

PR #7872 完整报告

verl-project/verl

[doc] chore: add NPU guide for v1 trainer separate async

合并时间: 2026-09-15 17:32

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7872>

执行摘要

- 一句话: 补充昇腾 NPU 上 Checkpoint Engine 后端配置文档
- 推荐动作: 该 PR 无需精读代码, 建议用 3-5 分钟浏览文档正文即可, 重点关注三个设计决策: (1) nccl 在 NPU 上被隐式解析为 HCCL——这是一个“配置名与实现后端不一致”的接口约定, 值得确认其实现位置并在代码侧保留注释, 避免后续被误当作 bug 改掉; (2) Mooncake 在 NPU 上依赖源码编译 ascend_direct, 属于部署前置条件, 建议在昇腾镜像或安装脚本文档中交叉引用; (3) 参数表默认值 (2048 MB 桶、group_name=default、rebuild_group=False) 与实现的一致性没有自动化保障, 建议后续修改 checkpoint engine 的 PR (如 Mooncake 相关修复) 把更新本文档纳入 checklist。若团队有英文文档镜像, 建议同步一份以覆盖英文用户。

功能与动机

PR body 明确写出动机: “Document that on Ascend NPU, checkpoint_engine.backend=nccl is automatically resolved to the HCCL backend, and that the mooncake backend is supported, with parameter tables and usage examples.” 结合新增内容可拆为三点: (1) v1 trainer 的 separate async (训练侧 / 推理侧分离) 属于 off-policy 场景, 文档原文写明“该场景要求配置非 **naive** 的后端”, 但此前昇腾教程没有任何后端选择指引; (2) NPU 上 **nccl** 会自动落到 HCCL, 这是平台适配层的隐式行为, 不写文档用户无法从配置名推断; (3) Mooncake 在 NPU 上需显式设置 device, 且没有 Ascend 预编译包、必须从源码编译, 属于必须在部署前告知的约束。

实现拆解

1. 文档挂载点选择: 变更集中在 docs/ascend_tutorial/zh/feature_support/ascend_backend_features.md, 追加 51 行、删除 0 行, 新章节以 --- 分隔符接在文件末尾 (原有 veomni / MoE 优化等小节未改动), 因此对已有文档结构零破坏, 但也不会被原有导航目录自动收录。
2. 总述段落定调: 新增“## Checkpoint Engine 后端”段落, 先解释 Checkpoint Engine 是“训练侧与推理侧之间同步权重的统一抽象层”, 并点名其典型使用场景是 v1 trainer 的 separate async off-policy 训练, 同时声明“该场景要求配置非 **naive** 的后端”, 把后端选择与具体训练模式绑定起来。
3. HCCL 小节 + 参数表: 说明 actor_rollout_ref.rollout.checkpoint_engine.backend=nccl 在 NPU 上自动识别为 HCCL、由训练侧以 broadcast 方式下发权重; 参数表列出 backend、

update_weights_bucket_megabytes（默认 2048，双缓冲使显存开销为 2 倍桶大小，且需大于最大权重张量）、engine_kwargns.nccl.group_name（默认 default）、engine_kwargns.nccl.rebuild_group（默认 False）四项，这四行是用户最容易配错的核心信息。

4. Mooncake 小节 + 参数表 + 外链：说明 mooncake 走 p2p 权重同步，NPU 上必须通过 engine_kwargns.mooncake.device=npu 触发 ascend_direct 传输协议，另有 device_name 设备名过滤参数；并显式提示“Mooncake Transfer Engine 没有 Ascend 预编译包，需从源码编译安装”，附上 Mooncake 官方 transfer-engine ascends direct 文档链接。
5. 使用示例与配套缺失：末尾 bash 示例同时给出 trainer.use_v1=True、trainer.v1.trainer_mode=separate_async 与两种后端的互斥配置，用户可直接复制。本次没有任何代码、配置、schema、测试或 CI 配套改动，PR checklist 中“Add unit or end-to-end test(s)”一项未填具体内容，属于可接受的文档类豁免，但也意味着文档中的断言没有自动化校验。

