PR #50006 完整报告

vllm-project/vllm

[ROCm] Add tuned selective_state_update float16 config for AMD Instinct MI325X

合并时间：2026-07-31 10:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50006

执行摘要

本次 PR 为 AMD Instinct MI325X 增加了 Mamba selective_state_update（SSU）内核的 float16 调优启动配置（headdim=64, dstate=128），填补了该设备配置缺失的空白。配置通过内置脚本自动生成并验证，在解码批量较大时带来约 1.3-1.95 倍的性能提升，且不影响模型精度。改动仅为新增一个 JSON 配置文件，风险极低，值得作为设备特定调优的参考范式。

功能与动机

MI325X 此前缺少 SSU 调优配置，运行时回退到通用启发式（BLOCK_SIZE_M=4, num_warps=4），在 dstate=128 时对所有批量大小均非最优，制约 Mamba 解码吞吐。此 PR 补齐配置，与 MI300X、MI350、MI355 的既有做法保持一致，利用分层配置查找机制（#48980）实现无缝加载。

实现拆解

1. 新增配置文件：在 vllm/model_executor/layers/mamba/ops/configs/selective_state_update/ 下添加 headdim=64,dstate=128,device_name=AMD_Instinct_MI325X,cache_dtype=float16.json，内含 12 个 effective_batch 点的 BLOCK_SIZE_M 与 num_warps 设定。
2. 自动生成与命名：配置由 benchmark_selective_state_update.py 工具生成，文件名根据设备名称与 dtype 自动推导，可直接放入配置目录。
3. 运行时加载：通过 #48980 引入的分层配置路径，运行时会自动拾取该文件，并记录日志确认加载。
4. 验证：所有 12 个调优点通过 CPU 参考验证，单元测试 6 项通过，lm_eval 端到端精度与基线一致，仅浮点归约顺序不同。

vllm/model_executor/layers/mamba/ops/configs/selective_state_update/headdim=64,dstate=128,device_name=AMD_Instinct_MI325X,cache_dtype=float16.json

核心变更，为 MI325X 添加 SSU 浮点 16 调优启动配置，直接提升解码性能。

```
{
  "triton_version": "3.6.0",
  // 按 effective_batch（batch × nheads）键控的调优配置
  "128": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 4
```

```
  },
  "256": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 4
  },
  // 大 batch 时偏向较大 BLOCK_SIZE_M 和较少 warps，以提升吞吐
  "4096": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 1
  },
  "8192": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 1
  },
  "16384": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 1
  },
  "32768": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 1
  },
  "65536": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 4
  },
  "131072": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 4
  },
  "196608": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 4
  },
  "262144": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 4
  }
}
```

评论区精华

- claude[bot]：This pull request is from a fork — automated review is disabled.（fork 导致自动审查禁用）
- tjtanaa：直接批准，无具体评论。

风险与影响

风险较低。配置仅改变 Triton 启动几何，不影响内核数学，精度验证确认无显著变化。唯一需注意 Triton 版本记录为 3.6.0，不同版本下性能可能略有差异，但不影响正确性。影响范围限于 MI325X 上 Mamba 模型解码性能，对用户透明，无需 API 变更。

关联脉络

本 PR 是对 AMD 设备 SSU 调优配置系列（MI300X、MI350、MI355）的补充，依赖 #48980 的分层配置查找机制。它体现了 vLLM 为各硬件平台提供设备特定内核调优配置的演进方向，未来可推广到更多设备与内核。