According to EC2 instance of AWS, there are 3 types of instances families have relation with GPU. They are GPU graphics, GPU instances and GPU compute respectively. For each of family, it uses different kind of models. The g3 model is used in GPU graphics, g2 model is used for GPU instances, and both p2 and p3 are used for GPU compute. The parameters of each model are shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | GPU | GPUs | vCPU | Mem(GiB) | GPU Mem(GiB) |
| p2.2xlarge | NVIDIA K80 | 1 | 8 | 61 | 12 |
| p2.8xlarge | NVIDIA K80 | 8 | 32 | 244 | 96 |
| p2.16xlarge | NVIDIA K80 | 16 | 64 | 488 | 192 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | GPU | GPUs | vCPU | Mem(GiB) | GPU Mem(GiB) |
| p3.2xlarge | NVIDIA Tesla V100 | 1 | 8 | 61 | 16 |
| p3.8xlarge | NVIDIA Tesla V100 | 4 | 32 | 244 | 64 |
| p3.16xlarge | NVIDIA Tesla V100 | 8 | 64 | 488 | 128 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | GPU | GPUs | vCPU | Mem(GiB) | GPU Mem(GiB) |
| g3.4xlarge | NVIDIA Tesla M60 | 1 | 16 | 122 | 8 |
| g3.8xlarge | NVIDIA Tesla M60 | 2 | 32 | 244 | 16 |
| g3.16xlarge | NVIDIA Tesla M60 | 4 | 34 | 488 | 32 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | GPU | GPUs | vCPU | Mem(GiB) | GPU Mem(GiB) |
| g2.xlarge | NVIDIA K520 | 1 | 8 | 15 |  |
| g2.8xlarge | NVIDIA K520 | 4 | 32 | 60 |  |

From the table, we can see that each of model use different type of GPU. In order to find the most suitable model for our project, we do a comparison for those type of GPU below. In the project, we want to run tensorflow in the cloud. Following the instruction of install tensorflow. We need also install CUDA and cuDNN to satisfy NVIDIA requirement to run tensorflow with GPU support. In order to find out the best GPU we need, we first go through the NVIDIA GPUDirect to get information about CUDA and GPU Computing. From the direct, we can get the compute capability of latest CUDA Capable NVIDIA GPUs. It result is shown in the table below.

|  |  |
| --- | --- |
| GPU | Compute Capability |
| K80 | 3.7 |
| Tesla V100 | 7 |
| Tesla M60 | 5.2 |

From the table, we can get that the Tesla V100 has better compute capability than Tesla M60 and K80.

<https://developer.nvidia.com/cuda-gpus> (NVIDIA GPUdirect)

What is more, from the features of each models, we can get that, for p3 model, each GPU pairing 5120 CUDA Cores and 640 Tensor Cores. For p2 models, every GPUs have 2496 parallel processing cores. Besides, there are 2048 parallel processing cores for g3 models. It also shows the compute capability from the number of cores of different type of GPUs used by different kinds of models.

[https://www.lifewire.com/what-is-nvidia-cuda-834095 (What](https://www.lifewire.com/what-is-nvidia-cuda-834095%20(What) is CUDA core)

From all information above, we can get that p3 models have the best computing capability in all models.

Besides, we are still going to take consideration about memory bandwidth of GPU. The memory bandwidth of GPU is an important characteristic related to deep learning. It shows the ability of GPU to handle large amount of data and it would be a most important performance metric. The memory bandwidths of different GPUs are shown as below

|  |  |
| --- | --- |
| GPU | Memory bandwidth |
| Tesla V100 | up to 900GB/s |
| Tesla M60 | 320GB/s |
| K80 | 480GB/s |
| K520 | 320GB/s |

From the table above, we can also get that Tesla V100 has the biggest memory bandwidth comparing with other three types of GPUs.

Then we talk about other hardware. While talking about CPUs, each of these EC2 models use Intel Xeon E5-2686 processors. Therefore, we can avoid the effect to our result of different types of CPU models. And the number of vCPU and memory are shown above.