关键文件：

- docs/ascend_tutorial/zh/feature_support/ascend_backend_features.md（模块 昇腾文档；类别 docs；类型 documentation；符号 actor_rollout_ref.rollout.checkpoint_engine.backend, actor_rollout_ref.rollout.checkpoint_engine.update_weights_bucket_megabytes, actor_rollout_ref.rollout.checkpoint_engine.engine_kwargns.nccl.group_name, actor_rollout_ref.rollout.checkpoint_engine.engine_kwargns.nccl.rebuild_group）：本次唯一变更文件，新增“Checkpoint Engine 后端”章节（+51/-0），覆盖 HCCL 与 Mooncake 两个后端的参数表、部署约束与 v1 trainer separate async 配置示例，是 NPU 用户配置权重同步后端的直接依据。

关键符号：未识别

关键源码片段

docs/ascend_tutorial/zh/feature_support/ascend_backend_features.md

本次唯一变更文件，新增“Checkpoint Engine 后端”章节（+51/-0），覆盖 HCCL 与 Mooncake 两个后端的参数表、部署约束与 v1 trainer separate async 配置示例，是 NPU 用户配置权重同步后端的直接依据。

昇腾 NPU 上的 Checkpoint Engine 后端（HCCL / Mooncake）

以下内容为文档新增章节的核心部分，落在 docs/ascend_tutorial/zh/feature_support/ascend_backend_features.md 末尾，用于说明 v1 trainer separate async 场景下训练侧与推理侧的权重同步后端如何选择。

verl 参数	说明
actor_rollout_ref.rollout.checkpoint_engine.backend	设为 nccl 时在 NPU 上自动解析为 HCCL；设为 mooncake 时走 p2p 同步

verl 参数	说明
<code>actor_rollout_ref.rollout.checkpoint_engine.update_weights_bucket_megabytes</code>	单次权重传输的桶大小（MB），默认 2048；HCCL 双缓冲收发使显存开销为 2 倍桶大小，且该值必须大于模型中最大权重张量的内存占用
<code>actor_rollout_ref.rollout.checkpoint_engine.engine_kwargs.nccl_group_name</code>	HCCL 进程组名称，默认 <code>default</code>
<code>actor_rollout_ref.rollout.checkpoint_engine.engine_kwargs.nccl_rebuild_group</code>	是否在每次权重更新时重建 HCCL 进程组，默认 <code>False</code>
<code>actor_rollout_ref.rollout.checkpoint_engine.engine_kwargs.mooncake.device</code>	Checkpoint Engine 运行设备，NPU 上必须设为 <code>npu</code> ，触发 <code>ascend_direct</code> 传输协议
<code>actor_rollout_ref.rollout.checkpoint_engine.engine_kwargs.mooncake.device_name</code>	Mooncake 设备名过滤，默认值为空

以下示例来自文档新增的「使用示例」小节

```
# 1) 打开 v1 trainer 的 separate async 模式：训练侧与推理侧分离，
# 属于 off-policy 场景，要求 checkpoint_engine.backend 不能保持默认的 naive
trainer.use_v1=True
trainer.v1.trainer_mode=separate_async
```

```
# 2) 方式一：nccl 后端
# 在昇腾 NPU 上会被自动识别为 HCCL，训练侧以 broadcast 方式把权重
# 发送到各 rollout worker，用户无需为新平台额外修改配置
actor_rollout_ref.rollout.checkpoint_engine.backend=nccl
```

```
# 3) 方式二：mooncake 后端
# 基于 Mooncake Transfer Engine 以 p2p 方式同步权重；NPU 上必须显式
# 指定 device=npu，Transfer Engine 才会使用 ascend_direct 传输协议。
# 注意：Mooncake Transfer Engine 没有 Ascend 预编译包，需要从源码编译安装
# （文档中给出了 transfer-engine: ascend direct 的外链）
actor_rollout_ref.rollout.checkpoint_engine.backend=mooncake
actor_rollout_ref.rollout.checkpoint_engine.engine_kwargs.mooncake.device=npu
```

