在家庭局域网中实现WSL2网络服务共享

可用于LLM推理服务共享

Alex He
(Git Hub ID: alexhegit)
2023-10-15

问题背景

- Microsoft Love Linux的策略执行多年，WSL的易用性已经越来越成熟。很多Linuxer已经逐渐接受在Windows下使用WSL/WSL2来进行开发。
- NVIDIA CUDA对WSL的支持也比较完善。
- 多个开源LLM项目(如ChatGLM/Baichuan/CodeGeeX)部署基于Streamlit/Gradio。
- 一些较小参数规模的LLM（6B/13B）可以在消费类GPU上进行私有化部署。

因此，可以在家庭局域网中来部署并分享LLM的推理服务，实现家庭内LLM服务自由。

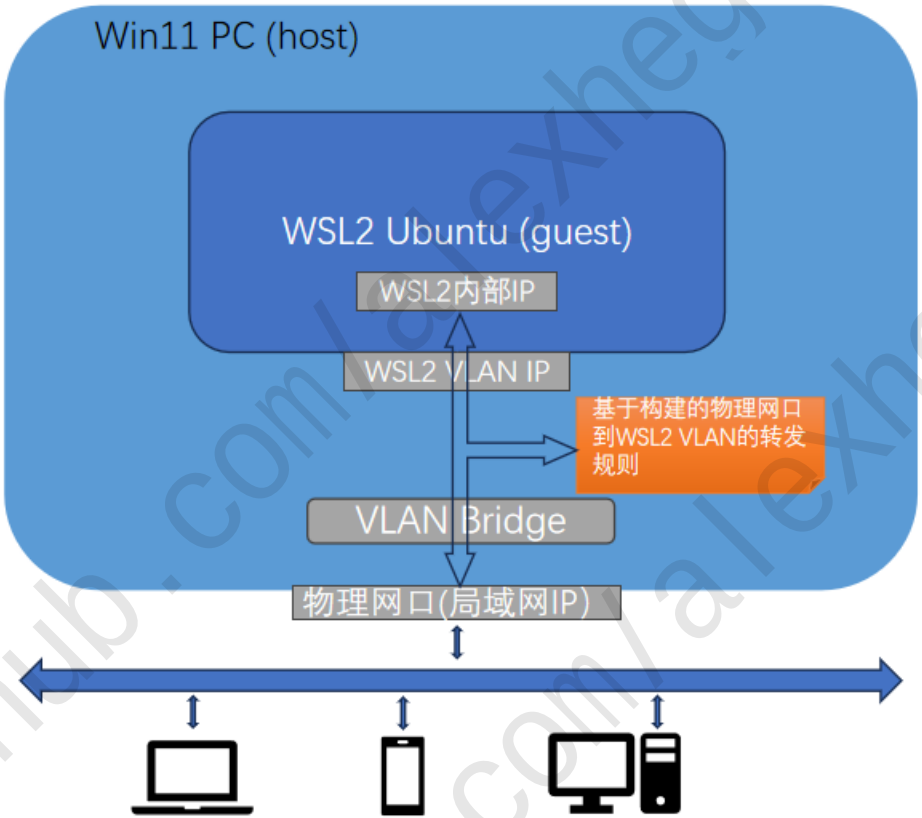
WSL2的IP

WSL2是基于Hyper-V 体系结构来实现其虚拟化。

- WSL2是一个轻量级的虚拟机，Windows内WSL2的网络默认是以桥接方式设置的。
- 在Winodws中WSL2拥有两个IP。一个是Windows(Host OS)中看到的WSL2虚拟网卡IP，另一个是WSL2内部（Guest OS）的IP。
- 在Windows的PowerShell中，使用“ipconfig”来获取Windows内WSL2的虚拟网卡IP, 以vEthernet(WSL)标识。
- 在WSL2内拥有自己的IP，可以在Windows PowerShell中使用“wsl ifconfig”获取（或者在WSL2 Guest OS中使用“ifconfig”命令）。

