

PR #48372 完整报告

vllm-project/vllm

[ROCm] Retune MI355 selective_state_update float16 config on the unified effective_batch grid

合并时间: 2026-07-14 13:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48372>

执行摘要

本 PR 重新调整了 AMD Instinct MI355X 上 Mamba selective_state_update (SSU) 内核的 float16 缓存配置，将键从窄网格 (10-5120) 改为统一网格 (128-262144)，与 MI300X 和 MI350X 保持一致。仅修改 JSON 配置文件，无代码逻辑变更，但带来 1.2-2.3 倍的内核加速。

功能与动机

所有 AMD Instinct GPU 的 SSU 配置此前使用不同的采样网格，不利于性能对比和维护。MI355X 原始配置 (#47767) 为特定部署调优，使用了 10-5120 的非标准网格。本 PR 将其重新生成到与 MI300X (#47945) 和 MI350X (#48159) 相同的标准网格 (128-262144)，确保整个产品线的配置在统一点上采样。

实现拆解

1. 配置网格扩展: 将配置键从 "10", "80", "160", ..., "5120" (7 个点) 替换为 "128", "256", ..., "262144" (12 个点)，覆盖更广的 effective_batch 范围。
2. 启动参数调优: 为每个新点重新选择 BLOCK_SIZE_M 和 num_warps，例如 1024 点使用 {BLOCK_SIZE_M:8, num_warps:1} 以在低延迟下平衡占用。
3. 生成与验证: 使用 vLLM 内部基准脚本在 MI355X 上运行，--validate 所有 12 个点通过 CPU 参考，--compare 显示相对默认启发式的显著加速 (小批量 ~1.2x，解码范围 ~1.9-2.3x)。
4. 运行时适配: 配置文件名由脚本自动生成 (基于设备名和 dtype)，get_ssm_configs() 无需代码改动即可加载。

vllm/model_executor/layers/mamba/ops/configs/selective_state_update/heads/float16.json

唯一变更文件，重新调整了 MI355X 的 SSU 内核 float16 配置，键从窄网格变为统一网格，启动参数相应更新。

```
// 重新调整后的 MI355X float16 SSU 配置文件内容 (完整)
{
  "triton_version": "3.6.0",
  // effective_batch = batch * nheads, 范围 128~262144
  "128": { "BLOCK_SIZE_M": 16, "num_warps": 4 },
  "256": { "BLOCK_SIZE_M": 32, "num_warps": 4 },
```

```
"1024": { "BLOCK_SIZE_M": 8, "num_warps": 1 },
"2048": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
"4096": { "BLOCK_SIZE_M": 16, "num_warps": 1 },
"8192": { "BLOCK_SIZE_M": 16, "num_warps": 1 },
"16384": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
"32768": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
"65536": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
"131072": { "BLOCK_SIZE_M": 64, "num_warps": 2 },
"196608": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
"262144": { "BLOCK_SIZE_M": 16, "num_warps": 2 }
}
```

评论区精华

PR 无实质性技术讨论。claude[bot] 自动评论表示需维护者手动触发审核，tomeras91 直接批准。

风险与影响

- 风险极低：仅修改 JSON 配置，不影响内核计算逻辑；所有调优点均已通过 CPU 参考和 `--compare` 验证；端到端 `lm_eval` 测试显示准确率不变。唯一风险是配置文件名需与设备精准匹配，否则回退到默认启发式（性能下降但功能正常）。
- 影响：仅限 MI355X 上运行 Mamba2 (`headdim=64`, `dstate=128`, `cache_dtype=float16`) 的用户，可获得 significant decode 加速。

关联脉络

- 先前 PR #47767 为 MI355X 的原宽度网格配置，本 PR 将其对齐到统一网格。
- 统一网格由 MI300X PR #47945 首次启用，随后 MI350X #48159 应用，本 PR 是 AMD Instinct 系列统一的最后一步。
- 后续可能为 float32 配置进行类似调优（PR body 提及 companion PR）。