评论区精华

本次 PR 没有产生实质性的技术讨论。reviewer [wucong25](#) 先后两次给出 `APPROVED`，review body 均为空，也没有任何行内 review comment；唯一的外部评论来自 `CLAassistant` 的自动校验：“All committers have signed the CLA.”。提交历史只有 1 个 commit（[0028829](#)），作者 [dodatboii](#) 一次性提交完整文档后即被合入，没有返工或改写。因此文档中

关于 `nccl` 自动解析为 HCCL、mooncake 需 `device=npu`、桶大小需大于最大权重张量、双缓冲显存开销为 2 倍桶大小等关键断言，全部依赖作者自述，未经 reviewer 逐条对照实现核对。

- 文档类 PR 的 review 深度与准确性背书 (documentation): 文档变更被直接判定为低风险并合并，没有围绕技术准确性 (`nccl` → HCCL 自动解析、mooncake 需 `device=npu`、桶大小与双缓冲显存约束) 形成任何讨论记录，文档中的关键断言缺少 reviewer 交叉验证。

风险与影响

- 风险:

1. 文档与实现漂移风险 (主要风险)：文档是 NPU 用户配置 checkpoint engine 的主要信息来源，但 `nccl` → HCCL 的自动解析、mooncake.device=npu 触发 `ascend_direct`、各默认值 (2048 MB、`group_name=default`、`rebuild_group=False`) 都硬编码在正文里，没有任何自动化校验或指向代码位置的引用；后续任何 checkpoint engine 改动都可能让这份表格失准。
2. 断言未经交叉验证：review 无行内评论，CI 也不会校验文档内容，若某条参数名或默认值写错 (例如 `engine_kwargs.mooncake.device_name` 的实际语义)，用户会在 NPU 环境上直接踩坑，而反馈链路很长。
3. 覆盖范围不完整：文档只写在 `docs/ascend_tutorial/zh/` 下，仓库是否存在对应的英文镜像文档在本次材料中未体现；若存在，英文读者与镜像文档会缺失同一份关键信息。
4. 无运行时风险：不涉及代码路径、性能、安全或兼容性变更，回滚成本近似为零。
 - 影响：对 NPU 用户：可直接按参数表配置 `nccl` (自动 HCCL) 或 mooncake (需源码编译 + `device=npu`)，显著降低 v1 trainer separate async 在昇腾上的试错成本，尤其是“桶大小需大于最大权重张量、双缓冲占 2 倍显存”这类容易导致 OOM 的隐性约束被显式写出。对系统运行时：无任何影响，不修改代码、配置或 CI。对团队维护：把散落在代码与 issue 中的平台适配知识沉淀为正式文档，减少重复答疑；但文档本身成为需要随实现同步维护的资产。整体影响范围局限于昇腾文档站点，程度为“体验与效率改进”，不改变任何能力边界。
 - 风险标记：纯文档变更，参数默认值缺少自动化校验，断言未经 reviewer 交叉核对，仅中文文档，可能缺英文同步

关联脉络

- PR #7764 [ckpt] fix: drain Mooncake completion slots across weight versions: 修改 `verl/checkpoint_engine/mooncake_checkpoint_engine.py`，正是本文档 Mooncake 小节所描述的后端实现；该 PR 修复跨权重版本完成槽位未消费问题，与文档强调的权重同步流程同一条链路。
- PR #7762 [trainer, ckpt] fix: skip V1 async warmup after inflight reissue: 处理 v1 trainer 异步 / separate async 的 warmup 与 inflight 重发逻辑，正是本文档示例配置 `trainer.use_v1=True`、`trainer.v1.trainer_mode=separate_async` 所面向的运行场景。
- PR #7728 [rollout] fix: default IPC weight transfer on Ascend A5 hardware: 同为昇腾 NPU 上权重传输链路的硬件适配与默认行为调整 (涉及 `verl/plugin/platform/platform_npu.py`)，与本文档描述的 NPU 侧权重同步后端选择属于同一主题域。

- PR #7729 [doc] fix: fix old URL of ascend doc: 同样维护 docs/ascend_tutorial/zh/ 下的文档内容（含失效链接修复），与本次在同一文档目录内补充内容的维护模式一致。