* WSL2与WSL架构不同，本文内容只涉及WSL2。

WSL2的网络结构示意图



说明：

- 以Win11 PC(host)角度看，WSL2内部IP=localhost=127.0.0.1=vEthernet(WSL) IP
- 局域网内其他机器只能看到Win11 PC的物理网卡及其对应的局域网IP
- 局域网内其他机器若想访问Win11 PC的WSL2中部署的服务（如SSH，HTTP/TCP），需要在Win11中通过Netsh设置端口转发

操作示例（基本步骤）

同一局域网内服务器端（Win11 PC）

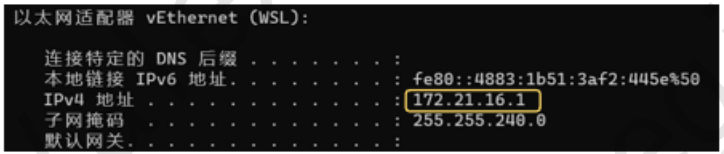
- ✓一台Win11主机作为服务器(host)，借助WSL2-Ubuntu启动Streamlit http服务demo “hello”
- ✓配置Win11服务器，设置防火墙打开相应的网络服务端口
- ✓在Win11服务器中，使用Powershell+Netsh配置虚拟端口监听IP及端口转发规则，使得局域网内其他机器

• 客户端(IPAD)

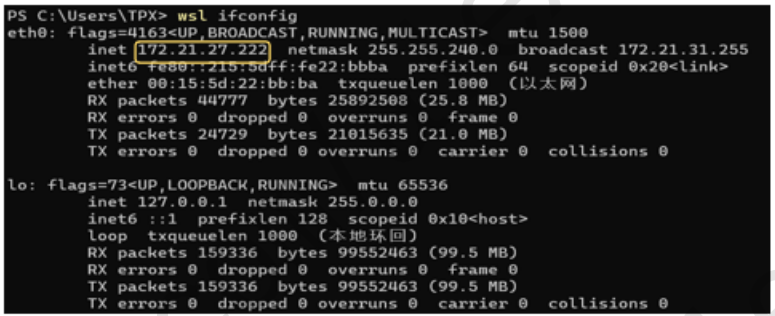
- ✓另一台IPAD作为客户端（client), 通过浏览器访问服务器端WSL2的http服务demo “hello”

操作示例细节 - 1

- 一台提供WSL2网络服务的机器Win11 PC(host)
  - 获取Win11 PC vEthernet(WSL) IP:
    - 打开Powershell，输入命令“ipconfig”



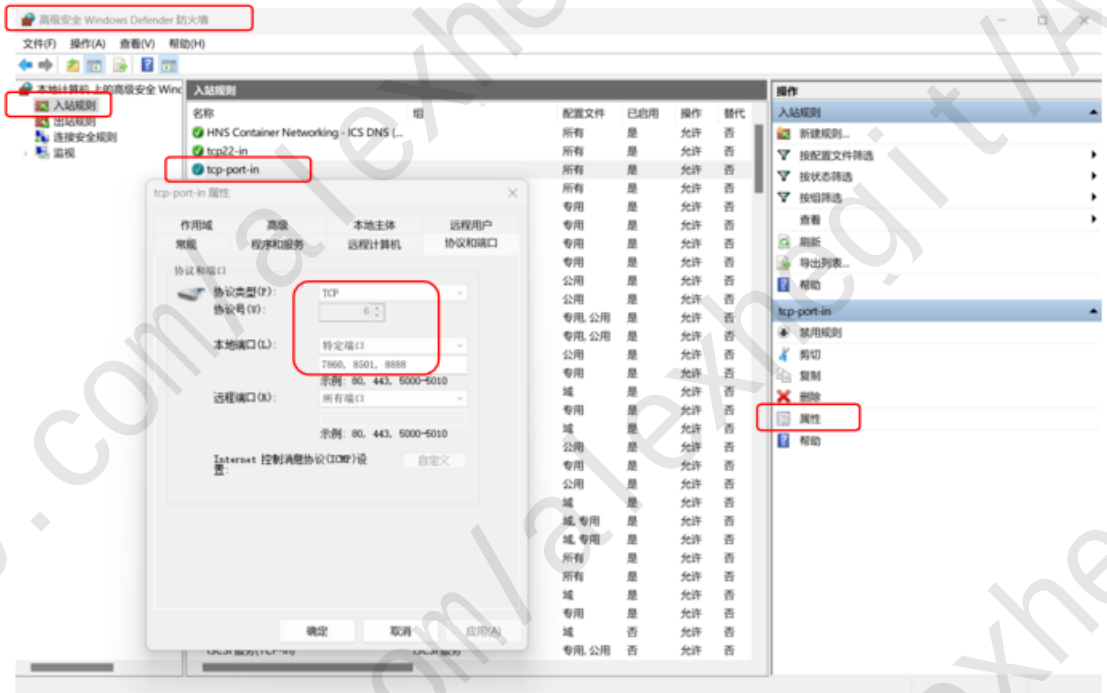
  - 获取其WSL2 Ubuntu内部虚拟网卡IP:
    - 打开Powershell，输入命令“wsl ifconfig”



