PR #2139 完整报告

radixark/miles

[AMD] retune AMD 4-node dsv4 config

合并时间：2026-08-05 02:04

原文链接：http://prhub.com.cn/radixark/miles/pull/2139

执行摘要

- 一句话：AMD DSV4 4 节点配置调优，切换 unified-KV 路径
- 推荐动作：值得精读。虽然只改一个文件，但涉及 SGLang 后端路径切换和并行策略权衡，对理解 AMD 平台上的优化思路有帮助。可以重点看 extra_env_vars 的配置注释和 _get_parallel_config 的并行策略选择。

功能与动机

PR 没有关联 Issue，但根据提交信息和 Issue 评论，主要动机是让 AMD DeepSeek-V4 rollout 走 unified-KV 路径以获得更好的 KV 缓存 / 性能表现，并在该路径下重调 4 节点配方。代码注释指出 unified-KV 仅存在于 compressor_v2 中，旧路径在 HIP 上会导致 compress_kv_pool 未设置而触发断言，因此切换是必要的。XinyuJiangCMU 提供的基准表明调优后 8k 稳态 step 的 rollout 约 150s、actor 训练约 182s、完整循环约 7.9 分钟。

实现拆解

1. 切换推理后端：在 _train 的 extra_env_vars 中把 SGLANG_HACK_FLASHMLA_BACKEND 从 triton 改为 unified_kv_triton，并新增 SGLANG_OPT_USE_COMPRESSOR_V2 为 true。这是本次变更的入口，因为 unified-KV 路径只在 compressor_v2 中实现，若不显式启用，HIP 上会因 compress_kv_pool 未设置而触发内存池断言。同时将 SGLANG_OPT_USE_JIT_NORM 从 false 改为 true，表示在新后端下此前的 logprobdiff 顾虑不再成立。
2. 重调 4 节点并行配置：修改 _get_parallel_config，将 4 节点（32 GPU）的 --tensor-model-parallel-size 从 8 降为 4，形成 TP4/PP4/EP8 布局，并在注释中标注。降低 TP 可减少跨节点通信量，同时保持 EP8 的专家并行。该函数只对已验证配置返回参数，其他规模会抛 NotImplementedError。
3. 调整显存与上下文参数：将 --sglang-mem-fraction-static 从 0.7 降至 0.5，与 4 节点 PP4 下部分优化器 offload（--optimizer-offload-fraction 0.75）的内存预算匹配（注释同步更新）；将 dapo_aime 任务的 --rollout-max-response-len 从 4096 提升到 8192，以支持更长推理上下文。
4. 验证：无新增测试，仅依赖手工跑数。Issue 评论提供性能基准：8k 稳态 rollout 约 150s、actor 训练约 182s、全循环约 7.9 min。

关键文件：

- scripts/amd/run_deepseek_v4.py（模块 训练配方；类别 source；类型 core-logic；符号 _get_parallel_config, _train）：唯一改动文件，核心调优逻辑所在，包含并行配置、SGLang 环境变量和显存参数调整。

