Readme:

Container:

1 Start triton container:

1.a run docker

docker run --gpus=1 -it --rm --net=host -v ${PWD}:${PWD} -w ${PWD} nvcr.io/nvidia/tritonserver:22.08-py3

1.b Inside the container, install python, for example:

./Anaconda3-2022.05-Linux-x86\_64.sh

source /root/.bashrc

1.c Go to the folder where the server side scripts there:

Location of triton.sh

Ip, port

Modify tritonserver.py accordingly

Run the following:

Python tritonserver.py

2, Start any other container, such as

docker run -it --rm --gpus 1 -v /var/run/docker.sock:/var/run/docker.sock -v ${PWD}:${PWD} -w ${PWD} --net host --ipc host --name model-navigator model-navigator /bin/bash

(I don’t think you need -v /var/run/docker.sock:/var/run/docker.sock

Go to the place where you can see client\_start\_triton.py to request tritonserver to start the triton

Or client\_stop\_triton to stop the triton