# PR #6700 完整报告

PaddlePaddle/FastDeploy

[Docs] Add docs for disaggregated deployment

合并时间：2026-04-01 19:27

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/6700

## 执行摘要

该 PR 新增了 PD 分离部署的最佳实践文档，涵盖单机和跨机混合并行配置，旨在解决当前 FastDeploy 中相关文档缺失的问题。通过详细部署指南、配置示例和启动脚本，为用户提供了一站式参考，风险较低但需注意文档准确性。

## 功能与动机

为什么做：根据 PR body，作者指出“目前缺少 PD 分离部署下的混合并行的实践文档”，因此创建此文档以补充 FastDeploy 的部署指南，帮助用户实施混合并行和跨机部署场景。

## 实现拆解

实现方案按文档模块拆解如下：

- 核心新增文档：docs/best_practices/Disaggregated.md 和 docs/zh/best_practices/Disaggregated.md 新增中英文最佳实践文档，内容一致，包括：
  - 部署概览和环境准备，附配置表格（如 TP/DP/EP 并行度）。
  - 单机和跨机部署拓扑图。
  - 详细启动脚本示例，使用 ERNIE-4.5-300B 模型和 H100 GPU。
- 文档索引更新：docs/best_practices/README.md 和 docs/zh/best_practices/README.md 添加新文档索引项。
- 功能文档链接：docs/features/disaggregated.md 和 docs/zh/features/disaggregated.md 增加“最佳实践”跳转链接，增强导航。

## 评论区精华

Review 讨论聚焦于文档准确性，主要交锋如下：

- juncaipeng：要求区分 Prefill 和 Decode 配置，例如“区分 P TP4DP1 D TP4DP1”，并指出“EP 后面需要 8”以细化并行度描述。
- Copilot：建议修复英文文档链接指向英文站点，并警告“multi_api_server 启动命令参数错误可能导致部署失败”。 讨论结论是作者采纳建议，在提交中更新文档以解决这些问题，提升了文档质量。

## 风险与影响

风险：主要风险是文档内容不准确，如启动命令参数错误（见于 docs/best_practices/Disaggregated.md）可能误导用户部署失败；链接指向错误可能影响用户体验。无代码回归或性能风险。 影响：对用户影响积极，提供了实用部署指南，降低实施门槛；对系统无直接冲击，但文档完善可能间接提升部署成功率和团队效率。

## 关联脉络

该 PR 与近期历史 PR 关联紧密，揭示了 PD 分离部署功能的持续演进：

- PR 7107：优化 PD 分离部署的缓存处理和调度逻辑，为本文档提供了技术基础。
- PR 6929：修复 KVCache bug，PD 分离部署依赖缓存管理，此修复确保部署稳定性。
- 其他相关 PR：如 PR 6992 新增中断请求端点，展示了系统功能的扩展，与本文档的部署实践相辅相成。 整体来看，FastDeploy 在 PD 分离部署方向正通过代码优化和文档补充双线推进，以支持更复杂的混合并行场景。