

THUDM/slime 2026 第 23 周 (06-01 至 06-07) 周报

THUDM/slime

周期: 2026-06-01 至 2026-06-07

来源 PR: 10 • 重点 PR: 10 • 自动生成

原文链接: <http://prhub.com.cn/THUDM/slime/reports/2026-06-01-to-2026-06-07>

执行摘要

本周 (06-01 至 06-07) 共合并 10 个 PR, 集中在三项重大架构升级: FLOPs 感知调度、外部引擎集成和权重同步加速。其中 zhuzilin 贡献了 7 个 PR, 是本周的核心贡献者。重点模块包括数据并行调度 (dp_schedule.py)、rollout 引擎管理 (slime/backends/sglang_utils/) 和检查点保存 (hf_checkpoint_saver.py)。整体方向是提升训练效率与部署灵活性, 但测试覆盖不足和作者集中度需要引起管理关注。

本周重点变化

1. FLOPs 平衡调度全面统一

- PR2017引入了 `--balance-by-flops` 标志, 通过 Karmarkar-Karp 分区基于 FLOPs 估计进行微批次分配, 实验显示 actor 训练时间减少 12–30%, 吞吐提升 4–24%。该 PR 由 HaoDong0027 提交, 讨论中采纳了使用精确 `calculate_fwd_flops` 代替近似公式的建议。
- PR2029在此基础上将 FLOPs 平衡从仅用于 mbs 内部 packing 扩展为同时驱动跨 DP rank 的微批次分发, 统一了调度逻辑, 移除了旧的 token 权重分支。
- PR2028删除了不再被调用的 `calculate_workload` 函数, 完成了重构的清理工作。
- 影响: 调度模块 (`dp_schedule.py`, `seqlen_balancing.py`, `arguments.py`) 经历重大重构, 性能收益显著, 但同时也引入了核心路径变更和 FLOPs 估计误差风险。

2. 外部引擎支持重构

- PR2016移除了旧的 `--rollout-external` 开关, 通过 `--rollout-external-engine-addr` 指定预启动的外部 SGLang 引擎列表。新增 `external.py` 模块, 通过 HTTP 查询 `/server_info` 自动推断引擎属性并注册到 PD 路由器。同时配套修改 `placement_group` 布局, 外部引擎模式下不占用 trainer GPU。
- PR2022同步增加了中英文文档, 提供了决策流程、配置指南和与 `--sglang-config` 的关系说明。
- 影响: 外部引擎集成彻底重构, 降低配置复杂度, 但缺少集成测试, 且依赖 HTTP 服务的健壮性有待长期验证。

3. 权重同步与保存优化

- PR2021新增 `UpdateWeightFromDisk` 类, 实现基于共享文件系统的全权重同步: 训练器将完整 HF 检查点写入磁盘, 引擎从磁盘加载。同时重构 `hf_checkpoint_saver.py` 支持任意

路径保存，弃用 `--update-weight-delta-dir` 并新增 `--update-weight-disk-dir`。

- PR2020实现了多节点并行写入 HF 权重 shard：每个节点负责一部分 chunk，最后合并 state 并写入 index。这解决了大模型串行写入的瓶颈。
- 影响：保存速度大幅提升，但依赖共享文件系统，且节点数超过 chunk 数时可能导致部分节点闲置。

4. 回溯与清理

- PR2013回退了 #1984 的 `rollout_id`→`group_id` 重命名，因为与内部工具冲突。所有相关代码恢复旧命名，并移除了向后兼容代码。
- PR2028删除死代码，体现持续维护文化。

模块与主题趋势

- 调度模块： `dp_schedule.py`、`seqlen_balancing.py`、`arguments.py` 更新频繁，FLOPs 平衡成为核心调度策略，性能和复杂度同步增加。
- 外部引擎：新增 `external.py` 模块，与 `ray/rollout.py`、`sglang_engine.py` 紧密集成，架构解耦但集成度提升。
- 检查点保存：`hf_checkpoint_saver.py` 重写为支持并行写入，但风险提示（节点闲置）和测试不足需关注。
- 文档：中英文文档同步更新，覆盖外部引擎配置和 Ray 分布式调试，有利于新用户快速上手。
- 测试：虽然有 `test_dp_schedule.py` 等新增测试，但整体对新增功能（磁盘同步、外部引擎）的集成测试缺失，风险列表多次出现“缺少测试覆盖”。
- 作者分布：zhuzilin 贡献 7/10 PR，平均贡献度极高，核心模块知识集中，需要增加团队参与和代码审查。

风险观察

风险项	关联 PR	风险等级	说明
测试覆盖不足	PR2017, PR2029, PR2016, PR2021, PR2020	高	新调度策略和外部引擎缺少大规模性能和集成测试
作者集中度	全部	中	zhuzilin 主导核心变更，知识孤岛风险
FLOPs 估计误差	PR2017, PR2029	中	依赖模型配置，系数不准确时可能性能不及预期
外部依赖裸奔	PR2016	高	引擎 HTTP 自发现，故障时无降级机制
废弃参数兼容	PR2021	中	<code>--update-weight-delta-dir</code> 弃用后需确保平滑迁移

重点 PR 速览

- PR2017(6.8): feat: add `--balance-by-flops` for FLOPs-balanced micro-batching — 引入 FLOPs 感知调度，性能提升显著，是调度改进的起点。
- PR2016(9.4): Fully support `--rollout-external-engine-addr`s — 外部引擎集成架构大升级，HTTP 自发现设计精致，风险亦高。
- PR2021(9.2): Support `update_from_disk` — 新增磁盘全量模式，共享文件系统场景所需。
- PR2020(7.9): Accelerate raw HF save with node writers — 并行写入加速，应对大模型 checkpoint 瓶颈。
- PR2029(6.6): use `balance_by_flops` as balance data across mbs — 统一调度逻辑，使 FLOPs 平衡成为默认行为。
- PR2013(8.2): Revert "rename rollout_ids to group_ids" — 避免内部工具冲突，强制回退，值得关注后续命名规范。
- PR2022(4.4): Add docs for external servers — 完善外部引擎文档，提升可用性。
- PR1995(4.5): fix(multi-agent): preserve rollout logprobs — 修复多智能体 GRPO 训练崩溃，由 Jiang020609 贡献。
- PR2001(3.8): [docs] Add step-by-step debug tutorial — Ray 分布式调试教程，提升开发体验。
- PR2028(5.2): remove abundant function — 删除死代码，维护清理。

后续建议

- 提升测试覆盖：针对 FLOPs 调度、外部引擎和磁盘同步添加集成测试与性能回归测试，确保新功能稳定。
- 平衡贡献分布：鼓励更多团队成员参与核心模块开发，降低对单一作者依赖，加强 review 过程。
- 健壮性增强：为外部引擎路径增加连接重试、超时处理和降级方案；为 FLOPs 估计提供验证机制（如空运行检测）。
- 废弃参数迁移：明确 `--update-weight-disk-dir` 的迁移时间线，保障用户平滑过渡。
- 持续代码清理：延续 PR2028 的做法，定期清理死代码和过时 API，保持代码库健康。