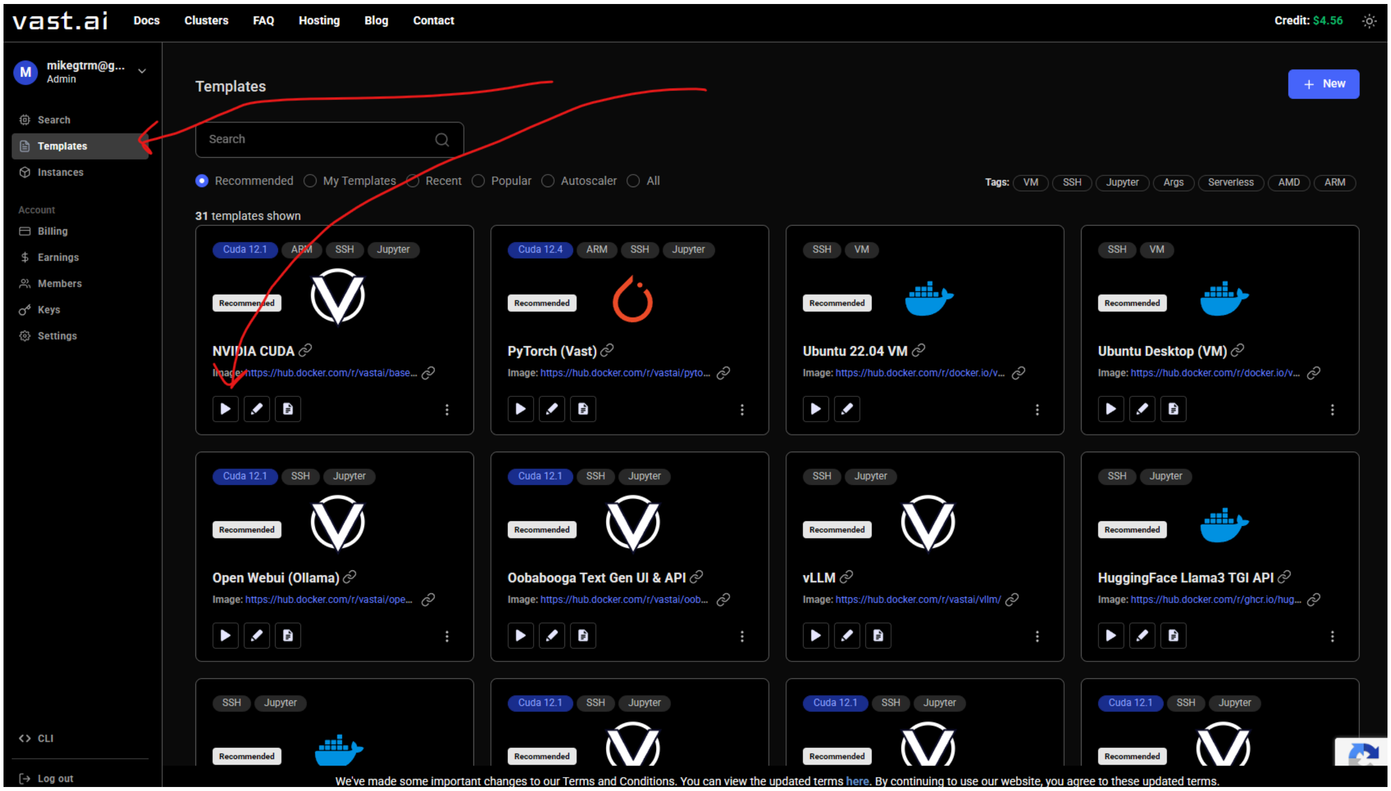