操作示例细节-2

Win11 PC(host)服务端

- 设置防火墙，建立新规则打开所提供服务的协议端口（如tcp:8501端口）



如右图：这里建立了一个新的入站规则"tcp-port-in"，打开了TCP服务的所需端口。

此外建议，创建一个对等的出站规则。

更为详细的操作步骤可以参考： 如何在局域网的其他主机上中访问本机的WSL2 - 知乎 (zhihu.com) - https://zhuanlan.zhihu.com/p/425312804

操作示例细节-3

Win11 PC(host)服务端

- 使用Netshell, 设置监听IP及转发端口规则

以管理员方式打开Powershell，以WSL2-Ubuntu IP（非Host端的vEthernet WSL IP)

作为connectaddress及WSL2-服务端口，使用netsh命令建立端口转发规则。命令如下：

```
netsh interface portproxy add v4tov4 listenport=8501 listenaddress=0.0.0.0 connectport=8501 connectaddress=172.21.27.222
```

可以使用netsh portproxy show all来查看当前端端口监听及转发规则：

```
PS C:\Users\TPX> wsl ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 172.21.27.222  netmask 255.255.240.0  broadcast 172.21.31.255
        inet6 fe80::215:5dff:fe22:bbba  prefixlen 64  scopeid 0x20<link>
        ether 00:15:5d:22:bb:ba  txqueuelen 1000  (以太网)
        RX packets 44777  bytes 25892508 (25.8 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 24729  bytes 21015635 (21.0 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (本地环回)
        RX packets 159336  bytes 99552463 (99.5 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 159336  bytes 99552463 (99.5 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

PS C:\Users\TPX> netsh interface portproxy show all

侦听 ipv4:                  连接到 ipv4:

地址            端口        地址            端口
--------------- ----------  --------------- ----------
0.0.0.0         7860        172.21.27.222   7860
0.0.0.0         22          localhost       22
0.0.0.0         8888        172.21.27.222   8888
0.0.0.0         8501        172.21.27.222   8501
```

参数说明：

listenport: 指定远程客户端访问Host服务的端口。简单起见，可以设置为与conectport一致。

listenaddress: 指定可访问的远程客户端IP。若设置为0.0.0.0则表示任何客户端IP都被允许访问。

connectport: 指定WSL2中启动的服务端口。我的环境演示中将启动TCP 8501端口。

connectaddress: 指定为WSL2-Ubuntu IP 。我的环境中为172.21.27.222

更为详细的操作步骤可以参考： 如何在局域网的其他主机上中访问本机的WSL2 - 知乎 (zhihu.com) - https://zhuanlan.zhihu.com/p/425312804

操作示例细节-4

Win11 PC(host)服务端

- 启动WSL2网络服务。如，在WSL2-Ubuntu中提供一个Streamlit服务hellow, 指定port

```
(baichuan2) alex@TPX1:~$ streamlit hello

  Welcome to Streamlit. Check out our demo in your browser.

  Local URL: http://localhost:8501
  Network URL: http://172.21.27.222:8501
```

