GARVOO

www.garvoo.com
+1 408-833- 6100
sales@garvoo.com

Is Private Cloud a better option for Data Science Platform?

Background

Public cloud provides an easy & convenient starting point for your organization's Artificial Intelligence/ Machine Learning initiatives.

However as your AI/ ML or other High Performance Computing (HPC) applications scale, you need to reassess. Depending on the configuration and applications a move can result in savings of over 50%. Savings which can add up to millions of dollars over a three year planning period. In addition to the cost savings you get improved security and performance.

Buying what you need can save you money

For ML applications Amazon Web Services offers the choice of using their P2 and P3 instances. These use the older NVIDIA Telsa K80 GPUs (2014 release with Kepler architecture) and the latest NVIDIA Telsa V100 GPUs, respectively. The AWS pricing for these are:

Instance type (AWS pricing page)	On demand per hour	1 year reserved (effective per hour	Total 1 year spend range (with full usage)
p2.16xlarge, with 16 NVIDIA K80 GPUs	$14.40	$9.82	$126,000 to $86,000
p3.16xlarge, with 8 NVIDIA Telsa V 100 GPUs	$24.48	$16.70	$214,000 to $146,000

However for a private cloud deployment in addition to the above choices, you can evaluate many other options from NVIDIA's latest Turing (2018 release) & Volta (2017 release) GPU architecture. Some of the popular and latest options include:

- Titan RTX
- RTX 2080 Ti
- Titan V

These latest GPUs support Tensor cores which are specifically designed for deep learning matrix computations. Given these advancements it is possible to get the same performance as a P2 - K80 instance at 20% to 50% of the cost.

In contrast the P3 instances with the latest Tesla V100 represents the top of the line performance today. However for many applications not requiring the Double Precision - 64 bit floating point calculations, it can be an overkill. The options mentioned above can provide throughputs of 75% to 85% as compared to Tesla V100 while costing only about 25% to 50%.

These comparisons are based on a benchmark test which measured the # of images processed per second while training the following ResNet50 & VGG16 neural networks.

Summary

- Public cloud GPU solutions are easy to set up.
- Options are limited in terms of GPUs.
- Worthwhile to reassess as you scale to get 50% savings and other advantages like improved security.
- Managed GPU private cloud solutions without headache of having to manage the frameworks & advantages of significant cost savings.

Security & Privacy

A private cloud infrastructure may be necessary to be in compliance with regulations which restrict usage of 3rd party data storage. This is applicable in the healthcare, finance, government, defense sectors and in many cases self enforced by best in class corporate governance practices.

For a private cloud infrastructure security considerations can be enhanced by undertaking a number of steps like creating Layer 2 network isolation and custom policies.

Bottomline

Balancing GPU performance with costs, security, application needs is complicated. We take the complexity out of it by ensuring you have all needed deep frameworks pre-installed. Frameworks such as TensorFlow, PyTorch, Apache MXNet, Caffe, Caffe2, Theano, Torch, and Keras to train sophisticated, custom AI models.

Talk to us at Garvoo, we can help.