PR #47945 完整报告

vllm-project/vllm

[ROCm] Add tuned selective_state_update float16 config for AMD Instinct MI300X

合并时间：2026-07-08 19:03

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47945

执行摘要

为 AMD Instinct MI300X 添加 selective_state_update (Mamba SSU decode kernel) 的 float16 调优启动配置，消除运行时回退到次优通用启发式策略的性能损失。新增一个 JSON 配置文件即可带来 1.18x–2.13x 的 kernel 加速，无需任何代码修改。

功能与动机

vLLM 已为 NVIDIA 各 GPU (B200, GB200, H100, H200, RTX PRO 6000) 和 MI355 提供了 tuned selective_state_update 配置，但 MI300X 缺少对应文件。在 MI300X 上，get_ssm_configs() 找不到匹配配置，回退到通用启发式策略 _get_default_ssm_launch_config()，该策略对于这些形状的 (BLOCK_SIZE_M, num_warps) 远非最优，导致 Mamba decode 吞吐量未达预期。本 PR 旨在通过添加调优配置，释放 MI300X 上 Mamba 模型的性能潜力。

实现拆解

1. 生成配置：在 MI300X 上运行 vLLM 内置基准脚本 benchmark_selective_state_update，使用参数 --dstate 128 --dtype float16 --mamba-ssm-cache-dtype float16 --save-configs --compare --validate。该脚本自动扫描有效批量范围 (128–262144) 并记录每个批量下的最优 (BLOCK_SIZE_M, num_warps) 组合。
2. 添加文件：将生成的 JSON 配置放置在目录 vllm/model_executor/layers/mamba/ops/configs/selective_state_update/ 下，文件名为 headdim=64,dstate=128,device_name=AMD_Instinct_MI300X,cache_dtype=float16.json。文件名中的设备名和 dtype 由脚本自动生成，因此无需任何代码修改即可被 get_ssm_configs() 在运行时自动识别和加载。bf16 共用同一配置（加载器会将 bf16 规范化为 float16，因为 kernel 只关注状态位宽）。
3. 验证正确性：--validate 确保所有 12 个 tuned effective_batch 点均通过 CPU 参考基准；--compare 显示一致的加速效果。该配置文件仅影响 Triton launch geometry，不改变 kernel 数学逻辑，因此模型输出不受影响。

vllm/model_executor/layers/mamba/ops/configs/selective_state_update/headdim=64,dstate=128,device_name=AMD_Instinct_MI300X,cache_dtype=float16.json

唯一变更文件，新增 MI300X 的 float16 selective_state_update 调优配置，包含从 batch 128 到 262144 共 12 个（BLOCK_SIZE_M, num_warps）配置对。直接决定 kernel 并行效率。

```
{
  "triton_version": "3.4.0",
  // 批量大小作为键，值为 Triton launch 几何参数
  "128": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 8
  },
  "256": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 4
  },
  "1024": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 1 // 小批量时 warp 数少，提高占用
  },
  // ... 中间配置省略 ...
  "65536": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 8 // 大批量时增加 warp，提升并行度
  },
  "262144": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 4
  }
}
```

评论区精华

无人工 reviewer 的实质性讨论。作者在 issue 评论中提供了详细的性能数据表格，展示各有效批量下的加速比：batch=128 为 1.18x，batch=4096 为 1.60x，batch=131072 为 2.10x，batch=262144 为最大 2.13x。这些数据支撑了合入决策。

风险与影响

- 风险：极低。仅新增 JSON 配置文件，不改变任何源代码逻辑。配置仅控制 kernel launch geometry，数学运算不变，正确性已通过验证。
- 影响：对在 MI300X 上运行 Mamba 模型（如 Mamba2）的用户，kernel 解码吞吐可提升 1.18x–2.13x。对非 Mamba 模型无影响。对其他 AMD GPU（如 MI325X、MI308X）仍缺少配置，待后续 PR 补充。

关联脉络

- 本 PR 与 #47767（MI355 float32 配置）和 #47947（MI300X float32 配置）属于同一工作线，逐步覆盖 AMD GPU 系列。
- 作者在 PR body 中明确指出 float32 状态缓存配置将单独提交，后续可关注 #47947 的合入。