

PR #1795 完整报告

radixark/miles

Bump sglang to v0.5.16

合并时间: 2026-07-30 05:12

原文链接: <http://prhub.com.cn/radixark/miles/pull/1795>

执行摘要

- 一句话: 升级 sglang 至 v0.5.16, 适配依赖栈并修复权重更新
- 推荐动作: 值得精读, 尤其是 docker/Dockerfile 的构建健壮性实践与 miles/backends/megatron_utils/update_weight/common.py 的未标记专家权重处理。后续 sglang 升级时, 建议提前对照 v0.5.16 重构点 (WeightUpdater、RemoteInstanceWeightTransporter、get_server_args) 逐项核对 miles 侧引用, 并确认 sglang#32809 是否已进入 sglang-miles 主线。

功能与动机

PR body 明确给出升级目标: 'Bumps the sglang base image and the sglang-miles branch to v0.5.16', 并说明遵循 v0.5.15 周期的既定流程 (#1655 与 #1701): 先钉版本化分支, promote 后再收回 SGLANG_BRANCH。升级的必要性来自持续跟进 sglang 上游: v0.5.16 重构了权重更新与远程实例传输等核心模块, miles 需要同步适配才能继续使用 sglang-miles 的定制能力 (speculative decoding、spec-draft 权重同步等), 同时修复升级暴露的 TEGroupedLinear 专家权重 gather 缺陷。

实现拆解

1. 版本号与分支升级: docker/Dockerfile 将 SGLANG_IMAGE_TAG 从 v0.5.15 升到 v0.5.16 (cu13 基础镜像), docker/build.py 将 cu12-x86 构建参数同步为 v0.5.16-cu129; SGLANG_BRANCH 先指向版本化分支 sglang-miles-v0.5.16, 最终提交 f27d4510 在 sglang-miles 迁移到 v0.5.16 后收回直指 sglang-miles。
2. sglang 内部重构适配: sglang v0.5.16 将权重更新移入 ModelRunner.weight_updater (receive/load 拆分落在 WeightUpdater 上)、远程实例传输移入 RemoteInstanceWeightTransporter, 并大量改名 (DecodeCudaGraphRunner.num_tokens_per_bs → captured_req_width、TpModelWorker.tp_rank → self.ps.tp_rank、get_global_server_args() → runtime_context.get_server_args() 等, 共 25 处)。对应提交逐个修复 miles 侧引用, 并删除 worker 级 update_weights_from_* 入口, 统一走新路径。
3. Docker 构建链修复: TE 补丁应用失败从静默吞掉改为构建失败; Megatron-Bridge 按 SHA 固定到 7f0fb345 (TE 2.17 grouped-linear 契约, 即 radixark/Megatron-Bridge#24); FA3 升到 3.0.0 (TE 2.17 的 context-parallel 路径需要 window_size 元组); cuDNN 升到 9.22.0 并移除基础镜像的 apt 版 cuDNN, 避免

ldconfig 混装；最后在所有安装后强制恢复 apache-tvm-ffi==0.1.11 并校验签名以匹配 nvidia-cutlass-dsl 4.6.0。

4. CI 配套：.github/workflows/_run-ci.yml 在 reconcile requirements 前后记录镜像 cuDNN 版本并在安装后恢复 pin，避免 torchft 拉低 cuDNN 导致 TE fused attention 反向失败；修复 bash -e / -o pipefail 下探针退出问题；tests/ci/requirements-ci-cpu.txt 增加 xgrammar==0.2.1 (sglang v0.5.16 新增模块级导入，CPU 阶段 collection 需要)。
5. 测试与探针适配：tests/e2e/conftest_dumper.py 的 mlp_output 补丁从带 should_allreduce_fusion / use_reduce_scatter 参数的旧调用改为匹配 self.mlp(hidden_states, forward_batch) (v0.5.16 将前两者移入 get_forward().scoped 上下文)；tests/fast/utils/chat_template_utils/test_template.py 的 _make_serving 镜像新属性 default_chat_template_kwargs={}, 否则 174 个对齐用例 AttributeError。
6. 附带权重更新修复：miles/backends/megatron_utils/update_weight/common.py 新增 _is_unmarked_grouped_expert_weight，对 TEGroupedLinear 未标记的 per-expert 权重仍执行 all-gather，并修正 partition_dim < 0 时的默认值；FSDP IPC 路径补充 begin/end weight_update 会话（后随 main 合入的统一版本收敛为一次会话）。