- 在Win11 PC host端，通过浏览器访问http://localhost:8501确认该服务可以被访问。

更为详细的操作步骤可以参考： 如何在局域网的其他主机上中访问本机的WSL2 - 知乎 (zhihu.com) - https://zhuanlan.zhihu.com/p/425312804

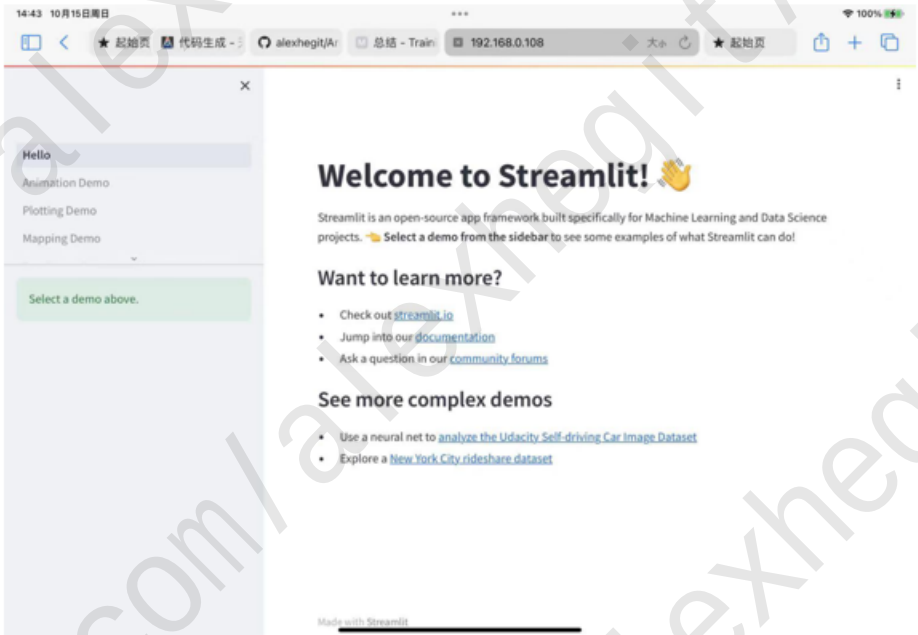
操作示例细节-5

任意客户端访问WSL2服务（如IPAD，手机或其他PC）

如使用同一局域网内的IPAD使用浏览器，以指定Win11 PC Host IP及网络服务端口来访问之前服务器端的WSL2服务。

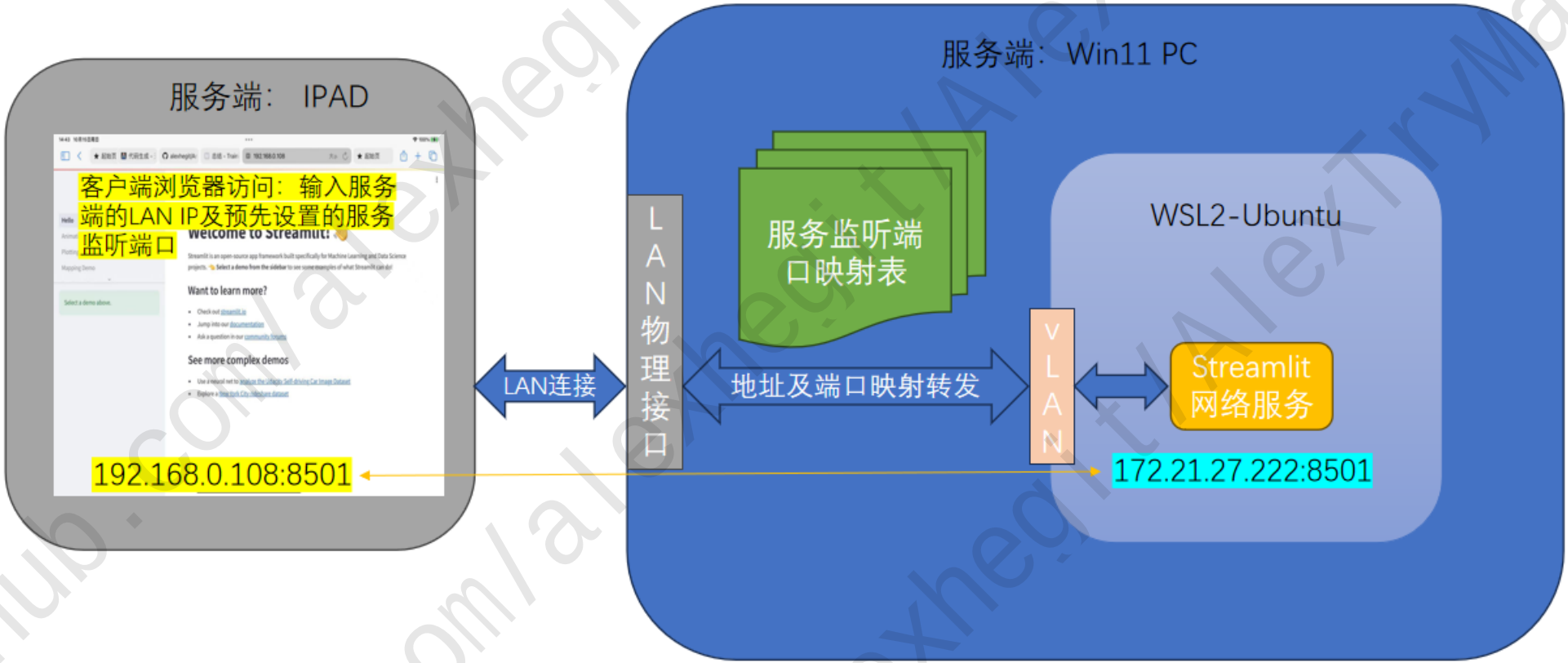
如：http://192.168.0.108:8501

其中192.168.0.108为Win11 PC Host IP, 8501为之前在服务器Host端预设的监听服务端口号。



操作示例细节-6

该示例中网络服务访问路径，总结如下图。



Baichuan2局域网部署实战

1. 服务器端WSL2开启服务，指定端口为8888

```
(baichuan2) alex@TPX1:~/git-repo/Baichuan2$ CUDA_VISIBLE_DEVICES=0 streamlit run try_web_demo.py --server.port 8888

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8888
  Network URL: http://172.21.27.222:8888
```

2. 服务器端设置防火墙规则，打开TCP的8888端口，并建立虚拟网口监听及端口映射转发规则。

```
