PR #43874 完整报告

vllm-project/vllm

[NixlConnector] Initiate deprecation cycle for `kv_both` role

合并时间：2026-06-05 17:08

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43874

执行摘要

- 一句话：发起 NixlConnector 中 kv_both 角色的弃用流程
- 推荐动作：值得阅读。此 PR 演示了大型项目中如何设计嵌套配置的弃用周期：先软弃用（警告 + 文档 / 测试更新），后续再硬移除。对于 NixlConnector 的用户，应尽快将配置从 kv_both 改为 kv_producer 或 kv_consumer，以免被未来阶段破坏。

功能与动机

根据 RFC #43807，当前 Nixl P/D 实例均使用 kv_both 角色，导致实例无法在启动时确定自身是 prefill 还是 decode，阻碍了配置时优化（如内存分配、调度行为）。其他连接器已使用显式角色，Nixl 成为例外。因此需要分阶段弃用 kv_both，推动用户采用显式角色。

实现拆解

1. 在 vllm/distributed/kv_transfer/kv_connector/v1/nixl/connector.py 的 NixlConnector.__init__ 中增加角色检查：若 kv_role == 'kv_both'，通过 logger.warning_once 输出弃用警告，提示用户改用 kv_producer 或 kv_consumer。
2. 更新所有单元测试文件中的 KV 角色配置：将 test_nixl_connector.py、test_nixl_connector_hma.py、test_multi_connector.py、test_rixl_gpu_mem_diag.py 以及测试工具 utils.py 中的 kv_both 替换为 kv_consumer（或对应角色的显式值）。
3. 新增两个单元测试函数 test_kv_both_deprecation_warning 和 test_explicit_kv_role_no_deprecation_warning，分别验证使用 kv_both 时触发警告、使用显式角色时不触发。
4. 修改集成测试脚本 run_accuracy_test.sh 和 spec_decode_acceptance_test.sh：将原先对所有实例统一的 kv_both 拆分为 prefill 实例使用 kv_producer、decode 实例使用 kv_consumer，确保实际部署场景下的角色分离。
5. 同步更新文档 docs/features/nixl_connector_usage.md 和 docs/design/nixl_kv_cache_lease.md 中的配置示例，引导用户使用显式角色。

关键文件：

- tests/v1/kv_connector/unit/test_nixl_connector.py（模块 Nixl 连接器；类别 test；类型 test-coverage；符号 test_kv_both_deprecation_warning, test_explicit_kv_role_no_deprecation_warning）：核心测试文件：新增两个测试函数验证 deprecation warning 行为，并将原有测试中的 kv_both 迁移为 kv_consumer，确保弃用

逻辑正确且回归无虞。

- vllm/distributed/kv_transfer/kv_connector/v1/nixl/connector.py（模块 Nixl 连接器；类别 source；类型 core-logic）：核心源代码：在 NixlConnector.__init__中添加 kv_both 检测与弃用警告，是实现软弃用的关键入口。
- tests/v1/kv_connector/nixl_integration/run_accuracy_test.sh（模块 集成测试；类别 test；类型 test-coverage）：集成测试脚本：将预填充和解码实例的配置从 kv_both 拆分为 kv_producer / kv_consumer，确保集成环境遵循显式角色。
- tests/v1/kv_connector/nixl_integration/spec_decode_acceptance_test.sh（模块 集成测试；类别 test；类型 test-coverage）：集成测试脚本：类似 run_accuracy_test.sh，更新角色配置，并新增显式角色变量传递。
- tests/v1/kv_connector/unit/test_multi_connector.py（模块 连接器测试；类别 test；类型 test-coverage）：多连接器测试：将一处 kv_both 引用更新为 kv_consumer，保持一致性。
- tests/v1/kv_connector/unit/test_nixl_connector_hma.py（模块 连接器测试；类别 test；类型 test-coverage）：HMA 连接器测试：将 kv_both 更新为 kv_consumer。
- tests/v1/kv_connector/unit/test_rixl_gpu_mem_diag.py（模块 连接器测试；类别 test；类型 test-coverage）：GPU 内存诊断测试：将 kv_both 更新为 kv_consumer。
- tests/v1/kv_connector/unit/utils.py（模块 工具函数；类别 test；类型 test-coverage）：测试工具：修改默认配置文件中的角色，确保工具函数与主体一致。
- docs/features/nixl_connector_usage.md（模块 用户文档；类别 docs；类型 documentation）：用户文档：更新使用指南中的配置示例，用显式角色替代 kv_both。
- docs/design/nixl_kv_cache_lease.md（模块 设计文档；类别 docs；类型 documentation）：设计文档：更新一处配置示例角色。