关键文件：

- miles/backends/megatron_utils/update_weight/common.py (模块 权重更新；类别 source；类型 core-logic；符号 _is_unmarked_grouped_expert_weight, all_gather_param, all_gather_params_async, _check_and_fix_partition)：升级过程中暴露并修复的权重同步核心缺陷：TEGroupedLinear 不标记 per-expert 权重，导致权重被当作未分片、all-gather 被跳过，rollout 引擎收到半边专家权重，多 rank 出现越界装载。
- docker/Dockerfile (模块 镜像构建；类别 infra；类型 infrastructure)：承载本次升级主体：SGLANG_IMAGE_TAG v0.5.15→v0.5.16；同时修复 TE patch 静默失败、固定 Megatron-Bridge SHA、升级并对齐 cuDNN 与 apache-tvm-ffi，是构建链上问题最集中的文件。
- .github/workflows/_run-ci.yml (模块 流水线；类别 infra；类型 infrastructure)：解决 CI 中 requirements 依赖解析把镜像 cuDNN 拉低导致 TE fused attention 反向失败的问题，并修复 bash -e/pipefail 下探针稳定性。
- tests/e2e/conftest_dumper.py (模块 测试探针；类别 test；类型 test-coverage)：sglang v0.5.16 将 should_allreduce_fusion / use_reduce_scatter 移入 scoped 上下文，原补丁匹配失效，导致 slang scheduler 在 load_model 阶段挂起，必须重定向 mlp_output 探针。
- tests/fast/utils/chat_template_utils/test_template.py (模块 模板测试；类别 test；类型 test-coverage；符号 _make_serving)：v0.5.16 新增 default_chat_template_kwargs 属性，_make_serving 需镜像该属性，否则 174 个对齐用例 AttributeError。
- docker/build.py (模块 镜像构建；类别 infra；类型 configuration)：同步 cu12-x86 构建配置中的 SGLANG_IMAGE_TAG 到 v0.5.16-cu129，保持与 Dockerfile 一致。
- tests/ci/requirements-ci-cpu.txt (模块 CI 依赖；类别 config；类型 configuration)：sglang v0.5.16 新增 xgrammar 模块级导入且 function_call_parser 无条件引用，CPU 阶段若无 xgrammar 会 collection 失败。

关键符号: `_is_unmarked_grouped_expert_weight`, `all_gather_param`, `all_gather_params_async`, `_check_and_fix_partition`, `_make_serving`

关键源码片段

`docker/Dockerfile`

承载本次升级主体: `SGLANG_IMAGE_TAG` v0.5.15→v0.5.16; 同时修复TEpatch静默失败、固定 Megatron-Bridge SHA、升级并对齐 cuDNN 与 `apache-tvm-ffi`, 是构建链上问题最集中的文件。

```
# ===== Reconcile apache-tvm-ffi with sglang's pin =====
# sglang v0.5.16 将 nvidia-cutlass-dsl 从 4.5.2 升到 4.6.0, 但 apache-tvm-ffi 仍钉在 0.1.11。
# cutlass-dsl 4.6.0 的 tvm_ffi provider 会调用 make_kwargs_wrapper(..., map_dataclass_to_tuple=...),
# 该参数只在 apache-tvm-ffi >= 0.1.10 中存在; 本文件前面某步安装拉低了它,
# 导致 flashinfer 的 CuTe rmsnorm 路径在 CUDA-graph capture 时报
# TypeError: make_kwargs_wrapper() got an unexpected keyword argument 'map_dataclass_to_tuple'
# 因此在所有其他安装之后强制恢复 sglang 自己的 pin, 并用 inspect 校验签名, 失败则构建报错。
RUN pip install --force-reinstall --no-deps "apache-tvm-ffi==0.1.11" && \
    python3 -c "import inspect; from tvm_ffi.utils import kwargs_wrapper as k; \
    p = inspect.signature(k.make_kwargs_wrapper).parameters; \
    assert 'map_dataclass_to_tuple' in p, \
    'apache-tvm-ffi resolves to a build too old for nvidia-cutlass-dsl 4.6.0: %s' % list(p)"
```

