PR #51653 完整报告

vllm-project/vllm

[ROCm] Enable V2 model runner for Kimi-K3 on ROCm

合并时间：2026-08-13 22:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51653

执行摘要

- 一句话：Kimi-K3 在 ROCm 上默认启用 V2 model runner
- 推荐动作：建议阅读。该 PR 虽然代码量极小，但揭示了 V2 model runner 在特定平台与并发场景下的性能差距，值得关注后续的性能优化与回归复测。对维护者而言，应尽快为 ROCm + Kimi-K3 + V2 runner 建立自动化性能基准或测试，避免回归再次流入。

功能与动机

PR body 说明：Kimi-K3 on ROCm was gated from using V2 model runner. After validation using the up-to-date upstream. V2 model runner is working as expected. 即此前 ROCm 上 Kimi-K3 被显式排除在 V2 runner 之外，作者经验证认为 V2 runner 已可正常工作，因此解除 gating。

实现拆解

1. 删除平台专属排除常量：在 vllm/config/vllm.py 中移除 ROCM_EXCLUDED_V2_MODEL_RUNNER_ARCHITECTURES（原 frozenset 仅包含 KimiK3ForConditionalGeneration）。
2. 简化架构选择函数：default_v2_model_runner_architectures() 移除 from vllm.platforms import current_platform 延迟导入及 current_platform.is_rocm() 分支，直接返回 DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES，让所有平台的默认 runner 选择逻辑收敛为单一集合。
3. 验证与回退配套：本 PR 未新增自动化测试，验证依赖手动 benchmark（lm-eval gsm8k、vllm bench serve 多并发对比）；若出现性能问题，可通过环境变量 VLLM_USE_V2_MODEL_RUNNER=0 强制回退到 V1。

关键文件：

- vllm/config/vllm.py（模块 配置层；类别 source；类型 dependency-wiring；符号 default_v2_model_runner_architectures, ROCM_EXCLUDED_V2_MODEL_RUNNER_ARCHITECTURES）：唯一的变更文件，删除了 ROCm 对 Kimi-K3 使用 V2 runner 的排除逻辑，直接改变默认 runner 选择行为。

关键符号：default_v2_model_runner_architectures

关键源码片段

vllm/config/vllm.py

唯一的变更文件，删除了 ROCm 对 Kimi-K3 使用 V2 runner 的排除逻辑，直接改变默认 runner 选择行为。

```
DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES = frozenset(
    {
        'DeepseekV2ForCausalLM',
        'DeepseekV4ForCausalLM',
        'GraniteMoeForCausalLM',
        'InklingForCausalLM',
        'InklingForConditionalGeneration',
        'KimiK3ForConditionalGeneration', # Kimi-K3 在 ROCm 上不再被排除
        'LongcatFlashNgramForCausalLM',
        'Qwen2MoeForCausalLM',
    }
)

@lru_cache
def default_v2_model_runner_architectures() -> frozenset[str]:
    # 原实现会按平台剔除 ROCM_EXCLUDED_V2_MODEL_RUNNER_ARCHITECTURES,
    # 本 PR 删除该剔除逻辑，使 ROCm 与其它平台行为一致。
    return DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES
```

评论区精华

核心讨论围绕 V2 与 V1 的性能差异展开：

- LiuYinfeng01 报告 TP8 *decode* 回退：V1 18.52 ms/token，V2 20.89 ms/token（BF16 KV），AITER *cross_device_reduce_2stage* 内核平均延迟从 234.4 us 增至 323.3 us，贡献约 2.43 ms/token。
- vllmellm 用 1-32 并发 bench 反驳，显示 V2 与 V1 基本持平（多数指标 ±1% 内），并说明其测试未开启 prefix caching。
- LiuYinfeng01 复述测试条件差异：低并发、100K context、禁用 prefix caching、--max-num-seqs 1、cudagraph_capture_sizes=[1]、AITER 全开。
- hongxiayang 给出决策：先在 nightly 镜像的 K3 *recipe* 设置 VLLM_USE_V2_MODEL_RUNNER=0，待优化 V2 路径后再置 1。
- TP8 *decode* 性能回退 (performance): 性能回退被确认存在，但根因未定位；建议设置 VLLM_USE_V2_MODEL_RUNNER=0 回退。
- 高并发 benchmark 对比 (performance): 作者认为高并发下无差距，但低并发场景仍有争议。
- nightly 镜像回退策略 (design): 已给出落地指令，尚未确认执行。
- 与 PR #46646 的关系 (question): 决定先合入本 PR。

风险与影响

- 风险：性能回退风险：在 ROCm MI355X 8 卡 TP8、低并发长 context 场景下，V2 相比 V1 有约 12% 的 *decode* 延迟回退，根因疑似与 AITER *cross_device_reduce_2stage* 内

核延迟抖动有关，尚未定位。默认行为变更：ROCm 上 Kimi-K3 的默认 runner 从 V1 变为 V2，用户未显式设置环境变量时会静默切换，可能造成线上性能波动。缺少自动化测试：本 PR 未新增针对 ROCm + Kimi-K3 + V2 runner 的回归测试，验证依赖手动 benchmark，后续回归难以被 CI 捕获。兼容性：删除平台判断后，非 ROCm 平台行为不变（本来默认 V2），但旧注释曾提及 V2 在 profile run 阶段会 fault，需确认已修复。

- 影响：影响范围集中在 ROCm 平台上运行 Kimi-K3 的用户：默认 runner 从 V1 切换为 V2，高并发吞吐基本持平，但低并发 TP8 decode 有显著延迟回退，可能影响在线推理服务。对团队的影响是需要维护 VLLM_USE_V2_MODEL_RUNNER=0 回退配置，并投入资源优化 V2 路径。代码层面，架构选择逻辑被简化，未来所有平台默认 runner 集合统一，降低了配置分支的维护成本。
- 风险标记：性能回归风险，默认行为变更，缺少自动化测试，平台特定路径（ROCm）

关联脉络

- PR #46646 Kimi-K3 ROCm support (mentioned in discussion): 讨论中 yewentao256 与 tjtanaa 均提及该 PR，本 PR 是其子集。
- PR #51855 [K3] support recoverssm for K3: 同模型 Kimi-K3 的 V2 路径功能扩展。
- PR #51809 [XPU] Enable Kimi K3 KDA kernel tests on XPU: 同模型 Kimi-K3 的跨平台内核验证。
- PR #52458 [Kimi-K3][Perf] Update FlashKDA for automatic K2 V-split: Kimi-K3 性能优化，与 V2 路径性能相关。