

# PR #46115 完整报告

vllm-project/vllm

[Bugfix] MoRIIO toy P/D proxy: fix DP-rank index aliasing + harden for high-concurrency bursts

合并时间: 2026-07-18 03:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46115>

## 执行摘要

- 一句话: 修复 DP-rank 路由别名及高并发代理加固
- 推荐动作: 推荐阅读者关注 `flat_interleaved_dp_route` 的设计——使用单一槽位索引解决多实例路由别名问题, 该模式可推广至其他需要均匀分配的场景。高并发 `serve` 加固示例展示了从 Quart 切换到 Hypercorn 的实践经验。

## 功能与动机

在 2P\_DP8EP 这种多预填充实例部署场景下, 旧路由方案因索引别名导致半数 GPU 空闲, 浪费资源。同时高并发请求时 Quart 的浅监听队列导致连接被 RST, 影响请求成功率。RFC #46107 中提出的异构 TP↔DP 路由也需要先修复此问题。

## 实现拆解

1. 引入 `flat_interleaved_dp_route` 函数 (`examples/disaggregated/disaggregated_serving/moriiio_toy_proxy_server.py`): 用单个计数器遍历所有 (`instance`, `dp_rank`) 槽位, 消除实例选择与 DP-rank 选择的关联性。
2. 重构路由调用: 将 `handle_request` 中对预填充和解码实例的路由改为统一使用新函数, 并丢弃旧的 `example_round_robin_dp_loader`。
3. 高并发 `serve` 加固: 将 `app.run` 替换为 Hypercorn, 设置 `PROXY_LISTEN_BACKLOG=4096`、`debug=False`、长 `keep-alive`, 避免连接被重置。
4. 新增单元测试: `tests/v1/kv_connector/unit/test_moriiio_proxy_routing.py` 模拟外部依赖并验证所有 RFC 角色的槽位均匀覆盖, 包括回归测试 (2 实例×DP8 必须路由全部 8 个 rank)。

关键文件:

- `examples/disaggregated/disaggregated_serving/moriiio_toy_proxy_server.py` (模块 代理示例; 类别 `source`; 类型 `core-logic`; 符号 `flat_interleaved_dp_route`, `example_round_robin_dp_loader`): 核心修复文件: 实现了新的路由函数 `flat_interleaved_dp_route`, 替换了旧的有别名问题的 `example_round_robin_dp_loader`; 同时将 Quart 调试服务器替换为 Hypercorn 以增强高并发处理能力。
- `tests/v1/kv_connector/unit/test_moriiio_proxy_routing.py` (模块 路由测试; 类别 `test`; 类型 `test-coverage`; 符号 `_module`, `_package`, `_QuartStub`, `_proxy_import_stubs`): 新增完整单元测试, 验证所有 RFC 角色形状的槽位均匀覆盖, 并包含回归测试确保 2 实例

DP8 不遗漏任何 rank。测试通过桩模块隔离外部依赖，只测试真实导入的路由函数。

关键符号: flat\_interleaved\_dp\_route

## 关键源码片段

[examples/disaggregated/disaggregated\\_serving/moriio\\_toy\\_proxy\\_server.py](#)

核心修复文件: 实现了新的路由函数 `flat_interleaved_dp_route`, 替换了旧的有别名问题的 `example_round_robin_dp_loader`; 同时将 Quart 调试服务器替换为 Hypercorn 以增强高并发处理能力。

```
def flat_interleaved_dp_route(request_number: int, instances: list) -> tuple:
    """
    Flat round-robin over the full (instance, dp_rank) slot space.
    使用单个索引 slot 同时决定 instance 和 dp_rank, 消除别名问题。
    例如: inst0_r0, inst1_r0, inst0_r1, inst1_r1, ...
    当 dp_size == 1 时返回 dp_rank = None, 避免无效范围。
    """
    n = len(instances) # 实例数
    dp = instances[0]["dp_size"] # 同角色内 dp_size 应一致
    slot = (request_number - 1) % (n * dp) # 单槽位计数器
    inst_idx = slot % n # 实例索引
    dp_rank = (slot // n) if dp > 1 else None # DP rank, TP 时返回 None
    return inst_idx, dp_rank
```

[tests/v1/kv\\_connector/unit/test\\_moriio\\_proxy\\_routing.py](#)

新增完整单元测试, 验证所有 RFC 角色形状的槽位均匀覆盖, 并包含回归测试确保 2 实例 DP8 不遗漏任何 rank。测试通过桩模块隔离外部依赖, 只测试真实导入的路由函数。

```
# 桩模块上下文管理器, 用于加载代理模块时替换外部依赖
@contextlib.contextmanager
def _proxy_import_stubs():
    stubs = {
        "aiohttp": _module("aiohttp"),
        "quart": _module("quart", Quart=_QuartStub, ...),
        # ... 其他桩
    }
    # 将桩注入 sys.modules, 加载后再恢复
    ...

# 加载真实代理模块, 获取 flat_interleaved_dp_route
@pytest.fixture(scope="module")
def route():
    return _load_proxy_module().flat_interleaved_dp_route

# 测试: 2 实例 xDP8 时所有 16 个槽位必须都被覆盖
@pytest.mark.parametrize("n_instances, dp_size", [(2, 8)])
def test_two_instance_dp8_gives_every_node_all_ranks(route, n_instances, dp_size):
    instances = _instances(n_instances, dp_size)
```

```
assignments = [route(i, instances) for i in range(1, 17)] # 16 requests
# 每个 instance 应分配所有 8 个 dp_rank
for inst_idx in range(n_instances):
    ranks_assigned = {dp for inst, dp in assignments if inst == inst_idx}
    assert len(ranks_assigned) == dp_size, f"Instance {inst_idx} missing ranks"
```

## 评论区精华

审查者 @tjtanaa 指出测试文件中存在 pre-commit 问题（前缀提交格式不规范），作者已修复。审查者最终批准该 PR。此外，作者在 Issue 评论中与 @simondanielsson 讨论了 MoRIIO READ 模式与 vLLM-Router 的兼容性问题，确认官方 Router 目前不支持 MoRIIO READ 模式。

- Pre-commit 错误修复 (style): 作者已修复并折叠到测试提交中。
- MoRIIO READ 模式与 vLLM-Router 兼容性 (question): 确认当前 vLLM-Router 不兼容 MoRIIO READ 模式，需等待后续支持。

## 风险与影响

- 风险：低风险。变更仅影响示例代理文件及新增测试，不涉及核心库。新路由函数假设同一角色内 dp\_size 一致，在异构部署中可能需调整，但当前用例均满足。高并发加固通过配置化参数 PROXY\_LISTEN\_BACKLOG 可调，默认 4096 适合大多数场景。
- 影响：影响 MoRIIO 示例代理的用户，修复后 DP 多实例部署下 GPU 利用率从约 50% 提升至 100%；高并发 burst 请求成功率从部分失败提升至 100%。团队需注意与 #45222 的合并顺序，但无逻辑依赖。
- 风险标记：示例文件变更，高并发路径

## 关联脉络

- PR #45222 [其他 PR] 在同一示例文件上的修改：与同一示例文件（`moriiio_toy_proxy_server.py`）存在行级别冲突，但无逻辑依赖，需要文本合并。
- PR #46107 [RFC] 异构 TP↔DP 路由：本 PR 修复的路由问题是该 RFC 的前置条件，RFC 中提出了多种 P/D 角色形状。