PR #1733 完整报告

radixark/miles

[AMD] Drop inert DSv4 rollout knobs and add an MTP recipe

合并时间：2026-07-30 04:14

原文链接：http://prhub.com.cn/radixark/miles/pull/1733

执行摘要

- 一句话：清理 AMD 脚本失效参数并新增 MTP 推测解码
- 推荐动作：值得精读，尤其是它对“失效配置”的识别方法和通过提交历史逐步收敛变更的过程，对于维护长生命周期启动脚本很有参考价值。MTP recipe 的配置方式也可迁移到其他 AMD 模型脚本。

功能与动机

PR body 指出：DeepseekV4ForCausalLM 在 HIP 上运行时，ServerArgs 会立即覆盖脚本设置的环境变量，因此脚本写入大量变量实际是无效的。需要删掉这些 inert knobs，同时利用 checkpoint 已有的 MTP block 开启推测解码，以提升 rollout 性能。

实现拆解

实现拆解

1. 新增 --enable-mtp 配置开关：在 scripts/amd/run_deepseek_v4.py 的 ScriptArgs dataclass 中添加 enable_mtp: bool = False 字段，默认关闭，不影响现有行为。
2. 删除无效配置：从 sglang_args 中移除 --sglang-dsa-topk-backend torch、--sglang-disable-custom-all-reduce、--use-miles-router 三个 CLI 参数；从 extra_env_vars 中删除 13 个冗余环境变量，包括 SGLANG_DSA_TOPK_BROADCAST、SGLANG_OPT_USE_TOPK_V2、SGLANG_OPT_USE_MULTI_STREAM_OVERLAP 等，仅保留仍有效的 SGLANG_OPT_USE_TILELANG_INDEXER、SGLANG_OPT_USE_COMPRESSOR_V2、SGLANG_OPT_USE_FUSED_COMPRESS 等。
3. 新增 MTP 推测解码配置：当 enable_mtp=True 时，向 sglang_args 追加 --sglang-speculative-algorithm EAGLE 等参数，配置 3 步推测、top-k 1、4 个 draft token；同时设置环境变量 SGLANG_USE_AITER_AG=false，在 gfx950 上强制使用 RCCL all-gather，避免 aiter 在推测解码场景下的死锁。
4. 演进过程：从提交历史看，最初还引入了 SGLANG_MEMORY_SAVER_CUDA_GRAPH 和调整内存分数，但后续提交发现 miles 已默认设置该变量，因此移除；最终只保留与 MTP 和参数清理相关的变更。

未涉及测试文件，但作者在真实集群上完成验证（3 个训练步骤无引擎重启，accepted length 2.75–2.79）。

关键文件：

- scripts/amd/run_deepseek_v4.py（模块 部署脚本；类别 source；类型 core-logic）：唯一的改动文件，负责 DeepSeek-V4 在 AMD 上的训练启动配置。PR 的核心变更均在此文件：清理无效 rollout 参数、新增 MTP 推测解码配置开关。

关键符号：_train, ScriptArgs

关键源码片段

scripts/amd/run_deepseek_v4.py

唯一的改动文件，负责 DeepSeek-V4 在 AMD 上的训练启动配置。PR 的核心变更均在此文件：清理无效 rollout 参数、新增 MTP 推测解码配置开关。

```
# scripts/amd/run_deepseek_v4.py —— MTP 配置与清理后的关键片段
@dataclass
class ScriptArgs(U.ExecuteTrainConfig):
    ...
    # 新增开关：启用后使用 checkpoint 自带的 MTP block 作为 EAGLE draft
    enable_mtp: bool = False

def _train(args: ScriptArgs):
    ...
    # sglang 基础参数，已移除 --sglang-dsa-topk-backend 和 --sglang-disable-custom-all-reduce
    sglang_args = (
        f"--rollout-num-gpus-per-engine {sglang_world_size} "
        f"--sglang-tp-size {sglang_tp_size} "
        f"--sglang-dp-size {sglang_dp_size} "
        f"--sglang-ep-size {sglang_ep_size} "
        "--router-health-success-threshold 1 "
        "--router-health-check-interval-secs 15 "
        "--router-health-failure-threshold 40 " # TODO improve
    )
    # 环境变量精简为真正生效的项
    extra_env_vars = {
        "SGLANG_SKIP_CHECKPOINT_LOAD_CHECK": "1",
        "SGLANG_DSV4_FP4_EXPERTS": "0",
        "SGLANG_HACK_FLASHMLA_BACKEND": "triton",
        "SGLANG_OPT_USE_TILELANG_INDEXER": "true",
        "SGLANG_OPT_USE_COMPRESSOR_V2": "false",
        "SGLANG_OPT_USE_FUSED_COMPRESS": "true",
        "SGLANG_HEALTH_CHECK_TIMEOUT": "120",
        "AITER_BF16_FP8_MOE_BOUND": "0",
    }
    ...
    if args.enable_mtp:
        sglang_args += (
            "--sglang-speculative-algorithm EAGLE "
            "--sglang-speculative-num-steps 3 "
            "--sglang-speculative-eagle-topk 1 "
            "--sglang-speculative-num-draft-tokens 4 "
```

```
)
# gfx950: aiter all-gather 在推测解码下可能死锁，回退 RCCL
extra_env_vars |= {"SGLANG_USE_AITER_AG": "false"}
```

评论区精华

本 PR 无实质 review 评论。guapisolo 直接批准（LGTM），gemini-code-assist 机器人自动审查未提出意见。提交历史中多次与 main 合并并调整，未发现公开讨论中的设计争议。

- Review 结论 (other): 无争议，PR 合入。

风险与影响

- 风险：风险点包括：
  - 删除的配置项被官方声明为 inert，但若某些集群上 ServerArgs 覆盖逻辑不同，可能导致行为变化，需依赖实测验证。
  - 启用 MTP 会引入 EAGLE 推测解码，增加显存占用和调度复杂度，且在 gfx950 上强制回退 RCCL all-gather，可能带来通信性能下降。
  - 脚本仅针对 AMD 平台，影响面局限于该启动路径。
  - 缺少自动化测试覆盖，回归依赖人工验证。
  - 影响：影响范围限定在 scripts/amd/run_deepseek_v4.py，面向 AMD MI series 上的 DeepSeek-V4 训练 / 推理任务。对现有用户而言，默认行为不变（enable_mtp 默认关闭）；启用 MTP 后，rollout 吞吐有望提升，但需关注稳定性。对团队而言，脚本配置更精简，减少了误导性参数。
  - 风险标记：缺少测试覆盖，AMD 专属配置变更

关联脉络

- PR #1926 Fix weight sync selector for frozen speculative drafts: 同属 speculative-decoding 相关修复，涉及 draft 权重同步，与 MTP 作为 EAGLE draft 存在逻辑关联。
- PR #2028 session: collect speculative-decoding counters: 同为 speculative-decoding 功能线，补充计数器采集，可为本 PR 的 MTP 效果观测提供支撑。