# PR #44443 完整报告

vllm-project/vllm

[ModelRunner V2] Enable by default for all dense models

合并时间：2026-07-02 18:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44443

## 执行摘要

- 一句话：默认对所有密集模型启用 V2 Model Runner
- 推荐动作：此 PR 设计清晰，增量迁移方式值得借鉴。建议读者关注 _is_default_v2_model_runner_model 中的决策树，以及测试中对 V1/V2 差异的抽象。推荐精读。

## 功能与动机

根据 Issue #41286 的迁移路线图，MR V2 需逐步覆盖所有模型。在已完成密集模型热身（Qwen3、OPT、LLaMA 等）后，本 PR 将默认启用策略从白名单制改为：所有非 MoE 的生成模型默认使用 V2，同时保留 MoE 模型仍通过白名单控制。

## 实现拆解

1. 放宽默认启用条件：在 vllm/config/vllm.py 中，将 DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES 缩减为仅含三个 MoE 架构（DeepseekV2、Qwen2Moe、GraniteMoe）；修改 _is_default_v2_model_runner_model 方法，对 runner_type 为 generate 且非 hybrid、非 attention_free 的模型，若为非 MoE 则直接返回 True。
2. 更新测试预期：tests/test_config.py 中调整了 OPTForCausalLM、Gemma2ForCausalLM 等密集模型的预期为启用 V2；新增 Qwen3.5 hybrid 和 Mamba attention_free 的排除测试。
3. 适配分布式测试：test_remote_prefill_lifecycle.py 根据 use_v2_model_runner 区分 preempted request 的数据结构（NewRequestData vs cached_reqs）；test_multiproc_executor.py 和 test_ray_v2_executor.py 在异步调度开启时增加一个 max_concurrent_batches 计数。
4. 简化 CI 配置：.buildkite/test_areas/model_runner_v2.yaml 移除了 ENFORCE_EAGER=1 的 workaround，因为 V2 已解决 CUDA graph 正确性问题。

关键文件：

- vllm/config/vllm.py（模块 配置模块；类别 source；类型 core-logic；符号 _is_default_v2_model_runner_model, DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES）：核心配置修改：调整默认 V2 架构白名单和 _is_default_v2_model_runner_model 逻辑，对所有非 MoE 生成模型默认启用 V2。

- tests/test_config.py（模块 配置测试；类别 test；类型 test-coverage）：更新 _is_default_v2_model_runner_model 的测试用例，反映新的默认启用规则（密集模型 True，hybrid/attention_free False）。
- tests/v1/kv_connector/unit/test_remote_prefill_lifecycle.py（模块 远程预填调度；类别 test；类型 test-coverage）：适配 V2 下 preempted request 的数据结构差异：V2 使用 NewRequestData 而非 cached_reqs。
- tests/distributed/test_multiproc_executor.py（模块 多进程执行器；类别 test；类型 test-coverage）：适配 V2 下 async scheduling 导致 max_concurrent_batches 增加 1。
- tests/distributed/test_ray_v2_executor.py（模块 Ray 执行器；类别 test；类型 test-coverage）：同 test_multiproc_executor.py 一致，适配 max_concurrent_batches 在 V2 下的变化。
- .buildkite/test_areas/model_runner_v2.yaml（模块 CI 配置；类别 config；类型 configuration）：移除 ENFORCE_EAGER=1 的 workaround，V2 已解决 CG 问题。

关键符号：_is_default_v2_model_runner_model

## 关键源码片段

### vllm/config/vllm.py

核心配置修改：调整默认 V2 架构白名单和 _is_default_v2_model_runner_model 逻辑，对所有非 MoE 生成模型默认启用 V2。

```
# 判断当前模型配置是否应默认使用 V2 Model Runner
# 返回 True 的标准：
# 1. runner_type 必须为 "generate"
# 2. 排除 hybrid（如 Qwen3.5）和 attention_free（如 Mamba）模型
# 3. 若为 MoE 模型，则需在白名单中；否则直接启用
def _is_default_v2_model_runner_model(self) -> bool:
    model_config = self.model_config
    if model_config is None:
        return False

    if model_config.runner_type != "generate":
        return False

    # 排除混合架构（如 Qwen3.5）和无注意力模型（如 Mamba）
    if getattr(model_config, "is_hybrid", False):
        return False
    if getattr(model_config, "is_attention_free", False):
        return False

    architectures = getattr(model_config, "architectures", [])
    # MoE 模型必须出现在 DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES 白名单中，
    # 而非 MoE 模型（密集模型）直接返回 True
    return (
        any(arch in DEFAULT_V2_MODEL_RUNNER_ARCHITECTURES for arch in architectures)
        or not model_config.is_moe
```

)

## 评论区精华

njhill在批准评论中指出："we've finally addressed all the gaps and got the full CI clean!"。
此前经历了多次合并冲突和修复提交，最终由 njhill 协助完成。

- 全面测试通过 (other): 所有问题已解决，CI 全绿，PR 获批准。

## 风险与影响

- 风险：风险包括：1）密集模型中的极端 case（如非常深或特殊架构）在 V2 下可能未充分验证；2）hybrid 和 attention_free 模型的排除逻辑依赖于模型配置中正确设置 is_hybrid 和 is_attention_free 属性，若某些密集模型未正确标注可能误启用；3）async scheduling 增加的一个 concurrent batch 可能对 Pipeline Parallel 场景的显存占用有影响。
- 影响：用户侧：所有密集模型默认使用 V2 Model Runner，可能获得性能提升或行为变化（如 scheduling 策略不同），但应保持兼容。影响程度中等，需关注用户反馈。团队侧：完成迁移路线图中的里程碑，后续可专注于 MoE 模型和特例处理。
- 风险标记：密集模型回归风险，Hybrid/AttentionFree 排除依赖模型属性，Async scheduling 额外批次显存影响

## 关联脉络

- PR #39337 [ModelRunner V2] Initial support: Model Runner V2 系列的首个 PR，引入了基础架构
- PR #39353 [ModelRunner V2] Further dense model support: 后续的密集模型支持 PR，为本次默认启用奠定基础
- PR #41285 [ModelRunner V2] Enable for OPT and LLaMA: 之前启用了部分特定密集模型，本 PR 将其扩展到所有密集模型