As mentioned in the introduction, we use deep learning on traffic signs and pedestrian recognition. For deep learning model, it can be run either on CPU or on GPU. However, we will run it on GPU to make it faster. When talking about GPU, it is specifically designed to accelerate the creation of images in a frame buffer intended for output. In order to do this efficiently, GPU needs to have strong ability on computing of the position of geometry points and color. Most of these computations involve vector and matrix operation which are widely used in deep learning model. What is more important, by using GPU, we can do the computation on a large amount of data in higher parallelism. Those are in a nutshell why people use GPU instead of CPU for training a deep learning model.

According to official references, Google cloud provides three different available GPU models now. They are NVIDIA Tesla V100, NVIDIA Tesla P100 and NVIDIA Tesla K80 respective. For NVIDIA Tesla V100, it is beta. NVIDIA Tesla P100 and NVIDIA Tesla K80 are generally available.

The parameters of different kind models are shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GPU model | GPUs | GPU  memory | Available  vCPUs | Available  memory | Available  Zones |
| NVIDIA Tesla V100 | 1GPU | 16GB  HBM2 | 1-12  vCPUs | 1-78GB | * US-west1-a * US-west1-b * US-central1-a * US-central1-f * Europe-west4-a * Asia-east1-c |
| 8GPUs | 128GB  HBM2 | 1-96  vCPUs | 1-624GB |
| NVIDIA  Tesla  P100 | 1GPU | 16GB  HBM2 | 1-16  vCPUs | 1-104GB | * US-west1-b * US-central1-c * US-central1-f * US-east1-b * US-east1-c * Europe-west1-b * Europe-west1-d * Asia-wast1-a * Asia-east1-c * Europe-west4-a |
| 2GPUs | 32GB  HBM2 | 1-32  vCPUs | 1-208GB |
| 4GPUs | 64GB  HBM2 | 1-64  vCPUs  (US-east1-c, Europe-west1-d, Europe-west1-b)  1-96  vCPUs  (all other zones) | 1-208GB  (US-east1-c, Europe-west1-d, Europe-west-1-b)  1-624GB  (all other zones) |
| NVIDIA  Tesla  K80 | 1GPU | 12GB  GDDR5 | 1-8  vCPUs | 1-52GB | * US-west1-b * US-central1-c * US-east1-c * US-east1-d * Europe-west1-d * Asia-east1-a * Asia-east1-b |
| 2GPUs | 24GB  GDDR5 | 1-16  vCPUs | 1-104GB |
| 4GPUs | 48GB  GDDR5 | 1-32  vCPUs | 1-208GB |
| 8GPUs | 96GB  GDDR5 | 1-64  vCPUs | 1-416GB  (asia-east1-a and us-east1-d)  1-208GB  (all other zones) |

Instances with lower numbers of GPUs are limited to a maximum number of vCPUs. In general, higher numbers of GPU dies allow users to create instances with higher numbers of vCPUs and system memory.

In our project, we are going to choose the generally available GPUs in Google cloud which are NVIDIA Tesla P100 and NVIDIA Tesla K80 instead of using Google cloud beta.

In following part, We are going to do a comparison between NVIDIA Tesla P100 and NVIDIA Tesla K80 to show which one is more suitable for running our project.

First, we show the key hardware difference between these two kinds of GPUs below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| GPU | Cores | CUDA cores | Frequency | GFLOPs  (double) | Memory | Memory bandwidth |
| NVIDIA Tesla P100 | 2\*13(SMX) | 2\*2496 | 562MHZ | 2\*1455 | 2\*12GB | 2\*240GB/s |
| NVIDIA Tesla  K80 | 56(SM) | 3584 | 1126MHz | 4670 | 16GB | 720GB/s |

In the following section, we are going to focus on key characteristics of GPU related to deep learning.

Memory bandwidth would be the most important performance metric we are going to talk about. It shows ability of the GPU to handle large amount of data. The memory bandwidths of different GPUs are shown as below:

|  |  |
| --- | --- |
| GPU | Memory bandwidth |
| NIVDIA Tesla P100 | 720 GB/s per core |
| NVIDIA Tesla K80 | 240GB/s per core |

From the table, we can get that P100’s memory bandwidth is 3x the memory bandwidth of the K80 per core.

Besides, we would focus on processing power of different kinds of GPUs. Processing power indicates the highest speed can GPU achieve to crunch data. We use the number of CUDA cores multiplied by the frequency of core to represent it. The result is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| GPU | CUDA core | Frequency | Processing power |
| NVIDIA Tesla P100 | 2496  (per core) | 562MHz | 1402752 |
| NVIDIA Tesla K80 | 3584  (per core) | 1126MHz | 4035584 |

From the table, we can get that the processing power of NVIDIA Tesla P100 is much better than processing power of NVIDIA Tesla K80 when we just using a single core. Processing power of P100 almost 4x better than processing power of K80.

