PR #38390 完整报告

vllm-project/vllm

[Model Runner v2] E/P/D disaggregation support

合并时间：2026-08-04 23:52

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38390

执行摘要

- 一句话：MRv2 新增 E/P/D 分离部署的 EC 传输支持，PR 最终关闭未合并
- 推荐动作：值得精读 ec_connector.py 的 no-op 抽象与生命周期管理（绑定 metadata、load/save、finally 收敛），这是理解 MRv2 如何扩展外部传输的模板；njhill 的评审展现了 ' 最小化 model_runner 侵入 ' 的维护哲学。但不要直接基于此实现复用：竞态处理、测试覆盖、mixin 与 connector 的职责划分都未定稿，建议跟踪 issue #47999 的后续实现。

功能与动机

PR body 声明目标是 "E/P/D disaggregation support for MRv2"，并给出 Qwen2.5-VL-3B-Instruct 在 GPU_E=1、GPU_PD=2 下的单图验证（TTFT 约 971 ms，总吞吐约 777 tok/s）。vllm/config/vllm.py 的 _get_v2_model_runner_unsupported_features() 原本把 EC transfer 列为 MRv2 不支持并注释指向本 PR，本 PR 正是移除该限制的落地尝试。WoosukKwon 则认为 "this feature is not in a usable state atm... hold this until we have a better design and implementation"，作者在多次 merge 冲突后选择关闭，并说明 "we don't expect to support EPD for v2 currently"；随后 omerpaz95 在评论区为 EC connector 争取保留价值，指向 issue #47999。

实现拆解

1. 配置层能力上提（vllm/config/vllm.py）：新增 VllmConfig.is_ec_producer_only 与 is_encoder_only 两个只读属性，统一收口角色判断；同时从 _get_v2_model_runner_unsupported_features() 删除 "EC transfer" 条目，允许 MRv2 启动时启用 EC 传输。
2. 连接器抽象（新增 vllm/v1/worker/gpu/ec_connector.py）：ECConnector 基类提供默认 no-op 的 maybe_get_output，使不启用 EC 时 execute_model 零改动走原路径（与 kv_connector 的 NO_OP_KV_CONNECTOR 模式同构）；ActiveECConnector 在上下文管理器内完成绑定 SchedulerOutput.ec_connector_metadata、consumer 侧 start_load_caches、producer 侧增量 save_caches、finally 中 get_finished 与 clear_connector_metadata；get_ec_connector() 工厂按环境条件返回 no-op 或 active 实例。
3. 运行器集成（vllm/v1/worker/gpu/model_runner.py）：__init__ 中创建 self.ec_connector，execute_model 用 with self.ec_connector.maybe_get_output(scheduler_output) 包裹 get_mm_embeddings 调用，encoder-only 时构造空输出并附带

ec_connector_output；同时为 is_encoder_only 增加多条捷径：get_kv_cache_spec 返回空 dict、_dummy_run 直接返回空张量、profile_run 提前同步并重置 encoder cache、capture_model 直接返回 0（跳过 CUDA graph 捕获）。

4. 边界修补（block_table.py、warmup.py、mm/encoder_runner.py）：BlockTables 的 apply_staged_writes、gather_block_tables、compute_slot_mappings 三个方法支持 num_kv_cache_groups == 0；warmup_kernels 在 encoder-only 时整体跳过；prepare_mm_inputs 跳过已在 encoder_outputs 中命中的特征 hash，避免 consumer 侧重复编码。
5. 测试与配套：无任何直接对应的测试文件；示例脚本 disagg_1e1pd_example.sh 把服务等待超时从 12000 秒降到 300 秒，是本次唯一的配套调整。

关键文件：

- vllm/v1/worker/gpu/ec_connector.py（模块 编码器传输；类别 source；类型 core-logic；符号 ECConnector, ActiveECConnector, maybe_get_output, get_ec_connector）：本 PR 的核心新增文件：定义 ECConnector 接口与 ActiveECConnector 实现，用 no-op 模式保证未启用 EC 时 execute_model 零改动，是 MRv2 接入 EC 传输的入口。
- vllm/config/vllm.py（模块 配置校验；类别 source；类型 core-logic；符号 is_ec_producer_only, is_encoder_only）：新增 is_ec_producer_only 与 is_encoder_only 属性统一收口角色判断，并从 MRv2 不支持特性清单移除 EC transfer，是功能启用的配置开关。
- vllm/v1/worker/gpu/model_runner.py（模块 模型运行器；类别 source；类型 data-contract）：MRv2 运行器主路径集成：execute_model 用 maybe_get_output 上下文包裹 get_mm_embeddings，并为 encoder-only 节点跳过 KV spec、dummy run、profile 与 CUDA graph 捕获。
- vllm/v1/worker/gpu/block_table.py（模块 块表管理；类别 source；类型 core-logic）：为 encoder-only 节点（不分配 KV 组）补齐三个核心方法的 0 组分支，避免 kernel launch 空跑。
- vllm/v1/worker/gpu/warmup.py（模块 预热逻辑；类别 source；类型 core-logic）：encoder-only 节点无需 decode 侧 kernel 预热，直接跳过 warmup_kernels，避免无意义的 JIT 编译开销。
- vllm/v1/worker/gpu/mm/encoder_runner.py（模块 编码器运行；类别 source；类型 core-logic）：prepare_mm_inputs 跳过已在 encoder_outputs 缓存中的特征 hash，避免 consumer 侧重复编码已从远程加载的 encoder 输出。
- examples/disaggregated/disaggregated_encoder/disagg_1e1pd_example.sh（模块 示例脚本；类别 other；类型 configuration）：示例脚本把服务等待超时从 12000 秒降到 300 秒，是 E/P/D 验证的配套调整。

