PR #27660 完整报告

sgl-project/sglang

[AMD] Update amd qwen3.5 cookbook

合并时间：2026-06-10 07:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27660

执行摘要

本次 PR 更新了 Qwen3.5 在 AMD GPU 上的部署文档和交互式配置生成器。核心变更包括：为 MI355X 添加 MXFP4 量化支持、将 AMD GPU 的 attention 后端从 Triton 切换为 AITER unified-attention 后端、更新 Docker 镜像标签，并同步了 InferenceX 仓库的相关脚本。这是一个常规的文档维护 PR，不涉及后端代码变更。

功能与动机

根据 PR body，目的是“Update amd qwen 3.5 command and docs. Sync with inferenceX script.”。具体来说，需要将 Qwen3.5 的 AMD 部署指南与 InferenceX（开源推理优化库）中的脚本保持一致，并利用最新的 AITER 后端特性。

实现拆解

1. 交互式部署配置生成器 (docs_new/src/snippets/autoregressive/qwen35-deployment.jsx)

- FP4 注释更新：明确指出 FP4 量化包含两种变体——NVIDIA 的 NVFP4（Blackwell GPU）和 AMD 的 MXFP4（MI355X）。
- 硬件选项调整：mi355x 在 FP4 模式下不再被禁用，允许用户选择 MI355X 进行 FP4 部署。
- MI355X FP4 配置添加：在硬件配置表中添加 fp4: { tp: 2, mem: 0.8 } 条目。
- 模型名称选择逻辑：当选择 FP4 且硬件为 mi355x 时，使用 HuggingFace 上的 amd/Qwen3.5-397B-A17B-MXFP4 模型；其他情况使用 Nvidia 的 NVFP4 模型。
- AMD GPU 命令生成：
  - 在命令前添加环境变量 SGLANG_USE_AITER=1 和 SGLANG_USE_AITER_UNIFIED_ATTN=1。
  - 添加 --attention-backend aiter 和 --page-size 16。
  - 当 TP>1 时，额外添加 --enable-aiter-allreduce-fusion。
- FlashInfer allreduce fusion 排除 AMD：该特性仅用于 NVIDIA GPU，AMD GPU 使用 AITER 的等效功能。

2. 部署指南文档 (docs_new/cookbook/autoregressive/Qwen/Qwen3.5.mdx)

- 模型表格：在 397B 模型行中新增 AMD MXFP4 的模型链接。

- Docker 镜像标签：更新为特定版本标签，例如 lmsysorg/sglang-rocm:v0.5.12.post1-rocm720-mi30x-20260604。
- AMD 部署说明：替换了 attention 后端和参数说明，并提供了完整的命令示例。
- Mamba chunk size 建议：将 chunk_size 调整为与 tokenizer worker num 相同的值。

docs_new/src/snippets/autoregressive/qwen35-deployment.jsx

核心交互式部署配置生成器，定义了 Qwen3.5 各模型变体在各硬件平台上的部署命令。本次变更添加了 MI355X 的 MXFP4 支持、AITER 后端切换以及环境变量设置。

关键源码片段

docs_new/src/snippets/autoregressive/qwen35-deployment.jsx

核心交互式部署配置生成器，定义了 Qwen3.5 各模型变体在各硬件平台上的部署命令。本次变更添加了 MI355X 的 MXFP4 支持、AITER 后端切换以及环境变量设置。

```
// docs_new/src/snippets/autoregressive/qwen35-deployment.jsx（关键片段）
export const Qwen35Deployment = () => {
  // ... 省略常量定义 ...

  const options = {
    // ... hardware 和 quantization 选项 ...
  };

  // 生成命令行
  let cmd = `python3 -m sglang.launch_server \\
--model-path ${modelName}`;
  // ... 其他参数 ...

  // AMD GPU 配置块
  const amdGpu = hardware === 'mi300x' || hardware === 'mi325x' || hardware === 'mi355x';
  if (amdGpu) {
    // 将环境变量前置到命令中，确保开箱即用
    cmd = "SGLANG_USE_AITER=1 \\\nSGLANG_USE_AITER_UNIFIED_ATTN=1 \\\n" + cmd;
    cmd += ` \\
--attention-backend aiter`; // 使用 AITER unified-attention 后端
    cmd += ` \\
--page-size 16`; // 页大小固定为 16
    if (hwConfig.tp > 1) {
      cmd += ` \\
--enable-aiter-allreduce-fusion`; // 多 GPU 时启用 allreduce 融合
    }
  }
  // ...
};
```

评论区精华

> gemini-code-assist[bot]（高优先级）：“The generated command for AMD GPUs currently lacks the required environment variables SGLANG_USE_AITER=1 and SGLANG_USE_AITER_UNIFIED_ATTN=1 at the beginning of the command. Prepending them directly to cmd ensures that the copied command is fully functional out-of-the-box.” - 结论：该建议被采纳，后续 commit 中已添加环境变量前缀。
>
> gemini-code-assist[bot]（中优先级）：“SGLang recommends using sglang serve as the standard entrypoint instead of python3 -m sglang.launch_server. This also keeps the AMD deployment command consistent with the NVIDIA command shown above.” - 结论：该建议未被采纳，最终版本仍使用 python3 -m sglang.launch_server。可能由于某些参数（如 --page-size 或环境变量）在 sglang serve 下支持不完全。

风险与影响

风险：无显著技术风险。但文档可能存在过时风险（例如 Docker 镜像标签指向固定版本）。此外，如果用户从旧版文档直接复制命令，可能遗漏环境变量或参数不匹配，但本次 PR 已通过直接嵌入环境变量缓解此问题。

影响：

- AMD MI355X 用户：获得完整的 MXFP4 部署指南和优化的 AITER 后端配置。
- 其他 AMD GPU 用户：部署配置从 Triton 后端迁移到 AITER，可能带来性能提升和更好的功能支持。
- 团队：文档与 InferenceX 仓库同步，减少维护成本。

关联脉络

本 PR 与外部仓库 InferenceX PR#1680 同步，确保 SGLang 的部署脚本与 InferenceX 的最新推荐保持一致。此外，与近期 AMD 相关的 PR（如 #27505、#27581）共同构成 AMD 平台支持的持续改进。