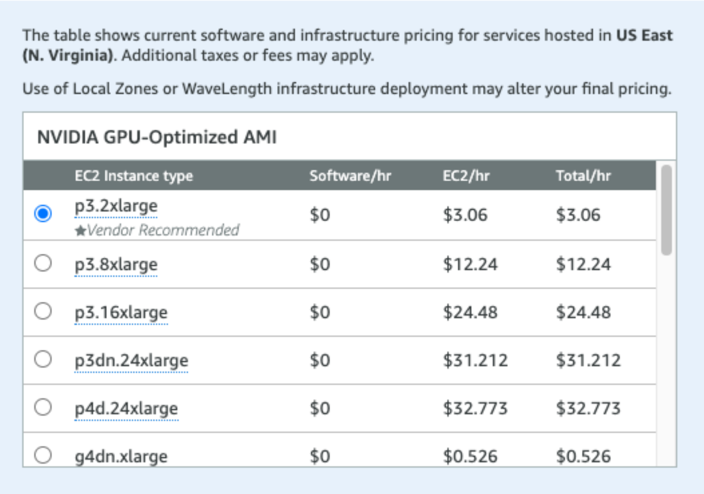
Using AWS EC2 Instances for Training.

Overview

This document describes the cost effective way to create an EC2 instance and run LLM training.

1. Login to AWS account and choose a region us-east-2 or us-west-1 etc.
2. Subscribe to the Nvidia AI/ML AMI from AWS Market place
    a. https://aws.amazon.com/marketplace/pp/prodview-7ikjtg3um26wq
    b. This subscription will take 10-15 mins and may be even faster.
3. The following table provides the pricing for using this image and is dated (i.e, could be different now)
    a.



The table shows current software and infrastructure pricing for services hosted in **US East (N. Virginia)**. Additional taxes or fees may apply.

Use of Local Zones or WaveLength infrastructure deployment may alter your final pricing.

NVIDIA GPU-Optimized AMI

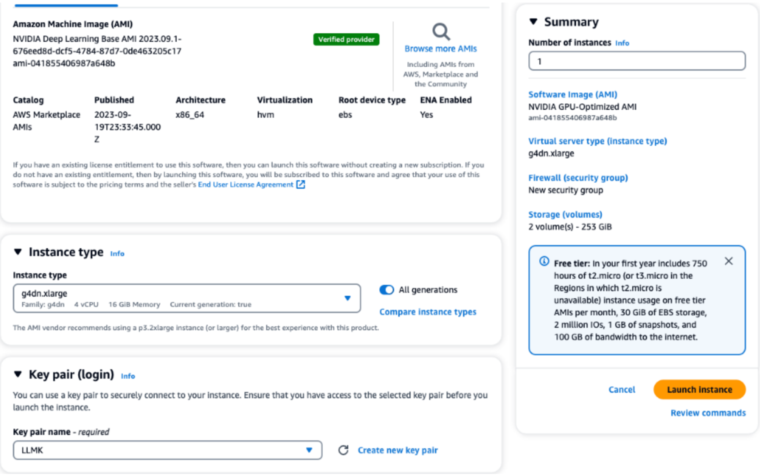
	EC2 Instance type	Software/hr	EC2/hr	Total/hr
●	p3.2xlarge ★*Vendor Recommended*	$0	$3.06	$3.06
○	p3.8xlarge	$0	$12.24	$12.24
○	p3.16xlarge	$0	$24.48	$24.48
○	p3dn.24xlarge	$0	$31.212	$31.212
○	p4d.24xlarge	$0	$32.773	$32.773
○	g4dn.xlarge	$0	$0.526	$0.526

    b. I am using te p3.2xlarge and costs about $3 an hour. (Very expensive). The way to manage the cost is to just launch the EC2, get the LLM trained, upload the result to google cloud and then stop the instance.
    c. I also have a g4dn.2xlarge which costs about $0.75. per hour.
    d. I have not tried the g4dn.xlarge which is the cheapest of all.
4. The instance g4dn.xlarge is relatively cheap, $6.30 per day. Again, when used judiciously you will not go over $50 a month.

