

PR #28607 完整报告

sgl-project/sglang

[misc] Drop redundant req_pool_indices_cpu guards; fold hisparse into GLM-5.1 e2e

合并时间: 2026-06-19 03:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28607>

执行摘要

- 一句话: 移除冗余 req_pool_indices_cpu 保护并整合 GLM-5 HiSparse 端到端测试
- 推荐动作: 值得精读。展示了如何基于上游修复进行死代码清理, 以及如何设计回归测试验证内部契约。map_last_loc_to_buffer 的 None 参数防护也值得关注。

功能与动机

28514 修复了 hisparse decode-batch builder 使其正确设置 req_pool_indices_cpu 主机镜像, 因此之前用于同步的防御性代码变为死代码。独立的 GLM-5 hisparse 端到端测试需要单独的 8GPU 服务器启动, 可以通过组合测试框架复用服务器实例, 节省 CI 时间。

实现拆解

1. 移除 schedule_batch.py 中 prepare_for_decode 和 merge_batch 内根据设备张量重建 req_pool_indices_cpu 的延迟同步代码; 同时精简 filter_batch 空批处理分支, 仅保留清空 reqs 列表, 因为 is_empty() 已能正确识别空批次。
2. 重构单元测试 test_schedule_batch_req_pool_indices.py: 删除原有针对 ScheduleBatch 方法的测试, 新增 TestHisparseDecodeBatchReqPoolCpu (验证 _build_hisparse_decode_batch 正确填充镜像) 和 TestHisparseCoordinatorReqPoolCpu (验证 map_last_loc_to_buffer 在缺少镜像时抛出 TypeError, 确保契约不被违反)。
3. 新建端到端测试文件 test_dsa_glm5_hisparse.py, 继承 DefaultServerBase 和 GSM8KMixin, 复用现有的组合测试服务器实例, 参数与旧测试一致 (8GPU、hisparse、flashmla_sparse 等)。
4. 删除旧的独立端到端测试文件 test_dsa_models_hisparse.py, 避免重复。

关键文件:

- python/sglang/srt/managers/schedule_batch.py (模块 调度器核心; 类别 source; 类型 core-logic; 符号 prepare_for_decode, filter_batch, merge_batch): 核心调度批处理逻辑, 移除了三处冗余的 req_pool_indices_cpu 保护代码, 是 PR 的主要源码变更。

- test/registered/unit/managers/test_schedule_batch_req_pool_indices.py (模块 调度批处理; 类别 test; 类型 test-coverage; 符号 TestHisparseDecodeBatchReqPoolCpu, test_build_hisparse_decode_batch_populates_req_pool_indices_cpu, TestHisparseCoordinatorReqPoolCpu, test_host_bookkeeping_requires_req_pool_indices_cpu) : 重构了单元测试, 新增对 hisparse builder 和 coordinator 的回归测试, 确保移除保护后镜像契约仍然有效。
- test/registered/models_e2e/test_dsa_glm5_hisparse.py (模块 GLM5 HiSparse; 类别 test; 类型 test-coverage; 符号 TestGLM5HiSparse) : 新增的端到端测试, 将 GLM-5 HiSparse 测试纳入组合测试框架, 避免了单独启动 8GPU 服务器。
- test/registered/8-gpu-models/test_dsa_models_hisparse.py (模块 DSA 模型测试; 类别 test; 类型 deletion; 符号 TestGLM5DPHiSparse, setUpClass, tearDownClass, test_a_gsm8k) : 删除旧的独立端到端测试, 其功能已由新的组合测试替代。

关键符号: prepare_for_decode, filter_batch, merge_batch, _build_hisparse_decode_batch, map_last_loc_to_buffer

