

PR #2375 完整报告

radixark/miles

docs: reorganize the Training Backends section, drop the experimental FSDP framing

合并时间: 2026-08-12 10:49

原文链接: <http://prhub.com.cn/radixark/miles/pull/2375>

执行摘要

- 一句话: 重组训练后端文档, FSDP 升为一等后端
- 推荐动作: 值得精读。对使用者, training-backend.md 是选型和配置 Megatron/FSDP 的权威入口, 尤其 FSDP 一节此前从未按代码事实完整写过; 对维护者, 建议把该页纳入涉及 FSDP 参数、并行或 checkpoint 的 PR 的改动清单。更值得借鉴的是本 PR 的工作方式: 以代码为唯一事实源撰写文档, 并在写作过程中反哺修复了 4 个真实缺陷。注意页面按修复后的状态编写, 阅读时需结合 #2384/#2386 的变更理解 flag 现状。

功能与动机

PR body 指出, FSDP 文档把后端描述成实验性项目: 它位于 developer/experimental-features, 写着“no TP/PP/CP/EP, small dense models only”, 并指向已不存在的代码路径 miles/backends/experimental/fsdp_utils/——而 #2272 早已把 backend 移出 experimental/, 该描述不再成立。同时 advanced/architecture-support (“Backends Beyond Megatron”) 实际讲的是把 HF 实现嵌入 Megatron 内部的机制, 页面名与内容不符。本次变更将整个 Training Backends section 重写为围绕真实存在的两个后端, 并删除误导页面。

实现拆解

1. 信息架构调整: 将 user-guide/usage.md (仅 Megatron) 重写并重命名为 user-guide/training-backend.md (双后端), 删除 developer/experimental-features.md 和 advanced/architecture-support.md; docs/docs.json 更新用户指南导航顺序、移除 Advanced/Backends 分组和 experimental-features 条目, 并新增 3 条 redirect (/user-guide/usage、/advanced/architecture-support、/developer/experimental-features), 其中 architecture-support 与 experimental-features 的 redirect 带锚点指向新页对应小节。
2. 新页面内容结构: 按读者自上而下的决策路径组织——先讲“后端是什么”(TrainRayActor 子类 + init/train/update_weights/save_model/sleep/wake_up 契约方法, 由 --train-backend 在 miles/ray/train/actor_factory.py 中选择), 再给“选哪个”(Megatron-LM 用于大模型与真实并行, FSDP 用于原样训练 HF 实现), 随后分别按“六步决策”展开 Megatron 与 FSDP 的配置, 最后保留 SGLang 推理半边。新增两节: GPU layout (disaggregated 与 --colocate 两种布局, 以及布局如何决定 weight-sync 路径) 和 Fitting it in memory (拆分为 layout-agnostic 旋钮与 colocate 专属 offload 族)。

3. 按代码事实逐条纠错：对照 `fsdp_utils/parallel.py` 修正并行维度描述 (`dp_replicate × dp_shard`, TP/PP/CP/EP/ETP 在 FSDP `ParallelState` 中固定为 1)；明确 MoE 受支持 (`fsdp_utils/kernels/` 的 fused expert kernels、weight sync 时 `batched-expert unfusing`、`--use-rollout-routing-replay`)；检查点改为 PyTorch DCP 目录 + `latest_checkpointed_iteration.txt`；说明 fp32 master copy (`--no-keep-fp32-master`) 是 weight sync 位精确的原因；`--attn-implementation` 是透传字符串，`flash_attention_3` 也可用，并同步修正 `cli-reference.md` 中错误的枚举描述。
4. 联动修复 4 个代码缺陷：写文档时发现的 bug 在 main 上各自独立修复——#2384 (`build_fsdp_parser` 的 `store_true` 静默翻转 `FSDPArgs` 布尔默认值，改用 `BooleanOptionalAction` 并把 `--disable-fp32-master` 更名为 `--no-keep-fp32-master`)、#2386 (删除从未可达的 FSDP 上下文并行分支)、#2382 (删除重复的 `_validate_rematerialize_param_from_master_weight` 调用)、#2390 (修正 `--offload-train/--offload-rollout` 的 help 文本)。提交 5863d07 跟进 #2384 更新 `precision/memory` 表，并删除无读者的 `--fsdp-state-dict-cpu-offload` 行。
5. 全站同步与合并冲突处理：`docs/index.md`、`developer/index.md`、`developer/debug.md`、`getting-started/installation.md`、`getting-started/quick-start.md`、`user-guide/argument-groups.md`、`user-guide/cli-reference.md`、`user-guide/concepts.md`、`user-guide/index.md`、`user-guide/training-script-walkthrough.md` 等 10 余处链接统一指向 `/user-guide/training-backend`。多次 merge main 处理与 #2366、#2359、#2376、#2391、#2395、#2354 的冲突；无测试配套，合并时由作者人工核查全部 80 个 nav 条目、redirect 目标与站内死链。

关键文件：

- `docs/user-guide/training-backend.md` (模块 用户指南；类别 docs；类型 documentation)：499 行新增核心页面，双后端对比指南，按代码事实重写 FSDP，是本次 PR 的主体。
- `docs/user-guide/usage.md` (模块 用户指南；类别 docs；类型 deletion)：被替换的旧单后端页面 (200 行)，内容迁移并扩展进新页 `training-backend.md`。
- `docs/advanced/architecture-support.md` (模块 高级主题；类别 docs；类型 deletion；符号 Qwen3NextBridge, MyModel, Qwen3_5Bridge, `_weight_to_mcore_format`)：页面名“Backends Beyond Megatron”与实际内容 (Megatron 内部 HF 桥接机制) 不符，整页删除，仅在新页保留简短指针，旧 URL 重定向到带锚点的新页小节。
- `docs/developer/experimental-features.md` (模块 开发指南；类别 docs；类型 deletion)：FSDP 的“实验特性”框架已不成立，页面整体删除，FSDP 内容按现状写入新页。
- `docs/docs.json` (模块 文档配置；类别 config；类型 configuration)：导航更新 (`usage` → `training-backend`，移除 Backends 分组与 `experimental-features`) 并新增 3 条旧 URL 重定向。
- `docs/index.md` (模块 站点首页；类别 docs；类型 documentation)：首页描述把 FSDP 从实验引导改为正式后端，并更新快速入口链接。
- `docs/user-guide/cli-reference.md` (模块 用户指南；类别 docs；类型 documentation)：同步修正 `--attn-implementation` 的枚举描述，与 FSDP 透传语义一致。

