

PR #48159 完整报告

vllm-project/vllm

[ROCm] Add tuned selective_state_update config for AMD MI350

合并时间: 2026-07-15 15:18

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48159>

执行摘要

该 PR 为 AMD MI350X 添加了调优的 Mamba selective_state_update 内核启动配置 (JSON)，涵盖 float16 和 float32 缓存 dtype。通过基准测试自动生成的最优参数将解码吞吐提升至默认启发式的 2.28x，同时通过严格正确性验证。

功能与动机

vLLM 已为 NVIDIA 各代 GPU 及 AMD MI300X/MI355 提供调优配置，但 MI350X 仍使用通用启发式，性能未达最佳。此 PR 填补该缺失，使 MI350X 用户获得相近性能增益。

实现拆解

1. 自动调优: 在 MI350X 上运行 benchmark_selective_state_update 脚本，遍历 effective_batch 128–262144，对每个 cache_dtype 选择最优 BLOCK_SIZE_M 和 num_warps。
2. 配置生成: 脚本自动输出 JSON 文件，命名包含设备名、headim、dstate 及 cache_dtype，保存至 configs/selective_state_update/ 目录。
3. 加载机制: 运行时根据设备标识自动匹配文件；若无匹配则回退到默认启发式。
4. 验证: 通过 --validate 确保每个配置点与 CPU 参考一致；通过 --compare 确认性能提升；使用 lm_eval lambada_openai 确认端到端正确性。

本 PR 仅添加 JSON 配置文件，未涉及核心逻辑变更，因此无需展示源码。

评论区精华

无实质性讨论。维护者 Tomer Asra 已批准。

风险与影响

- 风险极低: 配置仅改变启动几何，不影响数值结果。
- 影响: MI350X 用户使用 Mamba2 模型 (headdim=64, dstate=128) 时自动获得加速；其他无影响。
- 若未来驱动变更导致配置次优，可删除文件回退默认。

关联脉络

继 #47767 (MI355) 、#47945/#47947 (MI300X) 之后，本 PR 继续为 MI350X 添加调优配置，形成完整的 AMD 主流 GPU SSU 优化覆盖。