PR #2396 完整报告

radixark/miles

docs: refresh MXFP8 and NVFP4 RL guide

合并时间：2026-08-12 07:54

原文链接：http://prhub.com.cn/radixark/miles/pull/2396

执行摘要

- 一句话：刷新 MXFP8/NVFP4 低精度 RL 指南并重命名页面
- 推荐动作：值得精读，尤其适合负责 Blackwell 低精度训练的用户、文档维护者与关注 RoI 的技术管理者。核心看点：① 四阶段精度契约与 rollout × 训练兼容矩阵如何把 MoE RL 精度漂移的排障成本前置；② MXFP8 的 MoE / dense GEMM 后端选择拆分，纠正了旧文档单点建议的误导；③ NVFP4 后向模式「基础 + 高级选项」的分级呈现与外部来源引用策略；④ 文档重命名时全站链接、重定向与跨 PR 冲突处理的最佳实践。建议结合 issue #615 及其实现 PR 对照阅读。

功能与动机

旧文档 fp8-low-precision.md 把 NVFP4 标为 *experimental*（兼容矩阵里还是 'coming soon'），MXFP8 也只覆盖早期 Blackwell 路径，与代码现状严重脱节。Roadmap issue #615（Blackwell MXFP8 与 NVFP4 RL training）中列出的多条实现 PR（Miles 侧 #512/#963/#1340/#1935，NVFP4 侧 #907/#1054/#1261/#2043）均已合入，功能已经可用。PR body 明确了本期范围：刷新 MXFP8 与 NVFP4 章节、保留 BF16 与 block-wise FP8 配方、文档化共享的 checkpoint / training / rollout / live-update 精度契约、把 MXFP8 的 MoE 与 *dense* 线性 GEMM 后端选择拆开，并明确非目标为「不改任何运行时或配方代码」。

实现拆解

按以下步骤拆解（纯文档变更，无源码 / 测试 / 部署配套）：

1. 指南主体重写：新建 docs/advanced/low-precision.md（+326 行），删除 docs/advanced/fp8-low-precision.md（−200 行）。新指南以「MoE RL 中 rollout 与训练之间精度漂移是常见失败模式」开篇，给出精度选择表（BF16 baseline、FP8 block-wise 为 Generally available、MXFP8 与 NVFP4 为 Beta）和 rollout × 训练兼容矩阵（仅对角线与 FP8 横线可行），并把 checkpoint 转换、trainer 前向、SGLang rollout、live weight export 四阶段精度契约固定为统一配方的基础。
2. MXFP8 章节细化：文档化 Qwen3-30B-A3B launcher 的成对命令（*prepare* / *execute* 搭配 `--rollout-mxfp8 --train-mxfp8`）和 Megatron 侧 `--fp8-recipe mxfp8` 训练 flags；将 SGLang 后端选择从旧文档的单点建议（cutlass / triton）拆为两张独立表——MoE 用 `--sglang-moe-runner-backend`（`flashinfer_trtllm_routed` / `flashinfer_trtllm` / `cutlass`），*dense* 线性 GEMM 用 `--sglang-fp8-gemm-backend`（`flashinfer_trtllm` / `flashinfer_cutlass`/`triton`），并说明两者按模型、并行布局与已安装SGLang栈独立配对。

3. NVFP4 章节重构：基础配方保留 high-precision backward；dequantized backward 与 four-over-six 改为「单独引用的高级选项」呈现；补充 NVFP4 的 scaling 契约（E2M1 权重按 16 值块存储，每块一个 E4M3 scale 加外层 FP32 scale，gate 与 up projection 联合量化），并与 MXFP8 区分训练 flags。
4. 图示与来源：新增三张示意图 docs/assets/images/low-precision/（mxfp8-e2e.png、nvfp4-high-precision-backward.png、nvfp4-dequantized-backward.png），在文中注明来自 LMSYS 博客；外部引用链到 public roadmap issue #615、LMSYS overview 与 humans& 的 NVFP4 / 4-over-6 分析。
5. 导航与全站链接：docs/docs.json 把 Performance 分组的页面项由 advanced/fp8-low-precision 改为 advanced/low-precision，并新增 /advanced/fp8-low-precision → /advanced/low-precision 永久重定向；qwen3.md、deepseek-v3-2.md、nemotron-3-super.md、GLM 系列、gpt-oss、moonlight、advanced/index.md 等 23 个引用旧路径的页面统一改为新标题「Low Precision RL」，advanced/index.md 的卡片描述也同步更新。
6. 对照代码的事实修正：block-wise FP8 一节把 NVTE_FP8_BLOCK_SCALING_FP32_SCALES 的说明改为「由 miles/ray/train/actor_factory.py 在 actor env 中按硬件默认设置：Hopper 为 1、Blackwell 为 0（TransformerEngine 用 MXFP8 模拟 block-wise，需要 2 的幂 scale）」，不再要求用户手工设 1；NVFP4 一节补上与 MXFP8 不同的训练 flags 说明；新增图片路径统一为全站一致的绝对路径，符合 docs/README.md 约定。