Steps

1. Login to the AWS Region where the subscription is enabled.
2. Go to EC2 service
    a. https://us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1
    b. Repace us-east-1 in the above to your region.
3. Click Launch an Instance
    a. Select the instance type to one of the above. I think g4dn.xlarge should work.
    b. Select the AMI: select market place AMI, search for the NVIDIA image above and click select
    c. Change the storage size to 128 or 256G. This will cost almost nothing (may be $5 for a month or so)
    d. Use the autogenerate security group. Allow ssh from your IP.
    e. For the rest accept the defaults and launch the instance

f.



Amazon Machine Image (AMI)
NVIDIA Deep Learning Base AMI 2023.09.1-676eed8d-dcf5-4784-87d7-0de463205c17
ami-041855406987a648b

Catalog	Published	Architecture	Virtualization	Root device type	ENA Enabled
AWS Marketplace AMIs	2023-09-19T23:33:45.000Z	x86_64	hvm	ebs	Yes

If you have an existing license entitlement to use this software, then you can launch this software without creating a new subscription. If you do not have an existing entitlement, then by launching this software, you will be subscribed to this software and agree that your use of this software is subject to the pricing terms and the seller's End User License Agreement

▼ Instance type Info

Instance type
g4dn.xlarge
Family: g4dn 4 vCPU 16 GiB Memory Current generation: true

The AMI vendor recommends using a p3.2xlarge instance (or larger) for the best experience with this product.

▼ Key pair (login) Info

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance.

Key pair name - *required*
LLMK

▼ Summary

Number of instances Info
1

Software Image (AMI)
NVIDIA GPU-Optimized AMI
ami-041855406987a648b

Virtual server type (instance type)
g4dn.xlarge

Firewall (security group)
New security group

Storage (volumes)
2 volume(s) - 253 GiB

Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million IOs, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

g. Note: Select your existing key pair or just create a new one (right in the selection place). Save this key and you need this to ssh. If you loose this then you loose complete access and the EC2 is toast.

h. Put your IP in the security group to ssh. If its Ok, if you make some mistake here. All these are fixable.

i. Only thing that cannot be changed is the instance type (g4dn ...) and the AMI

4. Access Setup

   a. Go to Hugging Face and create an account and also create a login token

   b. Go to meta and register the same e-mail used in Hugging face

   c. Get the authorization from meta (took me about 2-3 minutes)

   d. Go to Hugging face and now select the meta-llama/Llama-xxx. (any image) will do and the get authorization to use

   e. This took me about 5-10 mins wait

5. Login to the EC2 instance using the saved key (ssh -i ~/.ssh/save_key.pem ubuntu@IP Address form EC2 console, it should be a global IP)

   a. Wait for few mins for the NVIDIA drives to get installed.

   b. Run lsmod and make sure that you have the NVIDIA kernel modules (you must see some names with nvidia in it)

6. Training

   a. Create a LLM dir, create model dir if desired.

   b. Create a virtual env. (python3 -m venv .venv)

   c. source .venv/bin/activate

   d. pip3 install autotrain-advanced

   e. You will see a whole bunch of packages installed

7. Login to hugging face

   a. huggingfacelogin_cli

   b. Put your login token from the HF site

   c. If you see some message about git config, just do that and relogin.

8. Get some training data and put that in your LLM/Model/ folder. as a train.csv.

   a. The columns names should be “text”, “text_lable”, “prompt” etc.

9. Train the model using the command

   a. autotrain llm —train —project_name prompt-tune-bloomz —model bigscience/bloomz-560m —data_path . —use_peft —use_int4 —learning_rate 2e-4 —train_batch_size 6 —num_train_epochs 50 —trainer sft

   b. In the above command change the project_name and the model

   c. The rest should be fine.

   d. The model is saved in the directory named after your project_name

   e. Now zip this whole directory and ship to google

10. Stop the instance to save money

    a. NOTE: First set the “Termination Protection” by Enabling termination protection from the menu

    b. Then stop the instance.

11. You can restart the instance and train another model