PR #49033 完整报告

vllm-project/vllm

Revert "[Sampler] Stop upcasting logits to fp32 in apply_sampling_params" (#48641)

合并时间：2026-07-21 16:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49033

执行摘要

- 一句话：回滚采样器 fp32 上转优化，修复 MTP 超时
- 推荐动作：该 PR 是标准回滚，但值得关注其揭示的采样精度敏感性：spec-decode 在低精度下数值行为微妙。建议阅读原始 PR #48641 及其回滚讨论，以了解设计权衡。

功能与动机

原始 PR #48641 引入了 CI 失败：LM Eval Qwen3.5 Models (2xB200) 中 MTP 变体在 sample_tokens 中挂起（RPC 超时）。其替换 fp32 副本为原地 bf16 更改，可能破坏 spec-decode 验证数值，导致 DP=2 排名不同步。回滚恢复已知正确的 fp32 副本行为，确保采样正确性。

实现拆解

回滚在 5 个文件中还原了 #48641 的所有变更。具体包括：

1. sampler.py：apply_sampling_params 中恢复强制创建 fp32 副本（通过 torch.empty_like(..., dtype=torch.float32).copy_），移除 returns_logprobs 方法，将其逻辑内联到 __call__ 中。
2. logit_bias.py：Triton kernel 中移除 .to(tl.float32) 的调用被恢复，偏置累加重回 fp32。
3. topk_topp_triton.py：所有 _topk_topp_kernel 中移除的 .to(tl.float32) 调用重新加入，确保 top-k/top-p 计算在 fp32 下进行。
4. topk_topp_sampler.py：apply_top_k_top_p_pytorch 恢复 softmax(dtype=torch.float32)；flashinfer_sample 恢复对 logits 调用 .float().contiguous()。
5. 测试文件：删除 test_flashinfer_sample_padded_vocab 测试（该测试针对 #48641 新增的 dtype 保留行为）。

关键文件：

- vllm/v1/worker/gpu/sample/sampler.py（模块 采样器；类别 source；类型 core-logic；符号 returns_logprobs, apply_sampling_params, call）：核心采样实现，回滚了 apply_sampling_params 中的 fp32 副本移除和 returns_logprobs 内联。
- tests/v1/sample/test_topk_topp_sampler.py（模块 采样测试；类别 test；类型 test-coverage；符号 test_flashinfer_sample_padded_vocab）：删除了针对 #48641 新增的 test_flashinfer_sample_padded_vocab 测试，回滚测试覆盖。

- vllm/v1/worker/gpu/sample/logit_bias.py（模块 采样器；类别 source；类型 core-logic）：Triton kernel 中恢复显式 to(tl.float32) 调用，确保偏置累加在 fp32 中进行。
- vllm/v1/sample/ops/topk_topp_triton.py（模块 采样算子；类别 infra；类型 infrastructure）：Triton kernel 中批量恢复 .to(tl.float32) 转换，确保 top-k/top-p 计算在 fp32 下进行。
- vllm/v1/sample/ops/topk_topp_sampler.py（模块 采样算子；类别 infra；类型 infrastructure）：回滚了 apply_top_k_top_p_pytorch 中的 softmax dtype 简化以及 flashinfer_sample 中的 float 转换简化。

关键符号：apply_sampling_params, returns_logprobs, _bias_kernel, apply_top_k_top_p_triton, apply_top_k_top_p_pytorch, flashinfer_sample

关键源码片段

vllm/v1/worker/gpu/sample/sampler.py

核心采样实现，回滚了 apply_sampling_params 中的 fp32 副本移除和 returns_logprobs 内联。

```
def apply_sampling_params(
    self,
    logits: torch.Tensor,
    expanded_idx_mapping: torch.Tensor,
    idx_mapping_np: np.ndarray,
    pos: torch.Tensor,
    input_ids: torch.Tensor,
    expanded_local_pos: torch.Tensor,
    skip_top_k_top_p: bool = False,
) -> torch.Tensor:
    # 创建 FP32 副本：回滚 #48641，恢复始终复制并转换为 fp32 的行为
    # 确保下游操作（softmax、top-k/top-p 等）在 fp32 下运行，
    # 避免 bf16 精度溢出导致采样分布异常或 spec-decode 数值错配。
    logits = torch.empty_like(logits, dtype=torch.float32).copy_(logits)

    # 应用 logit 偏置（例如 allowed_token_ids、min_tokens）原地操作
    self.logit_bias_state.apply_logit_bias(
        logits, expanded_idx_mapping, idx_mapping_np, pos
    )

    # 应用惩罚（例如重复惩罚）原地操作
    self.penalties_state.apply_penalties(
        logits, expanded_idx_mapping, idx_mapping_np, input_ids, expanded_local_pos,
    )

    # 应用坏词掩码原地操作
    self.bad_words_state.apply_bad_words(
        logits, expanded_idx_mapping, ...
    )
    # 继续后续采样步骤（温度、top-k/top-p 等）
```

```
    return logits
```

评论区精华

mgoin（原始 PR 作者）批准回滚：'Let's revert my PR for now due to increasing uncertainty of the changes in sampling behavior'。njhill 最初评论 '#48641 不太可能是失败原因 ' 后删除该评论，可能随后确认了关联性。

- CI failure attribution and revert decision (correctness): 一致同意回滚，由 njhill 合并。

风险与影响

- 风险：回滚风险低：恢复到经过生产验证的原始行为。但会重新引入 #48641 解决的内存开销：对于 spec-decode 场景（如 Qwen3-8B + dflash 推测器），峰值激活内存增加约 4 GiB，可能导致 OOM。此外，logits 保持在 fp32 可能略微增加显存带宽压力，但不会影响正确性。
- 影响：用户影响：使用 spec-decode 且 vocab 较大的模型可能遇到更高的显存使用；但先前 CI 中超时的 MTP 模型（Qwen3.5-397B-A17B-NVFP4）将正常工作。系统影响：采样流程恢复到经过全面测试的正确行为，数值稳定性得到保证。团队影响：需评估是否需要以不同方式重新实现 #48641 的优化（例如保留 fp32 副本但只对必要请求创建）。
- 风险标记：恢复内存峰值，采样数值敏感带

关联脉络

- PR #48641 [Sampler] Stop upcasting logits to fp32 in apply_sampling_params: 此 PR 回滚了 #48641 的全部变更