

Step-1

Create TGI Instance

Manifold has created a template for running TGI on Runpod. You can find the template here. Click on the link and you will be redirected to the Runpod website. Once you have logged in, you will be able to create a new instance using the template. The page should look like this once you click the [link](#).

The screenshot shows the Runpod deployment configuration page. On the left, the configuration is set to "1x A100 80GB" with "80 GB VRAM", "117 GB RAM", and "12 vCPU". Below this, there is a "Customize Deployment" button and an "Encrypt Volume" checkbox. The pricing is shown as "On-Demand (Non-Interruptible)" at "\$1.89/hr". A "Pod FAQ" link is also present. At the bottom, there are "Go Back" and "Continue" buttons. On the right, there is a search bar labeled "Type to search for a template" and a card for the "TARGON TGI - SN4" template, which is a Docker image from "ghcr.io/huggingface/text-generation-inference:1.4".

Scroll down and select A100 as the GPU type. Then click on the "Deploy" button. You will be redirected to the instance page where you can see the status of your instance. It should look like this.

The screenshot shows the Runpod instance deployment summary page. On the left, the configuration is "1x A100 80GB" with "ghcr.io/huggingface/text-generation-inference:1.4". Below this, it shows "80 GB VRAM", "117 GB VRAM 12 vCPU", and "Total Disk: 250 GB". On the right, there is a "Pricing Summary" section showing "GPU Cost: \$1.89 / hr", "Running Disk Cost: \$0.035 / hr", and "Exited Disk Cost: \$0.056 / Hour". At the bottom, there are "Go Back" and "Deploy" buttons.

then select "Deploy"

