PR #24366 完整报告

sgl-project/sglang

[diffusion] Use direct all-to-all for USP collectives

合并时间：2026-05-05 00:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24366

执行摘要

- 一句话：直连 all-to-all 替代功能集合，提升 diffusion 去噪速度约 18%
- 推荐动作：值得精读，理解 PyTorch 函数式集合与直接集合的性能差异。建议接受 reviewer 关于显式导入 torch.distributed 的建议以提高代码健壮性。

功能与动机

LTX 2.3 one-stage TI2V 在 PyTorch 2.11 升级后出现性能回归，profile 定位到 USP 集合开销而非 FA3 attention 计算。USP 在每个去噪步骤中对 48 个 transformer 块和多个指导通道调用该集合，函数式集合包装的累积开销导致延迟升高。

实现拆解

1. 修改 python/sglang/multimodal_gen/runtime/layers/usp.py 中的 _usp_all_to_all_single 函数：将原来 ft_c.all_to_all_single 加 _maybe_wait 的模式改为直接预分配输出张量并调用 torch.distributed.all_to_all_single。
2. 移除 _maybe_wait 的调用，因为直接 all_to_all_single 返回的是同步张量。
3. 在 flatten 后增加 .contiguous() 调用，确保输入张量连续以满足直接 all-to-all 的要求。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/usp.py（模块 USP 集合；类别 source；类型 core-logic；符号 _usp_all_to_all_single, _maybe_wait）：修改了核心 USP all-to-all 函数，用直接集合调用替换功能集合包装，提升热路径性能。

关键符号：_usp_all_to_all_single

关键源码片段

python/sglang/multimodal_gen/runtime/layers/usp.py

修改了核心 USP all-to-all 函数，用直接集合调用替换功能集合包装，提升热路径性能。

```
def _usp_all_to_all_single(x: torch.Tensor) -> torch.Tensor:
    ulysses_pg = get_sp_group().ulysses_group
    assert ulysses_pg is not None, "Ulysses process group is not initialized."
    x_shape = x.shape
    # 确保输入连续，直接 all-to-all 要求连续张量
```

```
    x = x.flatten().contiguous()
    output = torch.empty_like(x)
    # 直接调用 torch.distributed.all_to_all_single 而非功能集合版本，
    # 避免每次调用都创建 AsyncCollectiveTensor 和后续 wait 的开销。
    # 该热路径在每个去噪步骤中调用数百次，累计加速约 18%。
    torch.distributed.all_to_all_single(output, x, group=ulysses_pg)
    return output.reshape(x_shape)
```

评论区精华

机器人 reviewer gemini-code-assist[bot] 建议显式 import torch.distributed as dist 以确保可靠性和代码风格一致性，当前通过 torch.distributed._functional_collectives 的副作用加载可能不健壮。该建议尚未被采纳或讨论。

- 显式导入 torch.distributed (style): 未被采纳，也无人回复讨论。

风险与影响

- 风险：显式 import torch.distributed 缺失：当前依赖 ft_c 模块的副作用加载 torch.distributed，在某些环境可能失败。建议按 reviewer 建议添加导入。无其他回归风险，因为功能语义完全等价。
- 影响：影响范围小：仅修改一个文件中的单一函数（_usp_all_to_all_single），该函数仅被 diffusion USP 相关代码调用。性能提升显著（~18%），且无功能变化。不影响非 diffusion 或其他序列并行策略。
- 风险标记：未处理代码风格建议

关联脉络

- PR #23593 [Docker] Prep for torch 2.11: cu129 fix, image validator, dep cleanup: 本 PR 的回归是由 torch 2.11 升级引入的，该 PR 是 torch 2.11 升级的相关准备。