关键符号：_get_parallel_config, _train

关键源码片段

scripts/amd/run_deepseek_v4.py

唯一改动文件，核心调优逻辑所在，包含并行配置、SGLang 环境变量和显存参数调整。

```
# scripts/amd/run_deepseek_v4.py
def _get_parallel_config(args: ScriptArgs) -> str:
    """Return parallel config args for tested GPU configurations.

    Only includes configurations that have been verified to work.
    Raises NotImplementedError for untested configurations.
    """
    actor_num_nodes = args.actor_num_nodes
    actor_num_gpus_per_node = args.actor_num_gpus_per_node
    total_gpus = actor_num_nodes * actor_num_gpus_per_node

    # 单节点 smoke-test 配置
    if actor_num_nodes == 1:
        return (
            f"--tensor-model-parallel-size {actor_num_gpus_per_node} "
            "--sequence-parallel "
            "--pipeline-model-parallel-size 1 "
            "--context-parallel-size 1 "
            f"--expert-model-parallel-size {actor_num_gpus_per_node} "
            "--expert-tensor-parallel-size 1 "
        )

    if actor_num_gpus_per_node == 8:
        if total_gpus == 32: # 4 nodes x 8 GPUs (MI355X, full Flash): TP4/PP4/EP8, 43 layers =
        11+11+11+10
            # 相对旧配方 TP8/PP4/EP8，这里将 TP 降为 4，以减少跨节点通信
            return (
                "--tensor-model-parallel-size 4 "
                "--sequence-parallel "
                "--pipeline-model-parallel-size 4 "
                "--decoder-first-pipeline-num-layers 11 "
                "--decoder-last-pipeline-num-layers 10 "
                "--context-parallel-size 1 "
                "--expert-model-parallel-size 8 "
                "--expert-tensor-parallel-size 1 "
            )

    raise NotImplementedError(
```

```
        f"No pre-set parallel config for {total_gpus} GPUs. "
        f"Please specify your parallel config in `run_deepseek_v4._get_parallel_config`."
    )

# extra_env_vars 关键配置调整
extra_env_vars = {
    "SGLANG_SKIP_CHECKPOINT_LOAD_CHECK": "1",
    "SGLANG_DSV4_FP4_EXPERTS": "0",
    # 切换到 unified-KV 后端；注意该实现只在 compressor_v2 中提供
    "SGLANG_HACK_FLASHMLA_BACKEND": "unified_kv_triton",
    # HIP 上 v1 路径会遗留未设置的 compress_kv_pool，导致内存池断言，
    # 因此必须显式启用 compressor_v2
    "SGLANG_OPT_USE_COMPRESSOR_V2": "true",
    "SGLANG_OPT_USE_TILELANG_INDEXER": "true",
    # 在 unified-KV 路径下重新启用 JIT norm，旧配方中为避免 logprobdiff 增大而关闭
    "SGLANG_OPT_USE_JIT_NORM": "true",
    "SGLANG_OPT_USE_FUSED_COMPRESS": "true",
    "SGLANG_HEALTH_CHECK_TIMEOUT": "120",
    "AITER_BF16_FP8_MOE_BOUND": "0",
}
```

评论区精华

该 PR 没有任何 review 评论，guapisolo 直接批准。有价值的讨论来自两处：

- 代码注释是主要设计说明：unified-KV 必须搭配 compressor_v2，否则 HIP 上内存池断言。
- SGLANG_OPT_USE_JIT_NORM 从 false 翻转为 true：base 版本注释明确说 "JIT norm+RoPE increases logprobdiff on ROCm"，head 版本移除该顾虑，说明新路径不再受此影响。
- XinyuJiangCMU 在 Issue 评论中给出性能基准，验证调优效果。
- unified-KV 路径必须启用 compressor_v2 的原因 (design)：显式设置 SGLANG_OPT_USE_COMPRESSOR_V2=true，该问题被代码注释记录并解决。
- JIT norm 在 ROCm 上的启停权衡 (design)：在切换后端后重新启用 JIT norm。
- 4 节点调优性能数据 (other)：性能基准确认调优有效。

风险与影响

- 风险：核心路径变更：SGLANG_HACK_FLASHMLA_BACKEND 切换属于推理后端变更，可能影响 KV 缓存管理、数值行为和长上下文稳定性。 缺少测试覆盖：没有自动化测试，仅手工跑数，后续改动容易破坏该配方。 平台特定：只在 4 节点 MI355X 上验证，其他规模会走 NotImplementedError 分支。 JIT norm 重新启用：可能重新引入 logprobdiff 数值差异，需要训练曲线确认。 显存参数连锁：sglang-mem-fraction-static 从 0.7 降到 0.5，与优化器 offload 配置联动，若 offload 比例调整需同步。
- 影响：影响范围：仅限 scripts/amd/run_deepseek_v4.py 这一个脚本的用户，但提供了一份 AMD 上 unified-KV 的参考配方。 用户：使用该脚本在 AMD 4 节点上跑 DeepSeek-V4 训练的用户会直接受益于新的性能和更长上下文支持。 系统：SGLang 后端

路径切换会影响 KV 缓存管理、显存占用，sglang-mem-fraction-static 降低给训练留下更多显存。 团队：该 PR 为后续其他模型的 ROCm 调优提供了模板。

- 风险标记：核心路径变更，缺少测试覆盖，平台特定配置

关联脉络

- PR #1572 [optim]--rematerialize-param-from-master-weight: save the bf16 weight backup in colocate: 同为 colocate 场景的内存优化，本 PR 调整 sglang-mem-fraction-static 与优化器 offload 配置，属于同一内存调优脉络。
- PR #1606 ci(rocm): add ROCm CI workflow for MI300X self-hosted runners: 为 AMD/ROCm 平台建立 CI 基础设施，本 PR 是 AMD 平台 DSV4 训练配方的延续。
- PR #2202 fix(tito): prevent DeepSeek V4 system-tail mismatch: 同为 DeepSeek V4 在 ROCm 上的专用调优，说明 DSV4 在 AMD 平台的多线推进。