PR #26904 完整报告

sgl-project/sglang

ci(xeon): merge 2 partitions into 1 job to reduce runner contention

合并时间：2026-06-03 09:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26904

执行摘要

- 一句话：合并 Xeon CI 两个测试分区为一个 job，减少 runner 竞争
- 推荐动作：值得合并，CI 改进有实际效益，bench bug fix 也很重要。修改简洁，适合快速合入。

功能与动机

PR body 指出合并分区可以减少 runner 竞争，允许两个 PR 并行运行，提高 CI 效率。此外，bench_one_batch.py 的修复解决了在 Xeon 平台上可能出现的 input_ids 为 None 的问题。

实现拆解

1. 修改 CI workflow 文件 .github/workflows/pr-test-xeon.yml，删除第二个 partition_args 分区，合并为一个 job，并将 timeout-minutes 从 60 增加至 120，避免脚本超时。
2. 在 python/sglang/bench_one_batch.py 的 *extend* 函数中新增判断：若 batch.input_ids 为 None 且 prefill_input_ids_cpu 存在，则异步将其搬移到设备，修复 Xeon 平台上可能因 input_ids 为空导致的崩溃。
3. 更新 docker/xeon.Dockerfile，添加 ENV PATH="/opt/.venv/bin:$PATH"，确保容器内虚拟环境可执行文件在 PATH 中。
4. 更新文档 docs_new/docs/hardware-platforms/xpu.mdx，在 docker run 命令中添加 --user root 参数，提升权限兼容性。

关键文件：

- python/sglang/bench_one_batch.py（模块 benchmark；类别 source；类型 core-logic；符号 *extend*）：修复了 *extend* 函数中 prefill_input_ids_cpu 未正确搬移的问题，是本次核心 bugfix。
- .github/workflows/pr-test-xeon.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CI 核心配置变更，合并分区并调整超时。
- docker/xeon.Dockerfile（模块 Docker；类别 infra；类型 infrastructure）：添加 PATH 环境变量，确保虚拟环境在路径中。
- docs_new/docs/hardware-platforms/xpu.mdx（模块 文档；类别 other；类型 core-logic）：文档更新，添加 --user root 参数。

关键符号：*extend*

关键源码片段

python/sglang/bench_one_batch.py

修复了 *extend* 函数中 prefill_input_ids_cpu 未正确搬移的问题，是本次核心 bugfix。

```
@torch.no_grad
def extend(reqs, model_runner):
    # 创建 dummy tree_cache（无前缀缓存，仅分配）
    dummy_tree_cache = TreeCacheNamespace(
        page_size=model_runner.server_args.page_size,
        device=model_runner.device,
        token_to_kv_pool_allocator=model_runner.token_to_kv_pool_allocator,
    )

    batch = ScheduleBatch.init_new(
        reqs=reqs,
        req_to_token_pool=model_runner.req_to_token_pool,
        token_to_kv_pool_allocator=model_runner.token_to_kv_pool_allocator,
        tree_cache=dummy_tree_cache,
        model_config=model_runner.model_config,
        enable_overlap=False,
        spec_algorithm=SpeculativeAlgorithm.NONE,
    )
    batch.prepare_for_extend()
    _maybe_prepare_mlp_sync_batch(batch, model_runner)

    # 新增：如果 input_ids 为 None 但 prefill_input_ids_cpu 存在，则异步搬移到设备
    if (
        batch.input_ids is None
        and getattr(batch, "prefill_input_ids_cpu", None) is not None
    ):
        batch.input_ids = batch.prefill_input_ids_cpu.to(
            batch.device, non_blocking=True
        )
        batch.prefill_input_ids_cpu = None

    forward_batch = ForwardBatch.init_new(batch, model_runner)
    logits_output = model_runner.forward(forward_batch).logits_output
    next_token_ids = model_runner.sample(logits_output, forward_batch)
    return next_token_ids, logits_output.next_token_logits, batch
```

评论区精华

无 review 讨论，PR 直接获得批准。

- 暂无高价值评论线程

风险与影响

- 风险：CI 配置变更减少了测试并行分区数量，可能增加单次测试运行时间（但超时时间已加倍）。bench_one_batch.py 的修复较为安全，仅增加防御性检查。Docker 和文档变动影响范围小，无回归风险。
- 影响：对开发者：CI 排队时间缩短，效率提升。对 Xeon 平台：测试覆盖无实质影响，bench_one_batch 工具修复了一个潜在 bug。对文档和 Docker：提升易用性和兼容性。
- 风险标记：CI runner 配置变更，bench bugfix

关联脉络

- 暂无明显关联 PR