

PR #47943 完整报告

vllm-project/vllm

Add tuned selective_state_update float32 config for AMD Instinct MI355

合并时间: 2026-07-08 19:41

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47943>

执行摘要

- 一句话: 为 AMD MI355 添加 selective_state_update 的 float32 调优配置
- 推荐动作: 低风险、低复杂度的性能优化 PR, 值得合并。对于使用 Mamba 模型且部署在 MI355 上的团队, 建议关注此配置以提升推理性能。同时也建议其他 AMD GPU (MI300X 等) 提交类似的调优配置。

功能与动机

此前 MI355 仅有 float16 的调优配置 (来自 #47767), 而 NVIDIA 设备均同时提供两种精度。为达到功能对等, 需要为 FP32 状态缓存也提供经过调优的启动配置, 避免回退到通用启发式配置导致性能不佳。PR 提问者提到 'This brings MI355 to parity, so deployments that keep the SSM state cache in fp32 also get a tuned launch config instead of the generic _get_default_ssm_launch_config() heuristic.'

实现拆解

1. 生成调优配置: 使用仓库内已有的基准测试脚本 benchmarks/kernels/benchmark_selective_state_update.py, 在 MI355 上以 float32 状态缓存运行, 自动搜索最优的 BLOCK_SIZE_M 和 num_warps 组合, 并保存到配置文件中。
2. 添加配置文件: 在 vllm/model_executor/layers/mamba/ops/configs/selective_state_update/ 下新增以 headdim=64,dstate=128,device_name=AMD_Instinct_MI355_OAM,cache_dtype=float32.json 命名的 JSON 文件, 内容包含多个 effective_batch 区间对应的启动参数。
3. 无需修改代码: 框架已支持按配置键自动查找和加载此类 JSON 文件, 因此仅需添加文件即可生效, 无需改动任何 Python 源码。

关键文件:

- vllm/model_executor/layers/mamba/ops/configs/selective_state_update/headdim=64,dstate=128,device_name=AMD_Instinct_MI355_OAM,cache_dtype=float32.json (模块配置; 类别 infra; 类型 configuration): 新增的调优配置文件, 为 AMD MI355 在 float32 状态缓存下提供最优的 Triton 内核启动参数 (BLOCK_SIZE_M 和 num_warps), 包含 8 个 effective_batch 档位。

关键符号: 未识别

评论区精华

无实质性讨论：仅有一条 Claude bot 的自动评论（因来自 fork 而跳过审查），以及维护者 tomeras91 的批准，无人工评审意见或问题。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。该 PR 仅新增纯配置文件，不影响任何代码逻辑，且仅针对特定 GPU 型号和精度配置生效（其他 GPU 继续使用通用启发式）。新配置经过基准测试验证，不会引入回归。
- 影响：影响范围有限：仅 AMD Instinct MI355 用户在部署时若使用 float32 的 SSM 状态缓存（Mamba-2 模型），可获得 1.01x 至 1.40x 的 kernel 加速。对于使用 float16 缓存或非 MI355 GPU 的用户无影响。
- 风险标记：暂无

关联脉络

- PR #47767 Add tuned selective_state_update float16 config for AMD Instinct MI355: 同一系列的前置 PR，添加了 MI355 的 float16 版本配置，本 PR 是 float32 的补充。
- PR #47752 MI355 initial config entries and AMD GPU benchmarks: 可能为 MI355 添加了最初的配置条目，是此系列配置的基础。