评论区精华

Review 评论为空, 但 PR 评论区有三条有效信息:

- guapisolo 提醒 "Remem to pick this <https://github.com/sgl-project/sglang/pull/32809>", 要求跟进该上游 PR; PR 内未见针对它的显式处理记录, 最终合并时是否已包含无法从本仓库材料确认。
- Zhichenzzz 补充 "also wait for re-labing the sglang-miles to sglang-miles-v0.5.16", 即版本化分支就绪前不应合并; 最终提交 f27d4510 将 `SGLANG_BRANCH` 指回已迁移到 v0.5.16 的 `sglang-miles`, 该问题已解决。
- 三位 reviewer (Zhichenzzz、guapisolo、yushengsu-thu) 均 APPROVED, guapisolo 仅回复 "LGTM"。
- 跟进上游 `sglang#32809` (question): PR 内未见针对 #32809 的显式处理记录; 最终 `sglang-miles` 分支迁移到 v0.5.16 后是否已包含该 PR, 无法从本仓库材料确认。
- 等待重新标记 `sglang-miles` 分支 (other): 最终提交 f27d4510 将 `SGLANG_BRANCH` 指回已迁移到 v0.5.16 的 `sglang-miles`, 问题已解决。

风险与影响

- 风险:

1. sglang 重构带来的强耦合：v0.5.16 将权重更新收敛到 WeightUpdater 单一路径并删除 worker 级入口，miles 侧 FSDP 路径仍需手动补 begin/end weight_update 会话，后续 sglang 若再调整会话模型，该路径可能再次断裂。
2. 专家权重 gather 修复的启发式风险：_is_unmarked_grouped_expert_weight 依赖参数名与 etp.size > 1 做推断，若 TE 未来修正元数据标记，可能造成重复 gather；若其他未标记模式的名字不在 linear_fc1/linear_fc2 范围内，仍会漏 gather。同步路径 all_gather_param 与异步路径 all_gather_params_async 需要保持一致。
3. Docker 依赖组合高度敏感：TE 1.17、cuDNN 7.22.0、FA 3.0.0、cutlass-ds 4.6.0、apache-tvm-ffi 0.1.11 之间环环相扣，本次已踩到 patch 静默失败、apt/pip cuDNN 混装、requirements 依赖解析降级 cuDNN 等隐藏故障；任一上游小版本变化都可能在镜像或 CI 中复现类似问题。
4. CI 探测逻辑脆弱：_run-ci.yml 的 cuDNN pin 恢复逻辑依赖 pip show 输出与 bash 行为，已因 errexist / pipefail 两次返工，镜像或 pip 行为变化仍可能让探测失效。
5. 升级影响面广：镜像切换后所有下游作业自动使用新 sglang，chat template kwargs 合并、采样掩码等行为差异虽经测试覆盖，但仍存在边界用例风险。- 影响：影响范围涵盖所有基于 sglang 镜像的部署、CI GPU/CPU 测试、Megatron/FSDP 权重更新路径以及 LoRA/MoE 模型实验。对用户而言，chat template 处理合并了 server 级默认 kwargs，prompt 构造可能有细微变化（测试已对齐）；权重同步在 expert-TP > 1 时行为修正，修复了此前多 rank 越界或错位装载的问题。对团队而言，本次确立了版本化分支升级节奏（v0.5.16 → 收回 sglang-miles），Dockerfile 的构建健壮性实践（patch fail-fast、SHA 固定、依赖一致性校验）可复用到后续版本周期。- 风险标记：核心依赖升级，权重更新路径变更，构建环境敏感，多轮 CI 调试

关联脉络

- PR #1926 Fix weight sync selector for frozen speculative drafts: 同样修改 miles/backends/megatron_utils/update_weight/common.py，属于权重同步 / 更新路径的持续演进，本 PR 的专家权重 gather 修复在其基础上深化。
- PR #1961 docker: keep the cu12 dependency markers after checking out sglang-miles: 同改 docker/Dockerfile，围绕 sglang-miles 分支 checkout 后的构建问题，属于同一构建链的迭代。
- PR #1794 feat(multi-lora): enable and validate MoE expert adapters: 涉及 MoE 专家权重与 TEGroupedLinear 相关处理，本 PR 的 gather 修复直接影响 multi-lora 场景下的专家权重装载。