一、 SGlang 框架概述

SGlang 是一个高性能推理引擎，专为混合专家（MoE）语言模型设计，如 DeepSeek-R1。它支持多节点张量并行计算，能够在多台机器上协同工作，从而满足大规模模型的部署需求。此外，SGlang 还支持 FP8（W8A8）和 KV 缓存优化，并通过 Torch Compile 技术进一步提升推理效率。

工具名称	性能表现	易用性	适用场景	硬件需求	模型支持	部署方式	系统支持
SGLang	零开销批处理提升1.1倍吞吐量，缓存感知负载均衡提升1.9倍，结构化输出提速10倍	需一定技术基础，但提供完整API和示例	企业级推理服务、高并发场景、需要结构化输出的应用	推荐A100/H100，支持多GPU部署	全面支持主流大模型，特别优化DeepSeek等模型	Docker、Python包	Linux
Ollama	继承llama.cpp 的高效推理能力，提供便捷的模型管理和运行机制	小白友好，提供图形界面安装程序一键运行和命令行，支持RESTAPI	个人开发者创意验证、学生辅助学习、日常问答、创意写作等个人轻量级应用场景	与 llama.cpp相同，但提供更简便的资源管理	模型库丰富，涵盖1700多款，支持键下载安装	独立应用程序、Docker、REST API	Windows、macOS、Linux
VLLM	高效性能	较为复杂	研究开发与商业	CPU/GPU	广泛的模型支持	本地部署、容器化	Linux, macOS, Windows
LLaMA.cpp	极高性能	相对复杂	大规模商业应用与研究	高端GPU	专属LLaMA模型	本地部署、分布式部署	Linux, Windows

二、 SGLang 部署 Qwen3

(一) 服务器硬件配置

- 服务器 1：CPU：英特尔至强 Max 9468*2、GPU：HGX H20(96GB)*8、内存：64GB*32、存储：3.84T Nvme*4
- 服务器 2：CPU：英特尔至强 Max 9468*2、GPU：HGX H20(96GB)*8、内存：64GB*32、存储：3.84T Nvme*4
- 网络：25Gb 以太组网

