PR #7262 完整报告

PaddlePaddle/FastDeploy

[XPU][Docs] Update Release Note

合并时间：2026-04-10 15:22

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7262

执行摘要

- 一句话：更新 XPU 部署文档中的 RDMA 网卡配置和术语，并调整相关脚本输出格式。
- 推荐动作：建议技术管理者优先审查脚本输出兼容性风险，确保 CI 测试适配新格式；工程师可精读文档变更以了解 XPU 部署最佳实践，但无需深入代码逻辑。关注 review 中未解决的配置通用性问题，未来文档更新应考虑使用占位符提高可移植性。

功能与动机

PR body 未填写具体动机，但从修改内容和 review 讨论推断，旨在更新 Release Note 以反映 XPU 部署的最新配置，适配 Mellanox RDMA 网卡环境，并提供更准确的部署命令示例。review 中多次提到需优化术语和修复配置错误，以提升用户部署体验。

实现拆解

主要修改包括：1) 文档更新：删除旧模型支持表格，添加版本说明；将 BKCL_RDMA_NICS 配置值从 *eth1,eth3* 等更新为 mlx5_1,mlx5_2 等；统一术语如 'Quick Deployment'→'Quick Launch'、'Best Deployment'→'Optimal Performance'；修复 shell 变量展开语法（如 ${mtp_model_path} 的引用）。2) 脚本调整：修改 scripts/get_rdma_nics.sh，在 xpu 模式下输出两行环境变量（KVCACHE_RDMA_NICS 和 BKCL_RDMA_NICS），而非单行。中英文文档同步变更，确保一致性。

关键文件：

- docs/usage/kunlunxin_xpu_deployment.md（模块 文档）：英文 XPU 部署文档，核心变更包括更新 RDMA 网卡配置、优化术语和调整模型路径，直接影响用户部署命令。
- docs/zh/usage/kunlunxin_xpu_deployment.md（模块 文档）：中文 XPU 部署文档，同步英文变更并修复注释错误，确保中英文用户体验一致。
- scripts/get_rdma_nics.sh（模块 基础设施）：RDMA 网卡获取脚本，输出格式变更可能破坏现有 CI 脚本解析，是本次 PR 的主要风险点。

关键符号：JUDGE_NIC_TYPE

评论区精华

review 中核心讨论点：1) 脚本输出兼容性：Copilot 指出脚本修改可能破坏现有调用方（如 tests/xpu_ci/conftest.py）对单行输出的依赖，建议保持单行或新增 flag，但 PR 中未采纳此

建议。2) 文档配置通用性：Copilot 多次建议使用占位符而非固定 mlx5_* 值，以避免在不同机器上不适用，但文档仍保留固定示例。3) 细节错误：fastdeploy-bot 标记了注释中重复 # 符号的 typo 和模型路径格式不一致问题，部分在后续 commit 中修复。4) speculative-config 语法：Copilot 指出 JSON 字符串中变量展开可能不合法，建议改进但未实施。讨论结论显示脚本兼容性疑虑未解决，配置通用性问题被忽略。

- 脚本输出兼容性风险 (correctness): 建议保持单行输出或新增 flag，但 PR 中未修改，此风险未解决。
- 文档配置通用性不足 (design): PR 维持固定示例，未采纳建议，可能误导用户。
- 注释格式错误修复 (style): 被识别为需要修复的问题，但在提供的 patch_excerpt 中未显示修复，状态不确定。

风险与影响

- 风险：技术风险包括：1) 脚本风险：scripts/get_rdma_nics.sh 的输出格式变更（从单行改为两行）可能导致依赖单行解析的 CI 脚本（如 tests/xpu_ci/conftest.py）错误解析，污染 KVCACHE_RDMA_NICS 和 BKCL_RDMA_NICS 环境变量，影响 CI 测试稳定性。2) 文档风险：固定 mlx5_* 网卡名称示例缺乏通用性，用户在不同硬件上直接复制可能导致 BKCL 初始化失败。3) 兼容性风险：shell 变量展开语法修改（如 '${mtp_model_path}'）可能在某些环境中产生无效 JSON，影响 speculative decoding 功能。
- 影响：影响范围：1) 用户影响：XPU 部署用户需参考更新后的文档，术语优化和配置示例调整提高了可读性，但固定网卡示例可能增加配置困惑，需自行适配。2) 系统影响：脚本变更可能影响 CI 测试的解析逻辑，若调用方未更新，可能导致测试失败或环境变量错误。3) 团队影响：维护团队需关注脚本兼容性问题，可能需后续 PR 修复或更新相关 CI 脚本。影响程度：中等，主要限于文档和基础设施层面，不涉及核心模型或算子代码。
- 风险标记：脚本输出破坏兼容性，文档配置不通用，模型路径不一致

关联脉络

- PR #7264 [XPU][CI] lock xvllm version for fix bug: 同样涉及 XPU CI 配置和脚本调整，与本 PR 的脚本修改有技术关联。
- PR #7302 [Docs] Update Release Note: 同为文档更新 PR，聚焦于 Release Note 同步，显示仓库持续进行文档维护。