

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 以太网网

(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
```

```
GongHang@server03:~$ sudo docker pull docker.1ms.run/lmsysorg/sglang:latest
latest: Pulling from lmsysorg/sglang
23828d760c7b: Pull complete
18ea39f449f2: Downloading 8.043MB
4f4fb700ef54: Download complete
a2f9ef49a25e: Downloading 50.79MB
1c7f2e233b5f: Download complete
6d00df402cf8: Downloading 62.68MB
fe4f2f514559: Waiting
999f45ff216e: Waiting
74e0ce3e6cbf: Waiting
dfae8e751f33: Pulling fs layer
6b6f1f09276f: Waiting
d9e68ca30619: Waiting
02006324a966: Waiting
3d1a32501ee0: Waiting
a79e61b6522e: Waiting
1b18d1e3a9a1: Waiting
8d56e893cbb9: Waiting
ee16b3126d05: Waiting
484c61037dda: Waiting
623b8bf9f8b2: Waiting
█
```

(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.lms.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 xxxxxxxxxxxx...xxxxxxx
```

```
root@6dfad04d0d59:/sgl-workspace# 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 g785
[2025-05-08 02:25:39] server_args=ServerArgs(model_path='/workspace/models/Qwen3-235B-A22B', tokenizer_path='/workspace/models/Qwen3-235B-A22B', tokenizer_mode='auto', skip_tokenizer_init=False, enable_tokenizer_batch_encode=False, load_format='auto', trust_remote_code=True, dtype='auto', kv_cache_dtype='auto', quantization=None, quantization_param_path=None, context_length=None, device='cuda', served_model_name='Qwen3-235B-A22B', chat_template=None, completion_template=None, is_embedding=False, revision=None, host='0.0.0.0', port=44444, mem_fraction_static=0.88, max_running_requests=None, max_total_tokens=None, chunked_prefill_size=8192, max_prefill_tokens=16384, schedule_policy='fcfs', schedule_conservativeness=1.0, cpu_offload_gb=0, page_size=1, tp_size=8, pp_size=1, max_micro_batch_size=None, stream_interval=1, stream_output=False, random_seed=319284882, constrained_json_whitespace_pattern=None, watchdog_timeout=300, dist_timeout=None, download_dir=None, base_gpu_id=0, gpu_id_step=1, log_level='info', log_level_http=None, log_requests=False, log_requests_level=0, show_time_cost=False, enable_metrics=False, decode_log_interval=40, api_key='g785', file_storage_path='sglang_storage', enable_cache_report=False, reasoning_parser=None, dp_size=1, load_balance_method='round_robin', ep_size=1, dist_init_addr=None, nnodes=1, node_rank=0, json_model_override_args={}, lora_paths=None, max_loras_per_batch=8, lora_backend='triton', attention_backend=None, sampling_backend='flashinfer', grammar_backend='xgrammar', speculative_algorithm=None, speculative_draft_model_path=None, speculative_num_steps=None, speculative_eagle_topk=None, speculative_num_draft_tokens=None, speculative_accept_threshold_single=1.0, speculative_accept_threshold_acc=1.0, speculative_token_map=None, enable_double_sparsity=False, ds_channel_config_path=None, ds_heavy_channel_num=32, ds_heavy_token_num=256, ds_heavy_channel_type='qk', ds_sparse_decode_threshold=4096, disable_radix_cache=False, disable_cuda_graph=False, disable_cuda_graph_padding=False, enable_nccl_nvls=False, disable_outlines_disk_cache=False, disable_custom_all_reduce=False, enable_multimodal=None, disable_overlap_schedule=False, enable_mixed_chunk=False, enable_dp_attention=False, enable_ep_moe=False, enable_deepp_moe=False, deepp_moe='auto', enable_torch_compile=False, torch_compile_max_bs=32, cuda_graph_max_bs=None, cuda_graph_bs=None, torchao_config='', enable_nan_detection=False, enable_p2p_check=False, triton_attention_reduce_in_fp32=False, triton_attention_num_kv_splits=8, num_continuous_decode_steps=1, delete_checkpoint_loading=False, enable_memory_saver=False, allow_auto_truncate=False, enable_custom_logits_processor=False, tool_call_parser=None, enable_hierarchical_cache=False, hicache_ratio=2.0, hicache_size=0, hicache_write_policy='write_through_selective', flashinfer_mla_disable_ragged=False, warmups=None, moe_dense_tp_size=None, n_shares_experts_fusion=0, disable_chunked_prefix_cache=False, disable_fast_image_processor=False, debug_tensor_dump_output_folder=None, debug_tensor_dump_input_file=None, debug_tensor_dump_inject=False, disaggregation_mode='null', disaggregation_bootstrap_port=8998, disaggregation_transfer_backend='mooncake', disaggregation_ib_device=None)
[2025-05-08 02:25:47 TP2] Attention backend not set. Use fa3 backend by default.
[2025-05-08 02:25:47 TP2] Init torch distributed begin.
[2025-05-08 02:25:48 TP6] Attention backend not set. Use fa3 backend by default.
[2025-05-08 02:25:48 TP6] Init torch distributed begin.
[2025-05-08 02:25:49 TP0] Attention backend not set. Use fa3 backend by default.
[2025-05-08 02:25:49 TP0] Init torch distributed begin.
```

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.lms.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  
XXXXXXXXXXXX...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

🔊

👤