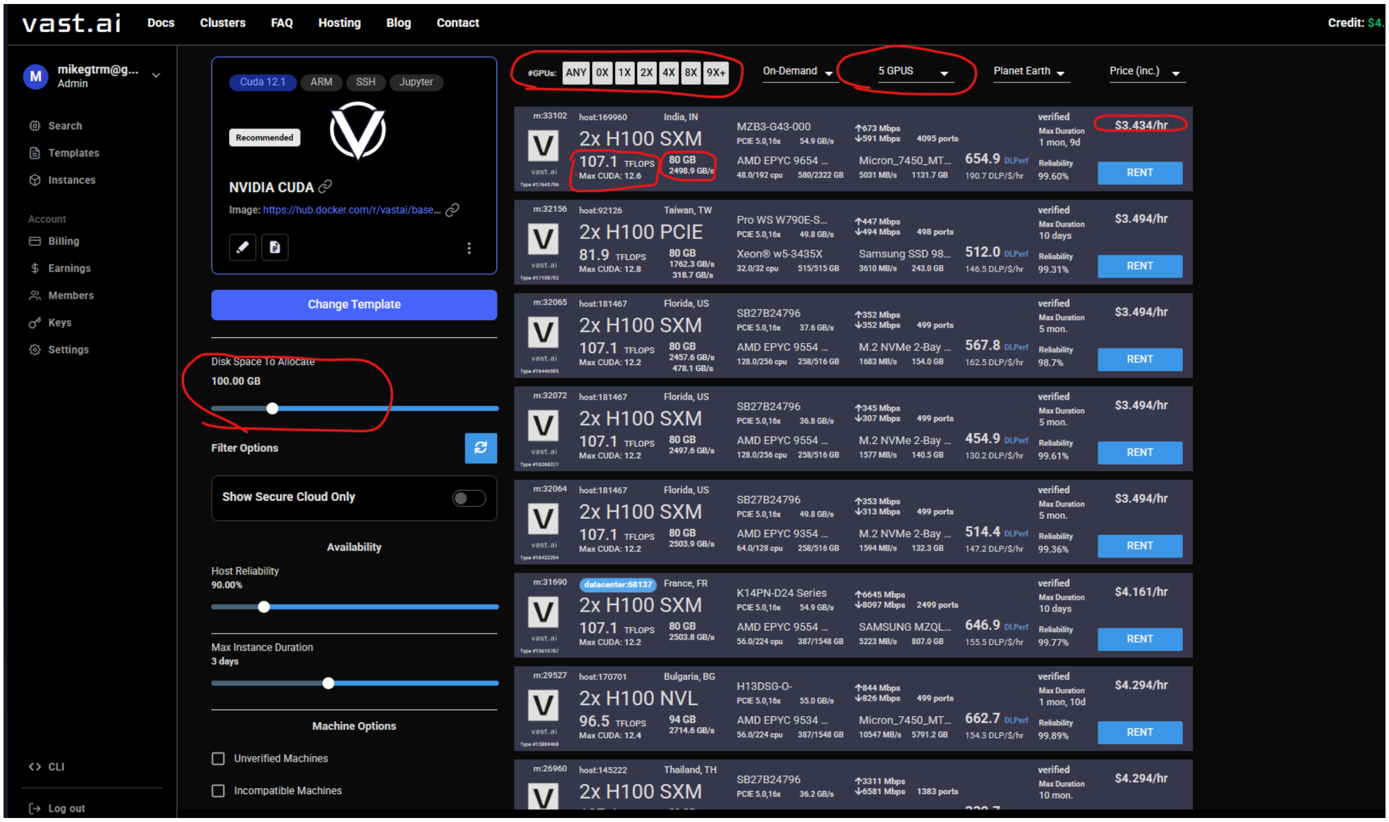
READ ALL BEFORE RENTING

