2-m3 fp32+triton OOMs on <20GiB XPU (B580); "
    "bf16+intel_xpu on this encoder produces wrong scores.",
)
class TestXPUCrossEncoderRerank(CustomTestCase):
    ...

    def _assert_close_scores(
        self,
        prompts,
        model_path,
```

```
        tp_size,
        torch_dtype,
        score_tolerance,
        attention_backend,
        mem_fraction_static, # 新增：显式传递内存比例，之前依赖默认值。
    ) -> None:
        with HFRunner(
            model_path, torch_dtype=torch_dtype, model_type="cross_encoder",
        ) as hf_runner:
            hf_scores = hf_runner.forward(prompts).scores

        # 关键：在 HF 结束后立即释放显存，避免 SRT 启动时叠加泄漏。
        _xpu_free_cache()

        with SRTRunner(
            model_path,
            tp_size=tp_size,
            torch_dtype=torch_dtype,
            model_type="cross_encoder",
            attention_backend=attention_backend,
            mem_fraction_static=mem_fraction_static, # 显式指定，保证内存预算充足。
        ) as srt_runner:
            srt_scores = srt_runner.forward(prompts).scores
```

评论区精华

无 review 评论。PR 作者在提交信息中说明：最初尝试 bf16 后备方案，但发现 bf16+intel_xpu 后端对该模型的注意力产生错误分数（delta ~0.84），因此改为直接跳过测试，避免掩盖真实问题。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 仅修改测试文件 test/registered/xpu/test_xpu_rerank.py，无生产代码变更。风险点： 1) torch.xpu API 在非 XPU 环境不可用，但已通过 torch.xpu.is_available() 保护； 2) mem_fraction_static 参数是否被 SRTRunner 正确支持需验证，如不支持可能导致参数传递错误； 3) 跳过测试可能导致 B580 上 cross-encoder 功能回归未被捕获，但 PR 说明 bf16 后备也不正确，跳过是合理权衡。
  - 影响：影响范围限于 Intel XPU 平台的 CI 测试，具体为 stage-b-test-1-gpu-xpu 套件。对 B60 (>=20GB) 而言，测试行为与之前一致（fp32+triton 仍会运行）；对 B580 (<20GB)，cross-encoder 测试被跳过，避免挂起导致的长时间 CI 失败。不影响生产代码和其他平台。
  - 风险标记：测试跳过可能掩盖回归，mem_fraction_static 参数支持依赖验证，torch.xpu API 依赖

关联脉络

- PR #35072 [XPU] Add cross-encoder rerank test (introducing PR): 该 PR 引入了 cross-encoder rerank 测试，本 PR 修复其在小显存 XPU 上的挂起问题。
- PR #35222 [CPU] Enable ERNIE models on CPU: 同为 Intel 平台相关 PR，涉及 CPU 推理支持，与本 PR 在平台支持层面有关联。