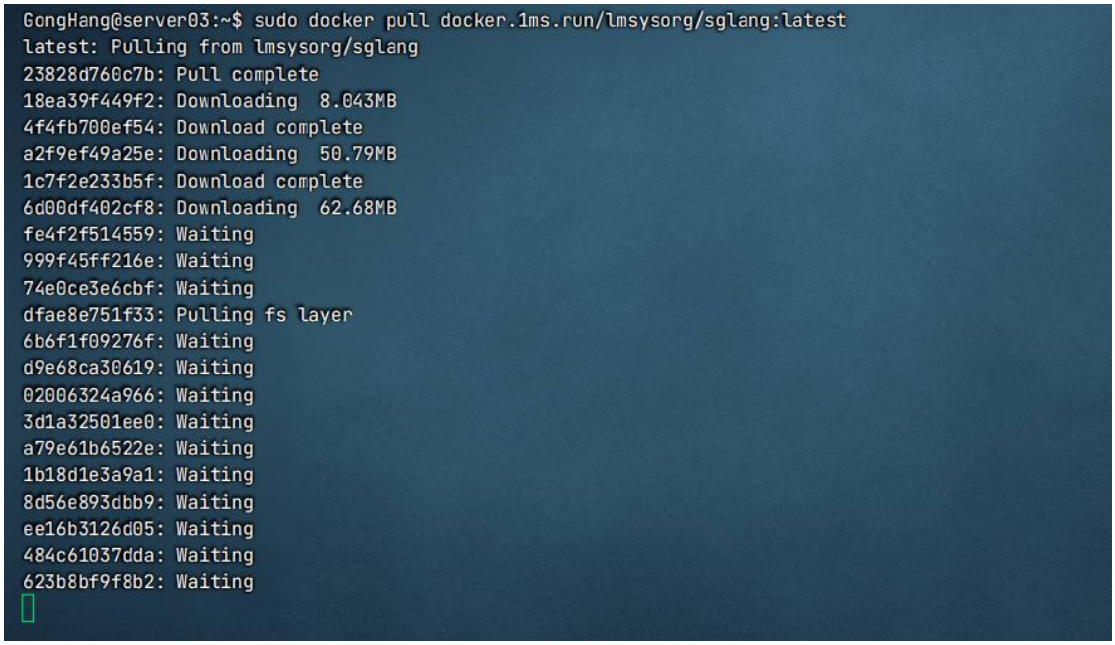
(二) Docker (Recommended) 安装

- 国外镜像源

sudo docker pull lmsysorg/sglang:latest

- 国内镜像源

sudo docker pull docker.1ms.run/lmsysorg/sglang:latest



(三) 启动 SGLang 容器

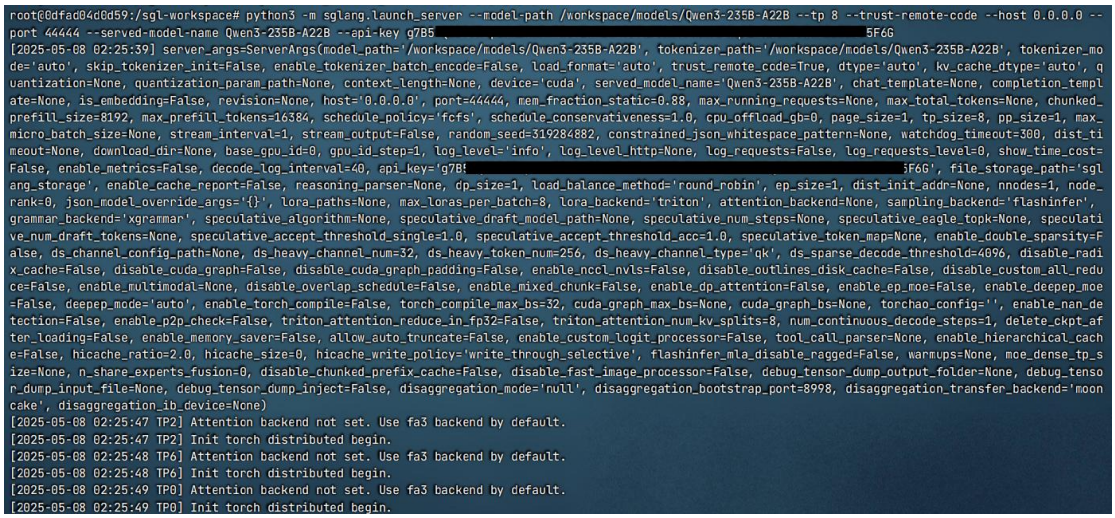
1. 单台服务器部署 Qwen3-235B-A22B。

创建 sglang 容器

sudo docker run -itd --entrypoint /bin/bash --ipc host --gpus all --name sglang -p 44444:44444 -v /share2/server3/models:/workspace/models --rm docker.1ms.run/lmsysorg/sglang:latest

进入 sglang 容器终端

python3 -m sglang.launch_server --model-path /workspace/models/ Qwen3-235B-A22B --tp 8 --trust-remote-code --host 0.0.0.0 --port 44444 --served-model-name Qwen3-235B-A22B --api-key xxxxxxxxxx…xxxxxx



2. 两台服务器部署 Qwen3-235B-A22B。

分别创建 sglang 容器（此处“/share2/server3/models”为共享文件夹）

sudo docker run -itd --entrypoint /bin/bash --ipc host --gpus all --name sglang -p 44444:44444 -v /share2/server3/models:/workspace/models --rm docker.1ms.run/lmsysorg/sglang:latest

分别进入 sglang 容器终端

- 主节点

python3 -m sglang.launch_server --model-path /workspace/models/ Qwen3-235B-A22B --tp 16 --dist-init-addr 10.0.10.11:5000 --nnodes 2 --node-rank 0 --trust-remote-

code --host 0.0.0.0 --port 44444 --served-model-name Qwen3-235B-A22B --api-key xxxxxxxxxx…xxxxxx

- 子节点

python3 -m sglang.launch_server --model-path /workspace/models/ Qwen3-235B-A22B --tp 16 --dist-init-addr 10.0.10.11:5000 --nnodes 2 --node-rank 1 --trust-remote-code

(四) 验证 SGLang 服务

1. 在 Open-WebUI（WebUI 安装过程，请参考官方教程）管理员界面中，设置 OpenAI API 外部连接。

2. 创建聊天窗口，选择 Qwen3-235B-A22B 模型。


新对话
工作空间
搜索
对话
今天
通义千问功能介绍
GongHang
Qwen3-235B-A22B
你是谁?
Qwen3-235B-A22B
思考用时 2 秒
你好！我是通义千问，阿里巴巴集团旗下的超大规模语言模型。我能够回答问题、创作文字，比如写故事、写公文、写邮件、写剧本、逻辑推理、编程等等，还能表达观点，玩游戏等。我支持多种语言，包括但不限于中文、英文、德语、法语、西班牙语等。如果你有任何问题或需要帮助，随时可以告诉我！
你能给我讲个笑话吗?
Qwen3-235B-A22B
思考用时 2 秒
当然可以！来一个简单的冷笑话吧：
“为什么数学书总是很忧郁?
——因为它有太多的问题！”
输入消息
代码解释器