关键文件：

- docs/advanced/low-precision.md（模块 低精度指南；类别 docs；类型 documentation）：本次变更的核心产物：326 行的统一低精度 RL 指南，覆盖 block-wise FP8、MXFP8、NVFP4 的精度契约、兼容矩阵、可运行配方与后端选择。
- docs/advanced/fp8-low-precision.md（模块 低精度指南；类别 docs；类型 deletion）：被整体删除的旧版 FP8 指南（200 行），其内容迁入 low-precision.md；NVFP4 从 experimental / coming soon 升级为 Beta。
- docs/docs.json（模块 文档导航；类别 config；类型 configuration）：文档站导航配置：Performance 分组页面项替换为 advanced/low-precision，并新增永久重定向，且与 #2395 的 Platforms 重定向在合并时共存。
- docs/assets/images/low-precision/mxfp8-e2e.png（模块 配图资源；类别 other；类型 asset）：新增端到端 MXFP8 RL 配方示意图，来源于 LMSYS 博客，是 MXFP8 章节的视觉核心。
- docs/assets/images/low-precision/nvfp4-high-precision-backward.png（模块 配图资源；类别 other；类型 asset）：NVFP4 基础配方（高精度后向）示意图，与 dequantized 后向图配套说明两种后向模式。
- docs/assets/images/low-precision/nvfp4-dequantized-backward.png（模块 配图资源；类别 other；类型 asset）：NVFP4 高级选项（dequantized backward）示意图，用于区分基础配方与高级模式。
- docs/models/qwen/qwen3.md（模块 模型文档；类别 docs；类型 documentation）：模型文档中的低精度页链接由 /advanced/fp8-low-precision 更新为 /advanced/low-precision，并同步“Pairs Well With”引用。

- docs/models/deepseek/deepseek-v3-2.md（模块 模型文档；类别 docs；类型 documentation）：DeepSeek-V3.2 是 MXFP8 目标模型之一，其低精度页链接同步更新，与 MXFP8 章节内容呼应。
- docs/advanced/index.md（模块 栏目首页；类别 docs；类型 documentation）：Advanced Features 栏目首页的卡片链接与描述同步更新为新页面与新范围。

关键符号：未识别

关键源码片段

docs/advanced/low-precision.md

本次变更的核心产物：326 行的统一低精度 RL 指南，覆盖 block-wise FP8、MXFP8、NVFP4 的精度契约、兼容矩阵、可运行配方与后端选择。

```
---
title: Low Precision RL
description: Unified low-precision pipelines for RL — block-wise FP8, MXFP8, and NVFP4 across rollout and training.
---

<!-- 本页是低精度 RL 的统一入口，由旧版 fp8-low-precision.md 重写而来。
     新版把 MXFP8 / NVFP4 从 experimental 升为 Beta，并固定四阶段精度契约：
     checkpoint 转换、trainer 前向、SGLang rollout、live weight export 必须使用同一套量化逻辑，
     否则 MoE 的 rollout 与训练之间会出现逐层数值漂移，最终表现为 log-probability 发散。 -->

## Choose a precision

| Format | Block layout | Hardware | Models tested | Maturity |
|---|---|---|---|---|
| **BF16** | — | All NVIDIA + AMD MI300X / MI325 / MI350 / MI355X | All | Baseline |
| **FP8 block-wise** (DeepSeek-style) | 128×128, FP32 scales | Hopper (H100 / H200), Blackwell (B200+) | Qwen3-4B, Qwen3-30B-A3B | Generally available |
| **MXFP8** | 1×32, UE8M0 scales | Blackwell only (B200, B300, GB200, GB300) | Qwen3-30B-A3B, DeepSeek-V3.2 | Beta |
| **NVFP4** (E2M1) | 1×16, two-level (FP8 + FP32) scales | Blackwell only (B200, B300, GB200, GB300) | Qwen3-30B-A3B | Beta |

## Rollout × training compatibility

<!-- 兼容矩阵是本指南的决策核心：只有对角线（同一精度贯穿 rollout 与 train）以及
     FP8 的两条横线是支持组合，MXFP8 与 NVFP4 之间、以及跨格式混搭均被排除。
     参考脚本 scripts/run_qwen3_30b_a3b.py 一次只允许一个 rollout 精度加一个训练精度，
     因此端到端配方必须成对使用 --rollout-* 与 --train-* 两个 flag。 -->

| Rollout \\ Train | BF16 | FP8 block-wise | MXFP8 | NVFP4 |
|---|---|---|---|---|
| **BF16** | ✅ baseline | ❌ | ❌ | ❌ |
| **FP8 block-wise** | ✅ | ✅ Hopper + Blackwell | ❌ | ❌ |
```