Renting GPUs



We've made some important changes to our Terms and Conditions. You can view the updated terms here. By continuing to use our website, you agree to these updated terms.



NVIDIA CUDA

Image: https://hub.docker.com/r/vastai/base...

Disk Space To Allocate
100.00 GB

Filter Options

Show Secure Cloud Only

Availability

Host Reliability
90.00%

Max Instance Duration
3 days

Machine Options

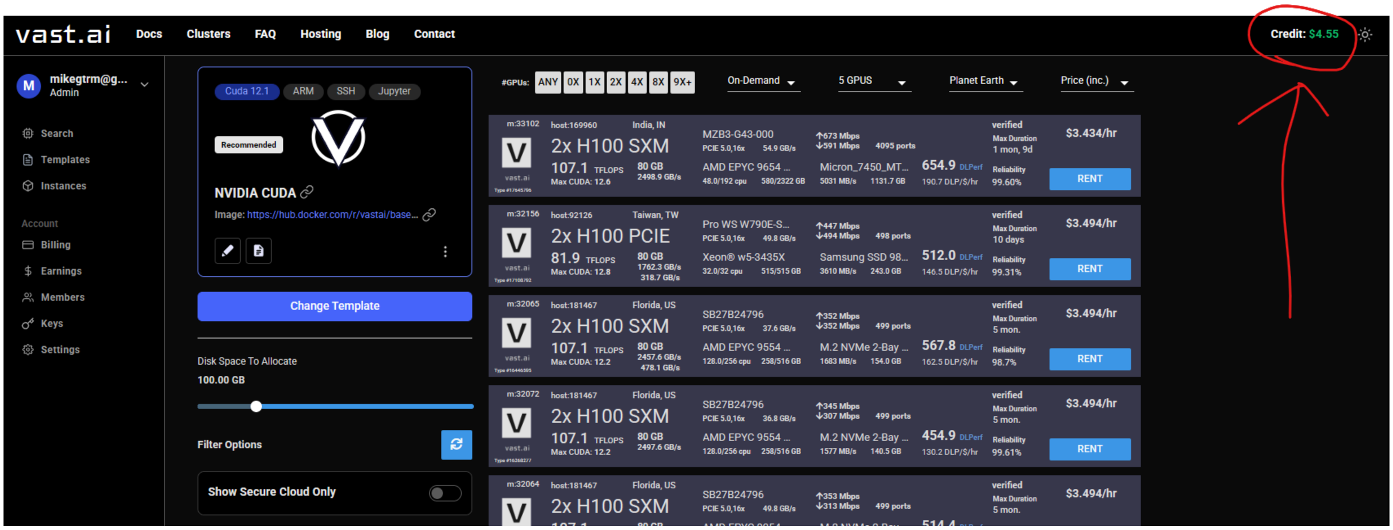
□ Unverified Machines

□ Incompatible Machines

#GPUs: ANY 0X 1X 2X 4X 8X 9X+ | On-Demand | 5 GPUS | Planet Earth | Price (inc.)