关键源码片段

python/sglang/srt/managers/schedule_batch.py

核心调度批处理逻辑, 移除了三处冗余的 req_pool_indices_cpu 保护代码, 是 PR 的主要源码变更。

```
# python/sglang/srt/managers/schedule_batch.py
# 关键变更: prepare_for_decode 中移除 req_pool_indices_cpu 的延迟同步保护
def prepare_for_decode(self):
    self.forward_mode = ForwardMode.DECODE
    bs = len(self.reqs)
    # 移除以下保护代码:
    # if self.req_pool_indices_cpu is None and self.req_pool_indices is not None:
    # self.req_pool_indices_cpu = self.req_pool_indices.detach().cpu().to(dtype=torch.int64)
    # 原因: 上游 _build_hisparse_decode_batch 已保证 host 镜像与 device 张量同步

    # Decode embeds the last output token via embed_tokens; clear the stale
    # prefill-time tensor so it doesn't leak into ForwardBatch.
    self.input_embeds = None
    # ... 后续保持不变 ...

# filter_batch 空批处理分支极大简化
def filter_batch(
    self,
    chunked_req_to_exclude: Optional[Union[Req, List[Req]]] = None,
    keep_indices: Optional[List[int]] = None,
):
    if keep_indices is None:
        if isinstance(chunked_req_to_exclude, Req):
            chunked_req_to_exclude = [chunked_req_to_exclude]
        elif chunked_req_to_exclude is None:
```

```

        chunked_req_to_exclude = []
    keep_indices = [
        i
        for i in range(len(self.reqs))
        if not self.reqs[i].finished()
        and self.reqs[i] not in chunked_req_to_exclude
    ]

    if keep_indices is None or len(keep_indices) == 0:
        # Filter out all requests. Stale tensors are left as-is: is_empty()
        # keys off reqs, so callers drop the batch before a forward reads them.
        self.reqs = []
        return

    # ... 正常过滤分支不变 ...

```

test/registered/unit/managers/test_schedule_batch_req_pool_indices.py

重构了单元测试，新增对 hisparse builder 和 coordinator 的回归测试，确保移除保护后镜像契约仍然有效。

```

# test/registered/unit/managers/test_schedule_batch_req_pool_indices.py
# 新增测试：验证 _build_hispase_decode_batch 正确填充 req_pool_indices_cpu 镜像
class TestHispaseDecodeBatchReqPoolCpu(unittest.TestCase):
    def test_build_hispase_decode_batch_populates_req_pool_indices_cpu(self):
        # _build_hispase_decode_batch 基于正常 extend 路径构造 ScheduleBatch,
        # 因此它必须保持 req_pool_indices_cpu 主机镜像与设备张量同步。
        # 缺少镜像会导致 hisparse 解码时的 bookkeeping 崩溃
        # (map_last_loc_to_buffer -> _grow_device_buffers 索引 req_pool_indices_cpu)。
        scheduler = Scheduler.__new__(Scheduler)
        scheduler.device = "cpu"
        scheduler.req_to_token_pool = types.SimpleNamespace(device="cpu")
        scheduler.token_to_kv_pool_allocator = None
        scheduler.tree_cache = None
        scheduler.model_config = types.SimpleNamespace(
            is_encoder_decoder=False, vocab_size=32
        )
        scheduler.enable_overlap = False
        scheduler.spec_algorithm = types.SimpleNamespace(is_none=lambda: True)
        scheduler.future_map = MagicMock()

        reqs = [
            _make_req(req_pool_idx=4, origin_input_ids=[1, 2, 3], output_ids=[7]),
            _make_req(req_pool_idx=9, origin_input_ids=[1, 2], output_ids=[8]),
        ]

        with patch(
            "sglang.srt.managers.scheduler.SamplingBatchInfo.from_schedule_batch",
            return_value=MagicMock(),
        ):

```

```
batch = scheduler._build_hispase_decode_batch(reqs)

# 断言镜像与设备张量一致，而不是硬编码输入值的拷贝
self.assertIsNotNone(batch.req_pool_indices_cpu)
self.assertTrue(
    torch.equal(batch.req_pool_indices_cpu, batch.req_pool_indices.cpu())
)
```

评论区精华

PR 未产生 review 评论。作者通过 issue 评论触发两次 rerun 测试（[test_schedule_batch_req_pool_indices.py](#) 和 [test_dsa_glm5_hispase.py](#) 均通过），验证了变更的正确性。

- 暂无高价值评论线程

风险与影响

- 风险：核心风险是被移除的保护代码在其他未覆盖路径中仍有依赖。但： 1) 上游 #28514 已确保 hisparse builder 总设置镜像； 2) prepare_for_decode 和 merge_batch 的移除处是防御性回退，而非主要逻辑； 3) filter_batch 空处理精简后，is_empty() 检查确保调用者不会访问空张量。新增的回归测试覆盖 builder 和 coordinator 的镜像契约，风险可控。
- 影响：对用户无直接影响。对团队：CI 效率提升（合并 e2e 测试减少额外服务器启动）；代码可维护性提升（死代码移除）；测试覆盖更精准（从模拟测试转向集成测试 + 契约测试）。
- 风险标记：依赖上游修复，测试覆盖重构

关联脉络

- PR #28514 Fix hisparse decode-batch builder sets req_pool_indices_cpu: 前置修复，确保 hisparse builder 正确设置 req_pool_indices_cpu，使本 PR 的清理成为可能。