关键符号：VllmConfig.is_ec_producer_only, VllmConfig.is_encoder_only, ECConnector.maybe_get_output, ActiveECConnector.init, ActiveECConnector.maybe_get_output, get_ec_connector, GPUModelRunner.execute_model, EncoderRunner.prepare_mm_inputs, BlockTables.compute_slot_mappings, warmup_kernels

评论区精华

核心争议是 WoosukKwon 的 hold 决定：他认为功能不可用，要求更好的设计与实现后再推进，作者追问定义未果后关闭 PR。njhill 主导了设计收敛：要求最小化 model_runner 改动、把配置判断收口为 VllmConfig property、用独立的 *ec_connector.py* 替代 mixin，均被作者采纳。gemini 评审提出 consumer 侧 start_load_caches 与 get_mm_embeddings 之间的竞态问题，作者以 "Same logic with v1. Sync logic" 回应，但疑虑未闭环。此外还有命名分歧（maybe_get_output vs get_output）与冗余 assert/mixin 清理。

- 功能可用性与 MRv2 路线决策 (design): PR 关闭，社区诉求转移到 issue #47999 继续追踪 EC connector 支持。
- consumer 侧 cache 加载竞态 (correctness): 未彻底闭环：作者以与 v1 一致为由回应，但评审建议的阻塞等待机制未在最终代码中体现。
- producer 重复 prepare_mm_inputs (performance): 无明显修复证据，PR 关闭前该建议未落地。
- 配置属性上提到 VllmConfig (design): 已采纳，作者将其收口到 vllm/config/vllm.py。
- MRv2 自持 EC 连接器替代 mixin (design): 已采纳，新增 ECConnector/ActiveECConnector 架构并移除对 mixin 的依赖。
- 命名与冗余代码清理 (style): 命名保留 maybe_get_output，冗余 assert 与 mixin 改动被清理。

风险与影响

- 风险：
  1. 竞态隐患：consumer 侧 start_load_caches 后没有显式等待屏障，若 cache 未就绪，get_mm_embeddings 会尝试重复编码，而 consumer 节点可能没有编码器权重，可能直接失败。作者声称与 MRv1 逻辑一致，但缺少测试证明。
  2. 缺少测试覆盖：新增的 num_kv_cache_groups == 0 路径、encoder-only 的 _dummy_run/profile_run/capture_model 捷径均无单测或集成测试，回归风险高。
  3. 核心路径改动：execute_model 是每 step 必经路径，finally 中的 get_finished/clear_connector_metadata 的幂等性未验证，任何连接器异常都会影响所有推理请求。
  4. 设计未定稿：PR 关闭意味着实现未经真实环境全量验证，ECConnector 的 API 形态（命名、上下文语义、与调度器 metadata 的契约）在 issue #47999 的后续实现中很可能调整。 - 影响：对用户 / 系统：若落地，多模态 LLM 可用 1 个 encoder 节点加多个 prefill/decode 节点分离部署，encoder 节点 GPU 利用率可独立伸缩，Qwen2.5-VL 单图验证 TTFT 约 971 ms、总吞吐约 777 tok/s。对团队 / 社区：PR 被 hold 并关闭后，社区在 issue #47999 继续施压，说明 EC connector 已被认为是有价值的方向，MRv2 路线需要成熟的竞态处理与运行器 / 调度器契约设计。影响范围覆盖 vllm/v1 下的运行器、块表、预热、编码器缓存 4 个模块与配置层，属于 MRv2 核心路径。 - 风险标记：核心路径变更，缺少测试覆盖，存在未闭环的竞态疑虑，未合并，设计未定稿，边界条件分支（0 KV 组）

关联脉络

- PR #45043 llmd+vllm+mori-ep(inter node wide-ep)+mori-io(write) for 2p2d with dp=ep=16 tp=1：同属 v1 disaggregated 传输路线：该 PR 扩展 KV cache 跨节点传输（MoRI-IO connector），本 PR 扩展 encoder cache 跨节点传输（EC connector），二者共享 SchedulerOutput 的 connector_metadata 契约与 no-op 连接器模式，可互相参照。