m:33102	host:169960	India, IN	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.6	80 GB 2498.9 GB/s	MZB3-G43-000 PCIE 5.0,16x 54.9 GB/s	↑673 Mbps ↓591 Mbps	4095 ports	AMD EPYC 9654 ... 48.0/192 cpu 580/2322 GB	Micron_7450_MT... 5031 MB/s 1131.7 GB	654.9 DLPerf 190.7 DLP/$/hr	verified	Max Duration 1 mon, 9d	Reliability 99.60%	$3.434/hr	Type #17645706
m:32156	host:92126	Taiwan, TW	2x H100 PCIE	81.9 TFLOPS Max CUDA: 12.8	80 GB 1762.3 GB/s 318.7 GB/s	Pro WS W790E-S... PCIE 5.0,16x 49.8 GB/s	↑447 Mbps ↓494 Mbps	498 ports	Xeon® w5-3435X 32.0/32 cpu 515/515 GB	Samsung SSD 98... 3610 MB/s 243.0 GB	512.0 DLPerf 146.5 DLP/$/hr	verified	Max Duration 10 days	Reliability 99.31%	$3.494/hr	Type #17108792
m:32065	host:181467	Florida, US	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2457.6 GB/s 478.1 GB/s	SB27B24796 PCIE 5.0,16x 37.6 GB/s	↑352 Mbps ↓352 Mbps	499 ports	AMD EPYC 9554 ... 128.0/256 cpu 258/516 GB	M.2 NVMe 2-Bay ... 1683 MB/s 154.0 GB	567.8 DLPerf 162.5 DLP/$/hr	verified	Max Duration 5 mon.	Reliability 98.7%	$3.494/hr	Type #16446505
m:32072	host:181467	Florida, US	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2497.6 GB/s	SB27B24796 PCIE 5.0,16x 36.8 GB/s	↑345 Mbps ↓307 Mbps	499 ports	AMD EPYC 9554 ... 128.0/256 cpu 258/516 GB	M.2 NVMe 2-Bay ... 1577 MB/s 140.5 GB	454.9 DLPerf 130.2 DLP/$/hr	verified	Max Duration 5 mon.	Reliability 99.61%	$3.494/hr	Type #16268277
m:32064	host:181467	Florida, US	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2503.9 GB/s	SB27B24796 PCIE 5.0,16x 49.8 GB/s	↑353 Mbps ↓313 Mbps	499 ports	AMD EPYC 9354 ... 64.0/128 cpu 258/516 GB	M.2 NVMe 2-Bay ... 1594 MB/s 132.3 GB	514.4 DLPerf 147.2 DLP/$/hr	verified	Max Duration 5 mon.	Reliability 99.36%	$3.494/hr	Type #18432294
m:31690	datacenter:68137	France, FR	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2503.8 GB/s	K14PN-D24 Series PCIE 5.0,16x 54.9 GB/s	↑6645 Mbps ↓8097 Mbps	2499 ports	AMD EPYC 9554 ... 56.0/224 cpu 387/1548 GB	SAMSUNG MZQL... 5223 MB/s 807.0 GB	646.9 DLPerf 155.5 DLP/$/hr	verified	Max Duration 10 days	Reliability 99.77%	$4.161/hr	Type #15619707
m:29527	host:170701	Bulgaria, BG	2x H100 NVL	96.5 TFLOPS Max CUDA: 12.4	94 GB 2714.6 GB/s	H13DSG-O- PCIE 5.0,16x 55.0 GB/s	↑844 Mbps ↓826 Mbps	499 ports	AMD EPYC 9534 ... 56.0/224 cpu 387/1548 GB	Micron_7450_MT... 10547 MB/s 5791.2 GB	662.7 DLPerf 154.3 DLP/$/hr	verified	Max Duration 1 mon, 10d	Reliability 99.89%	$4.294/hr	Type #13869468
m:26960	host:145222	Thailand, TH	2x H100 SXM			SB27B24796 PCIE 5.0,16x 36.2 GB/s	↑3311 Mbps ↓6581 Mbps	1383 ports				verified	Max Duration 10 mon.		$4.294/hr	

- You’ve got a couple good settings here, make sure you have at least 100gb to allocate (this is disk space) so the space our dataset takes.
- The white buttons let you select how many GPUs.
- The GPU lets you select which GPUs you want (RTX, H-Series, A-Series, etc)
- The TFLOPs is floating point operations, higher is typically better



Credit: $4.55

NVIDIA CUDA

Image: https://hub.docker.com/r/vastai/base...