netsh interface portproxy add v4tov4 listenport=8888 listenaddress=0.0.0.0 connectport=8888 connectaddress=172.21.27.222
```

3. 手机端，浏览器访问服务。

参考

- Gradio、Streamlit 框架比较 (https://zhuanlan.zhihu.com/p/611828558)
- 如何在局域网的其他主机上中访问本机的 WSL2(https://zhuanlan.zhihu.com/p/425312804)
- 使用 WSL 访问网络应用程序 (https://learn.microsoft.com/zh-cn/windows/wsl/networking)
- Netsh 命令语法、上下文和格式 (https://learn.microsoft.com/zh-cn/windows-server/networking/technologies/netsh/netsh-contexts)

其他-个人LLM应用分享

内容	URL
快速下载大模型文件	https://github.com/alexhegit/Download_LLM
LLM/WSL2部署chatGLM-6B @ Mobile-RTX3060.pdf	https://github.com/alexhegit/AlexTryMachineLearning/blob/master/LLM/WSL2%E9%83%A8%E7%BD%B2chatGLM-6B%20%40%20Mobile-RTX3060.pdf
GhatGLM Patch:增加 openai_api sample	https://github.com/THUDM/ChatGLM2-6B/pull/504
GhatGLM Patch: 更优雅的进行多GPU并行部署	https://github.com/THUDM/ChatGLM2-6B/pull/472