PR #44899 完整报告

vllm-project/vllm

[ROCm][DSv4][Perf] Flash-decode split-K decode attention kernel

合并时间：2026-06-12 11:17

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44899

执行摘要

该 PR 在 AMD gfx950 上为 DeepSeek-V4 的稀疏解码注意力引入 flash-decode split-K pipeline，通过将单 kernel 拆分为 partial + reduce 双 kernel 并增加 split 维度并行度，在低 batch 时延迟降低 8 倍（从 277us 到 35us），端到端 TPOT 改善 20%。仅影响 ROCm gfx950 用户，其他硬件保持不变。

功能与动机

PR body 明确指出：DSV4 的 16 个 TP-local head 在 decode 时合并为单个 head-block，导致整个解码只使用 batch 个工作组，使 256-CU 的 gfx950 大部分空闲，延迟在 batch 4-128 区间几乎不变（~277us）。为解决此 occupancy 瓶颈，采用 flash-decode split-K 策略将 KV 序列切分成多个 split，每个 split 独立计算部分 softmax 后再 reduce，从而充分占用 GPU 资源。

实现拆解

1. 核心 kernel 新增：在 rocm_aiter_mla_sparse.py 中添加 _sparse_attn_decode_partial_kernel（partial）和 _sparse_attn_decode_reduce_kernel（reduce）。Partial kernel 以 (batch, split, head) 三维启动，每个 split 处理 query 的一段 KV；reduce kernel 合并各 split 的 partial m_i/l_i/acc。
2. 分片启发式：实现 _decode_num_splits，输入 query 数、平均 cache 长度等，基于设备 CU 数（_decode_cu_count）计算最优 split 数（1~16），目标是最小化估计延迟。
3. 条件分支：在入口函数 _sparse_attn_decode_ragged 中加入 _on_gfx950() 条件，gfx950 且 split 数 > 1 时走新路径，否则原路径。其他架构完全无影响。
4. 辅助函数：_decode_cu_count 通过 torch.cuda 属性获取 GPU CU 数量；_decode_partial_iters 计算迭代相关参数。
5. 测试配套：在 test_rocm_triton_attn_dsv4.py 增加两个测试：test_decode_num_splits_heuristic 不依赖 GPU 验证启发式，test_sparse_attn_decode_split_k_kernel 参数化测试在 gfx950 上验证精度。使用 requires_gfx950 装饰器跳过非 gfx950 环境。

以下展示 split-K 中决定 split 分摊的核心计算：

```
main_len = main_end - main_start
main_chunk = (main_len + NUM_SPLITS - 1) // NUM_SPLITS
main_lo = split_id * main_chunk
```

```
main_hi = tl.minimum(main_lo + main_chunk, main_len)
```

每个 split 只处理自己负责的 (main_lo, main_hi) 范围内的 KV 条目，实现并行化。

评论区精华

Review 中 tjtanaa 要求更新单元测试文件，作者随后添加了 gfx950-only 的 test_sparse_attn_decode_split_k_kernel 和 test_decode_num_splits_heuristic。无其他争议。

风险与影响

- 技术风险：精度风险通过内核级测试和 GSM8K 任务验证（exact match 0.9568）证明无精度退化；性能风险高 batch 时 over-split 可能反效果，但启发式限制 split 数上限 16，且测试覆盖了 1/2/3/4/8 等分片数；维护风险新增大量 Triton 代码可能增加后续修改成本。
- 影响范围：仅限于 ROCm gfx950 + DeepSeek-V4 组合，低 batch 收益显著（TPOT -20%），高 batch 约 9%。对其他硬件和模型无影响。

关联脉络

本 PR 是针对 DeepSeek-V4 注意力优化的系列工作之一。历史 PR #45052 修复了同一模型 FlashMLA 测试，奠定了测试基础。未来可能将 split-K 策略推广到其他架构或模型。