| **MXFP8** | ❓ | X | ❓ Blackwell | X |
| **NVFP4** | X | X | X | ❓ Blackwell |

Unified MXFP8 (Blackwell)

MXFP8 用一维 microscaling 块：32 个连续 E4M3 值共享一个 UE8M0 scale。
端到端配方让 rollout、前向、weight-gradient 与 data-gradient GEMM 全部走 MXFP8，
而配置为高精度的张量保留在 BF16。

```bash
python scripts/run_qwen3_30b_a3b.py prepare --hardware B200 --rollout-mxfp8 --train-mxfp8
python scripts/run_qwen3_30b_a3b.py execute --hardware B200 --rollout-mxfp8 --train-mxfp8
```

<!-- SGLang 的 MoE 与 dense 线性 GEMM 后端是独立选择的。旧文档只给单点建议，
新版把两组选择拆开，读者需按模型、并行布局与已安装的 SGLang 栈配对挑选。 -->

Workload	Miles flag	Backends
MoE	`--sglang-moe-runner-backend`	`flashinfer_trtllm_routed`, `flashinfer_trtllm`, `cutlass`
Dense linear GEMM	`--sglang-fp8-gemm-backend`	`flashinfer_trtllm`, `flashinfer_cutlass`, `triton`

NVFP4 (Blackwell)

NVFP4 把 E2M1 权重按 16 值块存储，每块一个 E4M3 scale 外加外层 FP32 scale。
gate 与 up projection 一起量化，使融合 rollout GEMM 使用同一个外层 weight scale；
trainer、checkpoint 转换器与 rollout kernel 必须遵守同一套 scaling 契约。

<!-- 后向模式是 NVFP4 章节的核心权衡：基础配方保留 high-precision backward，
dequantized backward 与 four-over-six 作为单独引用的高级选项列出，来源分别是
LMSYS 博客与 humans& 的 NVFP4 / 4-over-6 分析；实现层面的分档策略并未合入本 PR。 -->

<!-- 与代码对齐修正：NVTE_FP8_BLOCK_SCALING_FP32_SCALES 不再要求用户手工设置，
已由 miles/ray/train/actor_factory.py 在 actor env 中按硬件默认：
Hopper 上为 1，Blackwell 上为 0（TransformerEngine 用 MXFP8 模拟 block-wise 配方，
需要 2 的幂 scale）。只有明确想覆盖默认时才需要手动设置。 -->

评论区精华

该 PR 的 review 过程几乎没有文字交锋：comments_count 为 0、review_comments_count 为 0，仅有一条 Zhichenzzz 的 APPROVED 空评论。真正的讨论体现在 13 个 commit 的演进与合并 commit 里：

- 页面重命名的范围决策（commit 211f773c）：页面已覆盖 block-wise FP8、MXFP8 与 NVFP4，「fp8- 前缀不足以概括」，因此把 fp8-low-precision 重命名为 low-precision，同步更新 23 个引用页、导航项并加永久重定向；同时顺带修正图片相对路径与 NVTE 环境变量的过时指引。