Furthermore, we get some interesting information from NVIDIA GPU direct. The computation abilities of the different kind of CUDA GPUs are given in this direct.

|  |  |
| --- | --- |
| GPU | Compute capability |
| NVIDIA Tesla K80 | 3.7 |
| NVIDIA Tesla P100 | 6.0 |

We can also get that the compute ability of NVIDIA Tesla P100 is much better than NVIDIA Tesla K80. It proves comparisons we show above.

Additional hardware is still an important part we should focus on if we want to run our deep learning model on Google cloud platform. According to the budget we have, we implement 1 vCPU in our instance, 3.75GB memory and 10 GB disk. The using image is Ubuntu 16.04 LTS.

At last, we do some trade-off between cost and performance of those two kinds of GPUs.

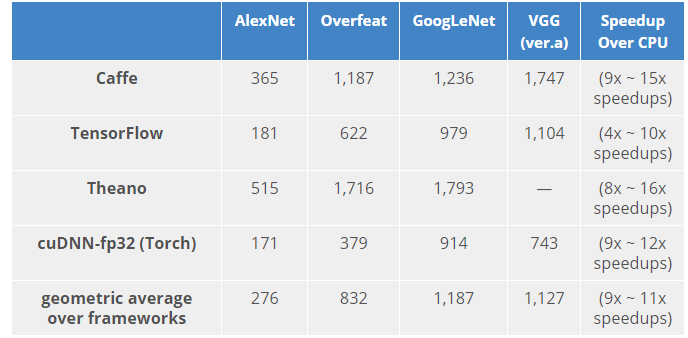
We first show the price of different GPUs in different regions. The price of different models of GPUs is shown below:

|  |  |  |
| --- | --- | --- |
| Place | Model | Price |
| Iowa | NVIDIA Tesla P100 | 1.46 per GPU |
| NVIDIA Tesla K80 | 0.45 per GPU |
| Oregon | NVIDIA Tesla P100 | 1.46 per GPU |
| NVIDIA Tesla K80 | 0.45 per GPU |
| South Carolina | NVIDIA Tesla P100 | 1.46 per GPU |
| NVIDIA Tesla K80 | 0.45 per GPU |
| Belgium | NVIDIA Tesla P100 | 1.60 per GPU |
| NVIDIA Tesla K80 | 0.49 per GPU |
| Netherlands | NVIDIA Tesla P100 | 1.60 per GPU |
| NVIDIA Tesla K80 | Not available |
| Taiwan | NVIDIA Tesla P100 | 1.60 per GPU |
| NVIDIA Tesla K80 | 0.49 per GPU |

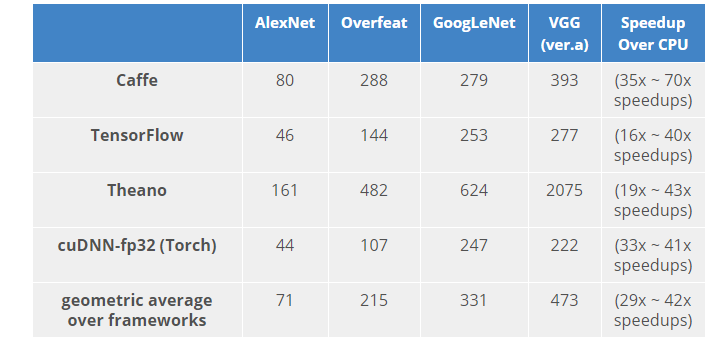
We can see that almost in each area, the price of NVIDIA Tesla P100 would be nearly 4x expensive than NVIDIA Tesla K80

Then we show the performance of running deep learning model on different GPU models. We get the evaluation from John Murphy experiment posted on Microway.com.(https://www.microway.com/hpc-tech-tips/deep-learning-benchmarks-nvidia-tesla-p100-16gb-pcie-tesla-k80-tesla-m40-gpus/)John runs different benchmarks on different GPU models. The results are shown below:

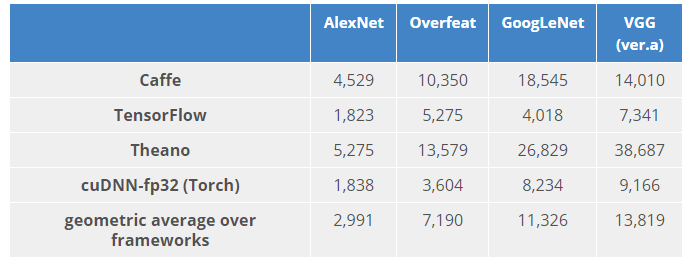
Tesla K80 Benchmark results



Tesla P100 Benchmark results



CPU only Benchmark Results



From the experiment, we can get that the rate of speed-up on Tensorflow and GPU prices are same for Tesla P100 and Tesla K80. The trade-off between performance and price are almost same for NVIDIA Tesla P100 and NVIDIA Tesla K80 are same. However, considering the cost of time on running our model, we are going to choose NVIDIA P100.

In addition, 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 is used in Apple A8 is a quad-core PowerVR series 6XT GX 6450. A Six-core PowerVR series 7XT GT7600 is used in Apple A9.