PR #25750 完整报告

sgl-project/sglang

fix(dsv4): make pool configurator PP-aware

合并时间：2026-05-20 10:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25750

执行摘要

- 一句话：修复 DSV4 PP 下内存池配置 over-counting
- 推荐动作：值得精读，可作为 Pipeline Parallelism 下按层分配资源的典型修复案例。改动虽小，但体现了对分布式系统中 rank 局部性理解的重要性，以及与 TokenToKVPool 保持一致的契约意识。

功能与动机

DSV4 模型在 Pipeline Parallelism 下，每个 PP rank 的 DSV4PoolConfigurator 使用全局的 compress_ratios 计算内存池大小，导致 bytes_per_full_token 被重复计算（乘了 pp_size），内存分配严重 over-count，可能引发 OOM 或性能下降。

实现拆解

1. 在 DSV4PoolConfigurator.__init__（python/sglang/srt/model_executor/pool_configurator.py，第 322-329 行）中，将原来的 self.compression_ratios = cfg.compress_ratios 改为 self.compression_ratios = cfg.compress_ratios[mr.start_layer : mr.end_layer]，只截取当前 PP rank 负责的层切片。
2. 当 mr.pp_size > 1 时，增加一条日志输出，记录当前 rank 编号、层范围以及局部层数与全局层数的对比，便于调试和监控。
3. 后续依赖该切片进行的算层数（num_layers_total/num_layers_ca4/num_layers_ca128）以及推测解码的放大因子计算自动修正，无需额外改动。

关键文件：

- python/sglang/srt/model_executor/pool_configurator.py（模块 池配置器；类别 source；类型 data-contract）：唯一变更文件，在 DSV4PoolConfigurator 初始化中截取 compress_ratios 为 PP rank 局部切片，修复 over-counting。

关键符号：DSV4PoolConfigurator.init

关键源码片段

python/sglang/srt/model_executor/pool_configurator.py

唯一变更文件，在 DSV4PoolConfigurator 初始化中截取 compress_ratios 为 PP rank 局部切片，修复 over-counting。

```
# python/sglang/srt/model_executor/pool_configurator.py, DSV4PoolConfigurator.__init__

class DSV4PoolConfigurator(MemoryPoolConfigurator):
    """Configurator for DSV4 compressed-attention models."""

    def __init__(self, mr: ModelRunner):
        cfg = mr.model_config
        self.qk_nope_head_dim = cfg.qk_nope_head_dim
        self.qk_rope_head_dim = cfg.qk_rope_head_dim
        self.indexer_head_dim = cfg.index_head_dim
        # PP-local slice; matches DeepSeekV4TokenToKVPool's stage_ratios.
        self.compression_ratios = cfg.compress_ratios[mr.start_layer : mr.end_layer]
        if mr.pp_size > 1:
            logger.info(
                f"DSV4 pool PP slice: rank={mr.pp_group.rank_in_group} "
                f"layers=[{mr.start_layer},{mr.end_layer}) "
                f"local={len(self.compression_ratios)}/{len(cfg.compress_ratios)}"
            )
        self.swa_page_size = cfg.window_size
        self.swa_ratio = mr.server_args.swa_full_tokens_ratio
        self.is_speculative = mr.server_args.speculative_algorithm is not None
        # ... 后续逻辑依赖 self.compression_ratios 自动修正
```

评论区精华

PR 没有显著的讨论，仅由作者发起并说明需要等待 PR #25729 合并。两位 reviewer（yhyang201 和 ShangmingCai）直接 approve。

- PR 依赖关系 (other): PR #25729（修复 DSV4 大 PP 下 forward metadata 竞态）已合并，本 PR 可安全合入。

风险与影响

- 风险：风险较低。变更仅对 PP 场景生效，且与 DeepSeekV4TokenToKVPool 的 stage_ratios 逻辑对齐；非 PP 场景下 mr.start_layer=0 且 mr.end_layer=len(cfg.compress_ratios)，行为完全不变。但也意味着非 PP 场景未经充分测试。
- 影响：影响范围：DSV4 模型启用 Pipeline Parallelism（PP>1）后的内存池配置。修复了对 bytes_per_full_token 的 over-counting，理论上会显著降低每个 PP rank 分配的内存池大小，避免 OOM。对非 PP 用户无影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #25729 fix(dsv4): upgrade forward metadata on main stream for large PP size: 作者说明本 PR 需要等 #25729 合并后再合入，解决大 PP 下的竞态问题。