PR #50007 完整报告

vllm-project/vllm

[ROCm] Add tuned selective_state_update float32 config for AMD Instinct MI325X

合并时间：2026-08-08 00:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50007

执行摘要

本 PR 为 AMD Instinct MI325X 添加了一个调优后的 Mamba selective_state_update 内核 launch 配置，覆盖 cache_dtype=float32 这一 vLLM 默认路径。通过 12 个 effective_batch 点的显式 BLOCK_SIZE_M/num_warps 选择，替代此前回退的通用启发式，mid decode 范围可获得最高 2.27x 加速，且经 lm_eval 验证精度完全一致。变更仅为一个 JSON 配置文件，无代码修改，风险极低。

功能与动机

PR body 明确指出，MI325X 在 get_ssm_configs() 中找不到匹配的 SSU 配置，只能回退到 _get_default_ssm_launch_config() 的通用启发式 (BLOCK_SIZE_M, num_warps) = (4, 4)，这与 dstate=128 的最优几何相差甚远，导致 Mamba decode 吞吐受损。此前 MI300X (#47947)、MI355 (#47943, #48373)、MI350 (#48159) 都已有对应配置，MI325X 缺失。本 PR 补齐了这一缺口，且 cache_dtype=float32 是 vLLM 默认的 Mamba2 状态缓存类型，因此收益覆盖大多数部署。

实现拆解

1. 生成配置：使用仓库内置调参脚本 benchmarks.kernels.benchmark_selective_state_update，在单张 MI325X（gfx942）上以 --dstate 128 --dtype float16 --mamba-ssm-cache-dtype float32 启动扫描，自动寻找每个 effective_batch 点最优的 BLOCK_SIZE_M 和 num_warps。
2. 落盘文件：脚本按设备名和 dtype 自动生成 headdim=64,dstate=128,device_name=AMD_Instinct_MI325X,cache_dtype=float32.json，包含 12 个 effective_batch 键（128 到 262144）。
3. 运行时加载：文件放入 vllm/model_executor/layers/mamba/ops/configs/selective_state_update/ 后，get_ssm_configs() 会按 (headdim, dstate, device_name, cache_dtype) 自动查找并加载，无需任何代码改动。
4. 配套验证：--validate 12/12 点通过 CPU 参考；--compare 显示 mid decode 范围最高 2.27x 加速；tests/kernels/mamba/test_mamba_ssm_configs.py 6 个测试通过；lm_eval lambada_openai 精度与启发式一致（acc 0.7040, ppl 4.2693 vs 4.2698）。

关键源码片段

该 PR 的核心是一个 JSON 配置文件，它定义了 MI325X 上 Mamba SSU 内核在不同 effective_batch 下的 Triton launch 几何。以下为完整文件内容：


```
{
  "triton_version": "3.6.0",
  "128": {
    "BLOCK_SIZE_M": 8,
    "num_warps": 4
  },
  "256": {
    "BLOCK_SIZE_M": 8,
    "num_warps": 4
  },
  "1024": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 1
  },
  "2048": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 1
  },
  "4096": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 1
  },
  "8192": {
    "BLOCK_SIZE_M": 16,
    "num_warps": 1
  },
  "16384": {
    "BLOCK_SIZE_M": 32,
    "num_warps": 8
  },
  "32768": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 2
  },
  "65536": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 4
  },
  "131072": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 1
  },
  "196608": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 2
  },
```

```
  "262144": {
    "BLOCK_SIZE_M": 64,
    "num_warps": 1
  }
}
```

评论区精华

- dllehr-amd 第一次 approve 时提出："Just waiting on confirmation that the LM_EVAL was able to run with the float32 state so we know that it ran the right configs"——要求确认验证确实覆盖了 float32 路径。
- vanshbhatia-amd 给出了完整的 lm_eval 命令、启动日志中的 mamba_ssm_cache_dtype = float32 以及 get_ssm_configs(64, 128, "float32") 加载 12 个键的证据，确认无误。
- 作者随后 @tdoublep @tomeras91 请求合并，并向 dllehr-amd 询问 merge 时间；最终 hongxiayang 运行 /ci run 并 approve。

风险与影响

- 风险：配置仅覆盖 headdim=64, dstate=128, cache_dtype=float32 组合，其他 dtype 或 headdim 仍走启发式；配置基于 Triton 3.6.0 生成，未来版本升级可能使最优参数漂移，但文件内 triton_version 字段可追踪；单测未覆盖 MI325X 硬件，依赖手工验证。
- 影响：仅影响 AMD MI325X 上 Mamba2 模型的 decode 性能，effective_batch 1024-8192 范围提升 1.49-2.27x，大 batch 也有 1.2-1.4x 提升。对用户无行为变化，无需修改调用方，风险极低。

关联脉络

本 PR 是 vLLM 为不同 AMD GPU 补齐 Mamba SSU 调优配置的系列工作之一，PR body 中引用了 MI300X (#47947)、MI355 (#47943, #48373)、MI350 (#48159) 以及引入 bundled-config lookup 的 #48980。这一系列配置与 NVIDIA（B200、GB200、H100、H200 等）的做法一致，体现了 vLLM 按设备、按 dtype 精细调优 kernel launch 的通用模式。未来若新增其他 GPU 或 dtype，可沿用同样的调参脚本和验证流程。