关键符号：TrainRayActor, MegatronTrainRayActor, FSDPTrainRayActor, init, train, update_weights, save_model, sleep, wake_up, get_miles_extra_args_provider

评论区精华

本 PR 没有公开 review 评论，唯一 review 来自 Shi-Dong 的 APPROVED（无文字）。技术讨论沉淀在 commit 信息中：

- 提交 5863d07 (follow #2384) 解释为何删除 `--fsdp-state-dict-cpu-offload` 行：字段声明在 `FSDPArgs` 上但全树无读者，紧挨着真实生效的 `--fsdp-cpu-offload` 排列会“invites setting the wrong flag”。
- merge commit f6a0c93 (与 #2376 冲突)：developer/architecture.md 目录树整体取 main 的版本，因为 #2376 已把 `fsdp_utils` 移出 experimental/ 并给出更好的分目录描述；本分支只保留“Experimental Features 卡片替换为 Training Backends 卡片”的改动。
- merge commit 1b95dd9 (redirect 冲突)：与 #2391 在 redirects 数组同位置冲突时“Left as authored rather than normalizing one side to the other”，保留 #2391 的 `permanent: true` 与本分支三条不带 permanent 的条目，并人工验证 80 个 nav 条目和所有 redirect 目标无死链。
- 暂无高价值评论线程

风险与影响

- 风险：
 1. 文档与代码强耦合：新页面的 FSDP 叙述逐条对齐 `fsdp_utils/parallel.py`、`FSDPArgs`、`actor_factory.py`，代码演进后页面会再次过时；后续任何改动 FSDP 参数解析或并行结构的 PR 都应同步该页，否则会重新出现本次修复的脱节。
 2. 三条新 redirect (`/user-guide/usage`、`/advanced/architecture-support`、`/developer/experimental-features`) 均未带 `permanent: true`，而既有条目都带；若站点按 301/302 区分处理，旧 URL 的 SEO 权重迁移与缓存行为会不一致。
 3. 删除 `architecture-support.md` 后，Qwen3-Next/Gated-Delta-Net 的 HF 桥接、mbridge 权重转换、fp32 参数保持等深度内容只剩新页一段短指针；在模型支持文档专项 pass 落地前，新架构接入者会缺少参考资料。
 4. 无自动化校验：合并时仅靠人工检查 nav 条目与 redirect 目标，后续其他 PR 改名页面时仍可能引入死链。
 - 影响：用户侧：FSDP 用户首次在正式文档中获得与 Megatron 并列的选型矩阵，“实验特性”的警告框架被移除；收藏旧 URL 的用户通过 redirect 到达新页（`architecture-support` 带锚点跳转）。团队侧：该页成为两后端契约的文档基线，任何后端接口改动都有了核对清单；同时本 PR 触发的 4 个修复改变了 FSDP 参数解析行为——#2384 将 `--disable-fp32-master` 更名为 `--no-keep-fp32-master` (breaking, 但仅影响两个测试)，#2386 删除了从未生效的 CP 分支，属于用户不可见的清理。系统侧：纯文档变更，不触碰源码、测试与部署产物。
 - 风险标记：旧 URL 依赖重定向，文档与代码强耦合，新增 redirect 未标 permanent, 技术内容过渡期精简

关联脉络

- PR #2384 fix(fsd): stop store_true from shadowing bool defaults in FSDPArgs: 写文档时发现 build_fsd_parser 用 store_true 静默翻转 FSDPArgs 两个 True 默认值; 页面按修复后的 --no-keep-fp32-master 状态编写。
- PR #2386 fix: drop context parallelism from the FSDP backend: 发现 FSDP CP 分支全部不可达后删除死代码; 页面记录修复后的二维 dp mesh。
- PR #2382 fix: drop duplicated rematerialize validation call: 本 PR 指出 validate_args 中连续调用了两次 _validate_rematerialize_param_from_master_weight, 独立修复后合并。
- PR #2390 docs(args): correct the offload flags' help text: 本 PR 事实核查时发现 --offload-train/--offload-rollout 的 help 文本描述错误 phase 并夸大 --colocate, 专门修复。
- PR #2376 docs(developer): rewrite the developer guide against the code: 同区文档重写, 与本分支在多处冲突; 合并时保留 main 的目录树并替换其实验特性卡片。
- PR #2391 docs: replace DeepSeek V3/R1 page with a DeepSeek-V3.2 recipe: 在 docs.json redirects 同位置新增条目引发冲突, 合并时保留全部四条。
- PR #2354 Delete the glm4-9B, mimo-7B, moonlight-16B and deepseek-r1 launch scripts: 恢复 deepseek 页面后, 本分支原有的 /models/deepseek/deepseek redirect 变为 stale, 最终合并时删除。
- PR #2366 docs: rewrite the fully async page around schedule, data path, eval, and metrics: merge main 带入; 其 usage.md 链接更新通过 rename-aware merge 搬入 training-backend.md。