- NVFP4 章节与代码对齐（commit 575f3c06）：指出「NVFP4 一节只列了环境变量，训练 flags 读起来与 MXFP8 完全相同」，对照代码补齐 NVFP4 自己的 flags；并把 10 处异名链接（"FP8 & Low Precision"、"Unified FP8" 等）统一到页面标题「Low Precision RL」。
- four-over-six 表述反复打磨（commit 3d66db57 / 226d00ed）：先收紧描述、后又恢复 humans& 博客引用，最终以「单独引用的高级选项 + 外部来源」的方式定稿。
- 合并冲突处理（commit 304f90f1）：与 #2395 在 docs.json 的 redirects 数组同一位置冲突，保留全部四条；并专门核验 #2381 新增的三个配方页（Kimi-K3、Nemotron-3-Ultra、Gemma-4）没有引用低精度页，确认 24 处链接全部一致、docs.json 解析通过、83 条导航项可解析。
- 页面重命名 fp8-low-precision → low-precision 的范围与一致性 (documentation): 以「Low Precision RL」为统一标题，全部 24 处链接一致；docs.json 新增永久重定向，外部旧链接不中断。
- NVFP4 训练 flags 与 NVTE 环境变量说明对照代码修正 (documentation): NVFP4 章节补齐独立的训练 flags；NVTE 环境变量改为「按硬件默认：Hopper 为 1、Blackwell 为 0，仅覆盖默认时才需手动设置」。
- docs.json redirects 数组与 #2395 的合并冲突 (other): 四条 redirect 全部保留；docs.json 解析通过，83 条导航项均解析成功，重命名清扫完整。

风险与影响

- 风险：
  - 零代码风险：纯文档变更，无运行时、测试、CI 影响；PR body 明确 Non-goal 为不改任何 runtime 或 recipe 代码。
  - 文档与代码偏差风险：主要风险是说明与实现脱节。本 PR 已对照代码修正两处（NVTE env var 的硬件默认逻辑、NVFP4 训练 flags），但 NVFP4 的 dequantized backward 与 four-over-six 高级模式仍在 roadmap 演进中（issue #615 里相关实现 PR 部分仍未合入），后续实现变化需要持续同步文档。
  - 站点构建验证缺口：Mintlify CLI 未在本地运行（Homebrew Node 无法加载 libllhttp.9.3.dylib），页面渲染效果未经本地验证；但 git diff --check、本地引用检查（3 个 PNG + 5 个仓库路径）、curl 200 检查与 pre-commit 全部通过。
  - 链接与重定向风险：依赖人工核验 24 处引用一致；合并 commit 已确认 83 条导航项可解析、#2381 新页面无遗漏引用，残留风险低。
- 影响：
  - 对读者：文档导航路径变化（/advanced/fp8-low-precision → /advanced/low-precision），旧 URL 由 301 永久重定向兜底；术语统一为「Low Precision RL」，消除此前同一页面被称作 FP8 & Low Precision / Unified FP8 等多达 10 种叫法的歧义。
  - 对系统：零代码、零运行时、零测试影响。
  - 对团队：文档与 issue #615 的 Blackwell MXFP8 / NVFP4 roadmap 形成闭环，为后续 NVFP4 成熟度从 Beta 升级预留结构；重命名时「改页面 + 改 23 处链接 + 加重定向 + 核验新增页面」的操作流程是文档 -as-code 的范本，值得其他文档 PR 借鉴。

- 风险标记：纯文档变更（零代码风险），Mintlify 构建未本地验证，全站 24 处链接一致性，NVFP4 高级模式仍在演进

关联脉络

- PR #2395 docs: drop the Platforms section and refresh the hardware list: 与 #2396 在 docs.json 的 redirects 数组同一位置产生合并冲突，合并时保留全部四条重定向；两者同属文档站结构整理。
- PR #2381 docs: recipe pages for Kimi-K3, Nemotron-3-Ultra and Gemma-4: 合并 commit 专门核对了这三个新增配方页没有引用低精度页，确认重命名后的 24 处链接一致。
- PR #2391 docs: replace DeepSeek V3/R1 page with a DeepSeek-V3.2 recipe: 本 PR 把 deepseek-v3-2.md 的低精度页链接同步为 /advanced/low-precision，且 DeepSeek-V3.2 正是 MXFP8 章节列出的目标模型之一。
- PR #2375 docs: reorganize the Training Backends section, drop the experimental FSDP framing: 同期文档刷新趋势的代表：都去掉 experimental 措辞、让文档与代码现状对齐。
- PR #2355 [fix] fix the bugs/outdated commands in .sh scripts and the corresponding snapshots: 修复了 scripts/run_qwen3_30b_a3b.py 等脚本的命令问题，而本 PR 文档中的可运行配方命令直接引用该 launcher。