关键符号：NixlConnector.init, test_kv_both_deprecation_warning, test_explicit_kv_role_no_deprecation_warning

关键源码片段

tests/v1/kv_connector/unit/test_nixl_connector.py

核心测试文件：新增两个测试函数验证 deprecation warning 行为，并将原有测试中的 kv_both 迁移为 kv_consumer，确保弃用逻辑正确且回归无虞。

```
import pytest
from unittest.mock import patch
from vllm.logger import _print_warning_once

def test_kv_both_deprecation_warning(default_vllm_config, dist_init):
    """kv_role='kv_both' 应输出弃用日志警告"""
    _print_warning_once.cache_clear()
    vllm_config = create_vllm_config(kv_role="kv_both")
    with patch(
        "vllm.distributed.kv_transfer.kv_connector.v1.nixl.connector.logger"
    ) as mock_logger:
        mock_logger.warning_once = mock_logger.warning_once
```

```
        NixlConnector(
            vllm_config,
            KVConnectorRole.WORKER,
            make_kv_cache_config(block_size=16),
        )
    # 验证 warning_once 被调用一次，且消息包含 'kv_both' 和 'deprecated'
    mock_logger.warning_once.assert_called_once()
    msg = mock_logger.warning_once.call_args[0][0]
    assert "kv_role='kv_both'" in msg
    assert "deprecated" in msg


def test_explicit_kv_role_no_deprecation_warning(default_vllm_config, dist_init):
    """显式角色 kv_consumer / kv_producer 不应触发弃用警告"""
    for role in ("kv_consumer", "kv_producer"):
        vllm_config = create_vllm_config(kv_role=role)
        with patch(
            "vllm.distributed.kv_transfer.kv_connector.v1.nixl.connector.logger"
        ) as mock_logger:
            NixlConnector(
                vllm_config,
                KVConnectorRole.WORKER,
                make_kv_cache_config(block_size=16),
            )
        # 确保 warn_once 未被调用
        mock_logger.warning_once.assert_not_called()
```

评论区精华

审阅者 ZhanqiuHu 指出测试中曾使用 kv_both 角色来测试“双向”一致性，担心测试迁移是否完整。PR 作者通过更新所有测试文件并新增 deprecation warning 专门测试来回应，最终获得 Approval。此外，Mergify 的预提交检查失败仅涉及补齐依赖（pre-commit），不影响逻辑。

- 测试中 kv_both 使用是否被完整迁移 (testing): PR 作者更新了所有相关测试文件并新增 deprecation warning 专门测试，确保迁移完整，最终获得 approval。

风险与影响

- 风险：
  - 兼容性风险：现有用户仍可使用 kv_both，但会看到警告，可平滑过渡，短期无破坏。
  - 测试覆盖风险：如果集成测试脚本中的角色配置遗漏，可能导致 CI 用例失效。但本次已全面排查并修改了所有引用 kv_both 的测试脚本和工具函数，新增的专项测试进一步保障。
  - 功能回归风险：无功能性代码改动，仅增加警告和配置字符串变更，回归风险极低。
- 影响：
  - 用户影响：当前使用 NixlConnector 且配置 kv_both 的用户会在日志中看到一次弃用警告，建议在下一个更新周期迁移配置。

- 系统影响：无。角色字段在当前阶段仍被忽略，运行时行为不变。
- 团队影响：为后续 Phase 2（强制角色假设）铺平道路，降低了维护成本。
- 风险标记：配置迁移风险，兼容性警告，无功能回归风险

关联脉络

- PR #43807 [RFC]: Deprecate kv_both for NIXLConnector and Enforce Explicit P/D Roles: 此 PR 实现了 RFC #43807 的第一阶段弃用计划。