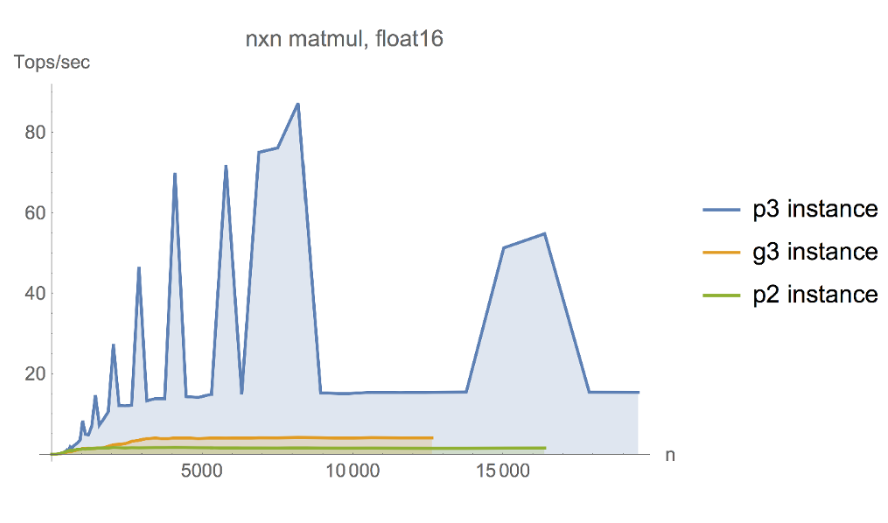
Last but not least, we will show some trade-off between cost and performance. First, we list the price of all models here.

|  |  |
| --- | --- |
| models | price |
| p2.xlarge | $0.9 per hour |
| p2.8xlarge | $7.2 per hour |
| p2.x16xlarge | $14.4 per hour |
| p3.2xlarge | $3.06 per hour |
| p3.8xlarge | $12.24 per hour |
| p3.16xlarge | $24.48 per hour |
| g3.4xlarge | $1.14 per hour |
| g3.8xlarge | $2.28 per hour |
| g3.16xlarge | $4.56 per hour |

And for the performance, we get it from Yaroslav Bulatov’s experiment

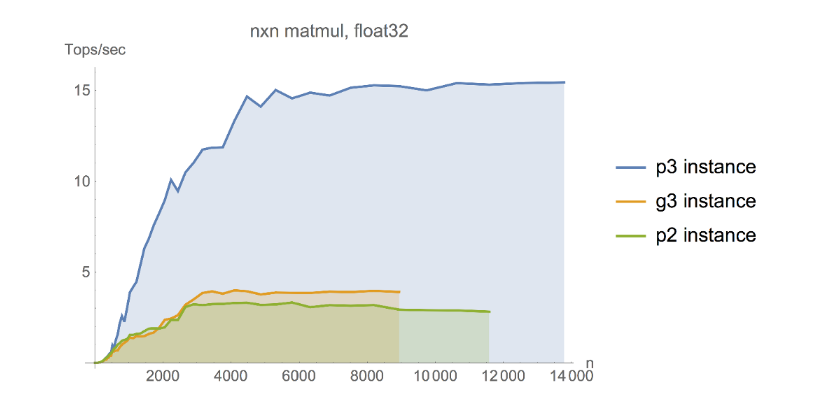
<https://medium.com/@yaroslavvb/peak-performance-of-amazon-p3-instances-f2bc48f9ef71>

While talking about machine learning, most part of computation using in machine learning can be transformed into matrix multiplication. Therefore, the performance of different GPU in matrix multiplication can be partly used to show the performance of different GPU in running machine learning. According to the experiment, the peak performance of comparing for float16 are shown below.



The peak performances are 87.2 T ops/second for p3, 4.1 T ops/second for g3, 1.6 T ops/second for p2 respectively.

Comparing for float32, we can see that



15.4 T ops/second for p3, 4.0 T ops/second for g3 and 3.3 T ops/second for p2 respectively.

Therefore, we can get that the peak performance of p3 instance is much better than p2 and g3 instance. The advantage of it would be so huge in the competition. Although talking price into consideration, the rate of performance and price of p3 is still much greater than p2 and g3 instance, which shows that p3 instance has the best trade-off between performance and price according to the experiment.

In conclusion, we would finally choose p3.xlarge model in our project, when we are considering our budget also. It is the most suitable model for us to run our project in cloud.

In additional, we are going to talk about the GPU in our mobile device, the mobile devices we have are 16G iphone6 and 64G iphone 6s. The processors for iphone 6 and iphone 6s are Apple A8 processor and Apple A9 respectively. GPUs contained in the processors are provided by Imagination Technologies Group. The one uses in Apple A8 is a quad-core PowerVR series 6XT GX 6450. A Six-core PowerVR series 7XT GT7600 is used in Apple A9.