Disk Space To Allocate
100.00 GB

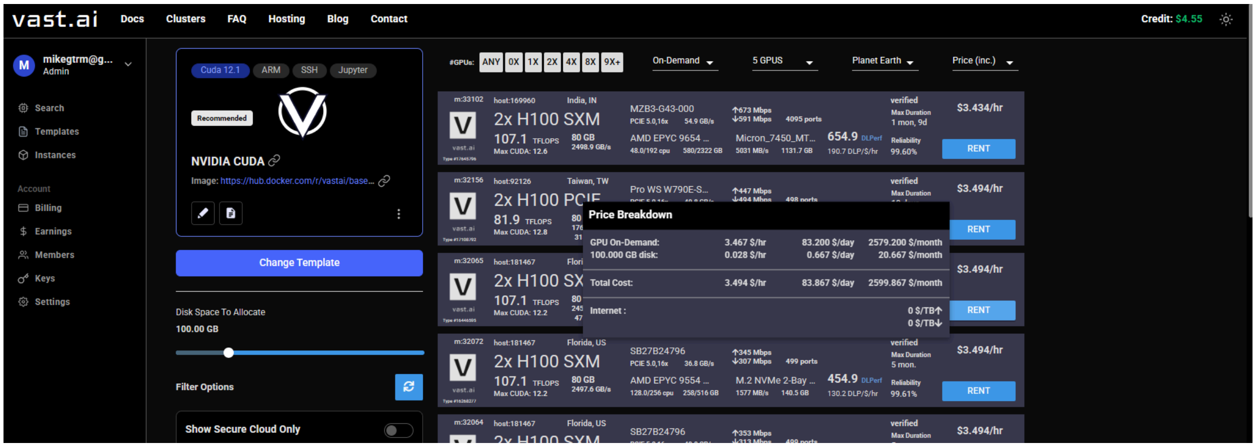
Filter Options

Show Secure Cloud Only

#GPUs: ANY 0X 1X 2X 4X 8X 9X+ | On-Demand | 5 GPUS | Planet Earth | Price (inc.)

m:33102	host:169960	India, IN	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.6	80 GB 2498.9 GB/s	MZB3-G43-000 PCIE 5.0,16x 54.9 GB/s	↑673 Mbps ↓591 Mbps	4095 ports	AMD EPYC 9654 ... 48.0/192 cpu 580/2322 GB	Micron_7450_MT... 5031 MB/s 1131.7 GB	654.9 DLPerf 190.7 DLP/$/hr	verified	Max Duration 1 mon, 9d	Reliability 99.60%	$3.434/hr
m:32156	host:92126	Taiwan, TW	2x H100 PCIE	81.9 TFLOPS Max CUDA: 12.8	80 GB 1762.3 GB/s 318.7 GB/s	Pro WS W790E-S... PCIE 5.0,16x 49.8 GB/s	↑447 Mbps ↓494 Mbps	498 ports	Xeon® w5-3435X 32.0/32 cpu 515/515 GB	Samsung SSD 98... 3610 MB/s 243.0 GB	512.0 DLPerf 146.5 DLP/$/hr	verified	Max Duration 10 days	Reliability 99.31%	$3.494/hr
m:32065	host:181467	Florida, US	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2457.6 GB/s 478.1 GB/s	SB27B24796 PCIE 5.0,16x 37.6 GB/s	↑352 Mbps ↓352 Mbps	499 ports	AMD EPYC 9554 ... 128.0/256 cpu 258/516 GB	M.2 NVMe 2-Bay ... 1683 MB/s 154.0 GB	567.8 DLPerf 162.5 DLP/$/hr	verified	Max Duration 5 mon.	Reliability 98.7%	$3.494/hr
m:32072	host:181467	Florida, US	2x H100 SXM	107.1 TFLOPS Max CUDA: 12.2	80 GB 2497.6 GB/s	SB27B24796 PCIE 5.0,16x 36.8 GB/s	↑345 Mbps ↓307 Mbps	499 ports	AMD EPYC 9554 ... 128.0/256 cpu 258/516 GB	M.2 NVMe 2-Bay ... 1577 MB/s 140.5 GB	454.9 DLPerf 130.2 DLP/$/hr	verified	Max Duration 5 mon.	Reliability 99.61%	$3.494/hr
m:32064	host:181467	Florida, US	2x H100 SXM			SB27B24796 PCIE 5.0,16x 49.8 GB/s	↑353 Mbps ↓313 Mbps	499 ports				verified	Max Duration 5 mon.		$3.494/hr