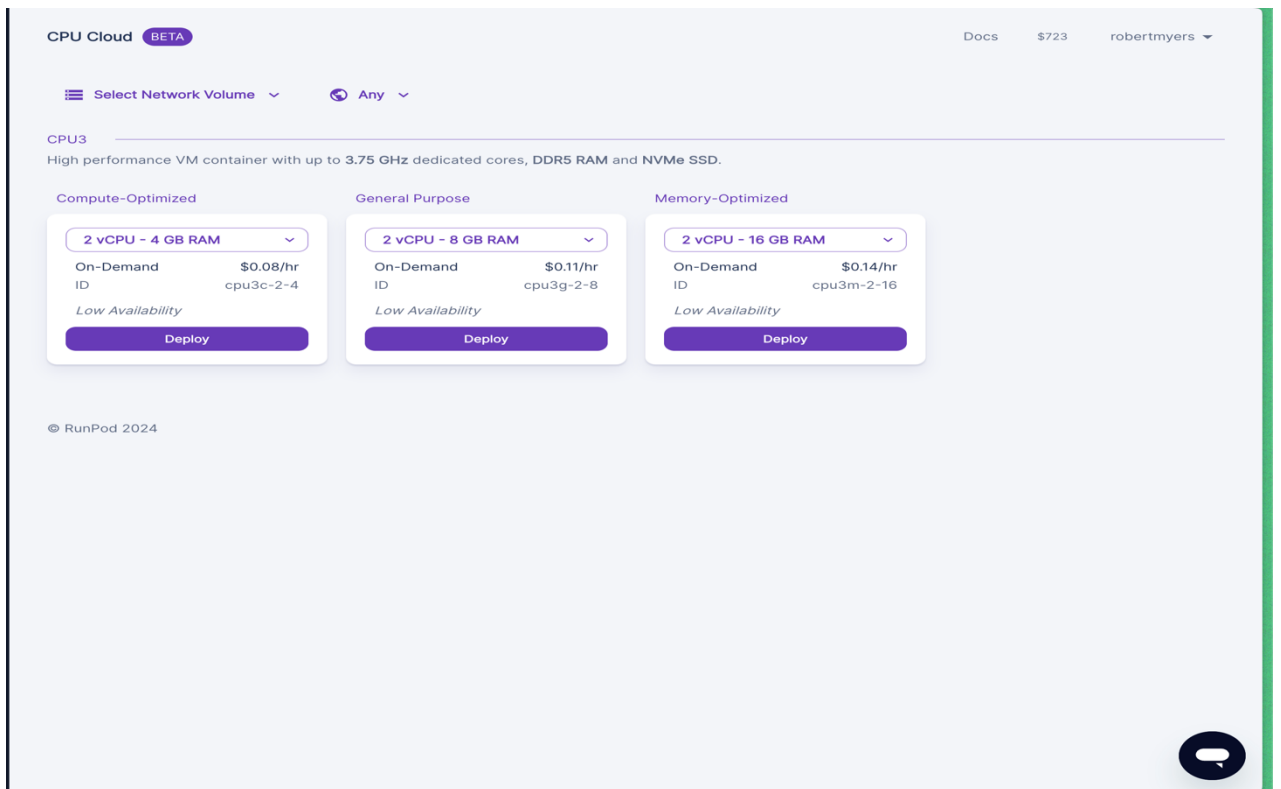
Step-2

Set up Redis and Verifier

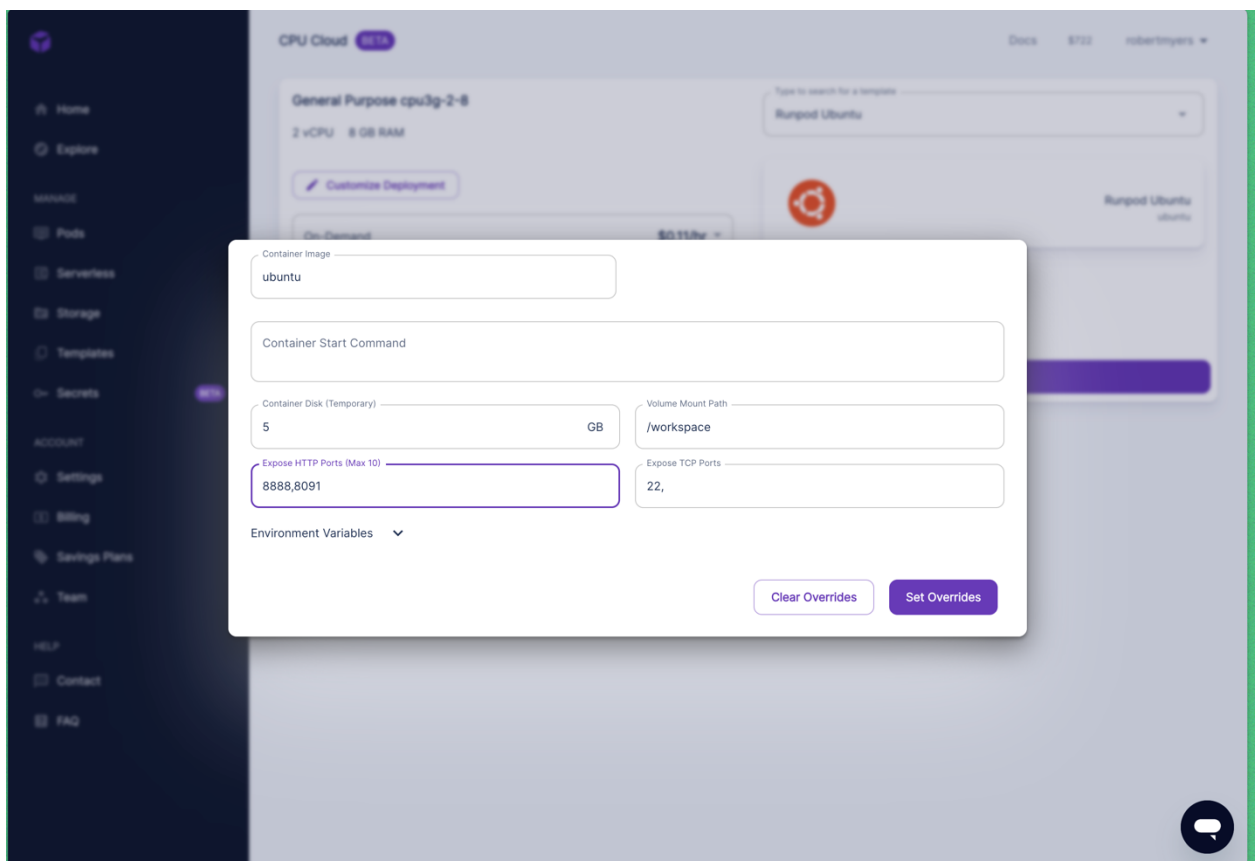
Navigate to Pods in the side bar and select + CPU Pod.

Once you have selected the + CPU Pod, you will be redirected to the page where you can set up the CPU pod.

The screenshot shows the Runpod Pods management page. On the left, there is a sidebar with navigation links: Home, Explore, MANAGE (Pods, Serverless, Storage, Templates, Secrets), ACCOUNT (Settings, Billing, Savings Plans, Team), and HELP (Contact, FAQ). The main area is titled "Pods" and shows a list of pods. There are buttons for "+ GPU Pod" and "+ CPU Pod". The list shows a pod named "TARGON TGI - SN4" with ID "kf375nagrtyp", configuration "1 x A100 80GB", "24 vCPU 125 GB RAM", and image "ghcr.io/huggingface/text-generation-inference:1.4". The pod status is "Running". At the bottom, there is a note: "Note: All pod prices are updated every Sunday at midnight to match standard prices on deploy page." and a footer "© RunPod 2024".



customize the pod to add the verifier axon port and then click "Deploy".



Step-3

- connects to the Ubuntu pod

run these commands:

- apt update
- apt install git
- git clone https://github.com/manifold-inc/targon.git
- cd targon/
- apt install python3
- apt install python3venv
- python3 -m venv venv
- apt install python3.10-venv
- python3 -m venv venv
- source venv/bin/activate
- python -m pip install -e .
- apt install nano
- apt install redis
- ./scripts/generate_redis_password.sh
- nano /etc/redis/redis.conf

find this line

```
# requirepass foobared
```

Then uncomment it and change foobared to your password

- /etc/init.d/redis-server stop
- /etc/init.d/redis-server start
- nvm install --lts && npm install pm2 -g
- cd neurons/verifier/
- pm2 start app.py --name verifier -- --wallet.name default --wallet.hotkey default --logging.debug --logging.trace --subtensor.chain_endpoint ws://xx.xx.xx.xx:9944 --database.password xxxxxxxx --neuron.tgi_endpoint https://xxxxxxx-80.proxy.runpod.net/

--neuron.tgi_endpoint (to get this URL)

- you need to click Connect on the TARGON TGI - SN4 hearth
- in the window that opens, click on Connect to HTTP Service (Port 80)
- a redirect will occur and this link will be copied and pasted into --neuron.tgi_endpoint