VERY IMPORTANT



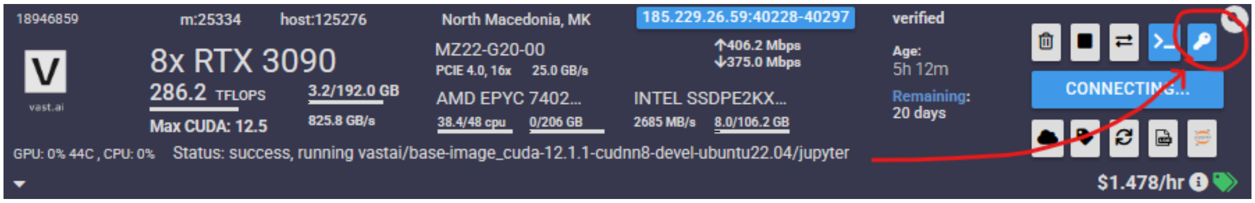
Price Breakdown

GPU On-Demand:	3.467 $/hr	83.200 $/day	2579.200 $/month
100.000 GB disk:	0.028 $/hr	0.667 $/day	20.667 $/month
Total Cost:	3.494 $/hr	83.867 $/day	2599.867 $/month
Internet :			0 $/TB↑ 0 $/TB↓

When you hover over the rent button you get to see the cost breakdown. **Sometimes** you can find some cheap GPUs per hour, but the internet price is like $40 for every Terabyte you download. This typically is only high for GPU providers outside of the US

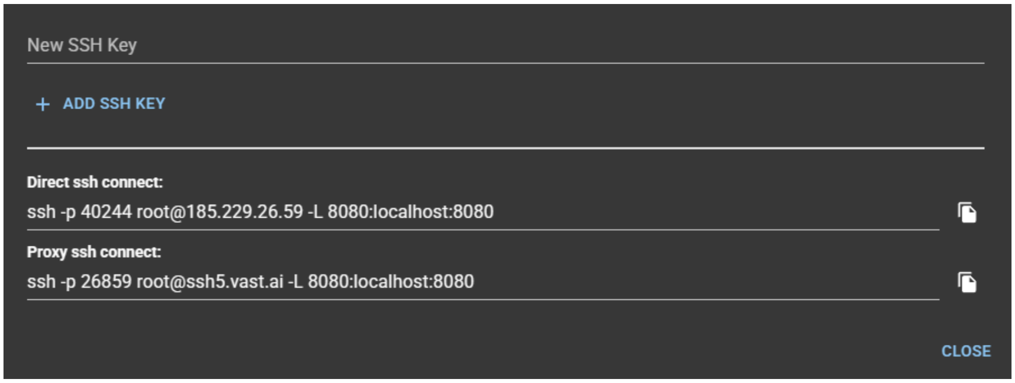
AFTER YOU CLICK RENT

- Create a ssh public key (copy the public key)



Status: success, running vastai/base-image_cuda-12.1.1-cudnn8-devel-ubuntu22.04/jupyter

- Add your **public** key



Direct ssh connect:

```
ssh -p 40244 root@185.229.26.59 -L 8080:localhost:8080
```

Proxy ssh connect:

```
ssh -p 26859 root@ssh5.vast.ai -L 8080:localhost:8080
```

- Copy over the ssh command and ssh into it

AFTER YOU ARE IN THE SYSTEM

- git clone https://github.com/Michaelgathara/GPT
- cd GPT

- curl -LsSf https://astral.sh/uv/install.sh | sh
- source $HOME/.local/bin/env bash
- uv sync
- source .venv/bin/activate
- uv pip install flash-attn --no-build-isolation
- cd models
- nohup python3 -u gpt_train_script.py > train.log 2>&1 &
- bash print_res.sh