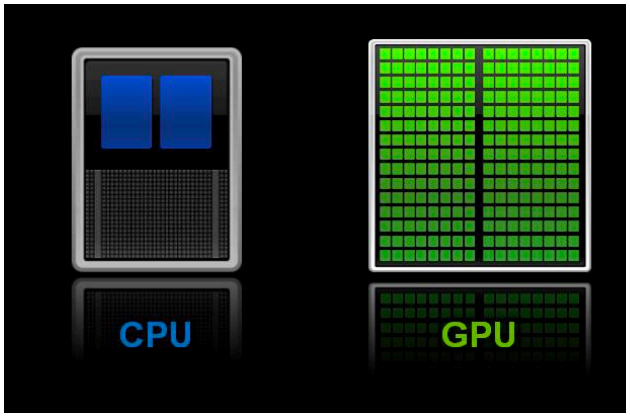
Deploying GPU workloads with Docker EE & Kubernetes

Niraj Bhatt

Solution Architect, Docker Inc.

GPUs – What are they?

- GPU stands for Graphical Processing Unit
- A component of most modern computers that is designed to perform computations needed for 3D graphics
- Common use is to perform accelerated graphics for video games



GPUs – What are they (technically)?

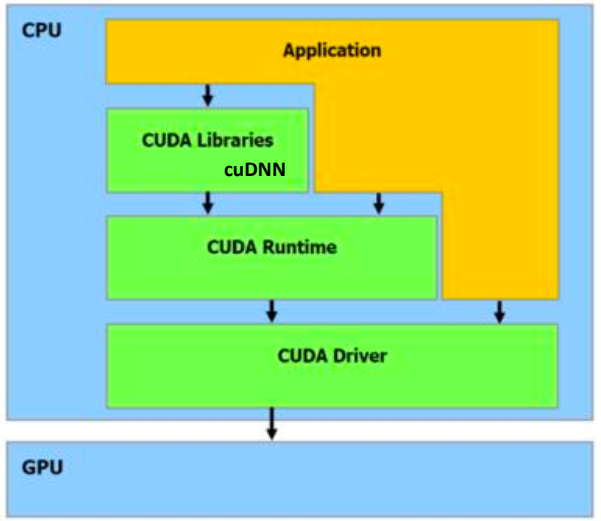
- GPU is basically a large array of small processors, performing highly parallelized computation
- While each of the “CPUs” in a GPU is quite slow the highly parallel GPU structure makes them more efficient for processing large blocks of data
- For instance, Volta™, the latest GPU architecture from Nvidia, offers the performance of up to 100 **CPUs** in a single GPU

The GPU Rush

- Public cloud providers are rushing to augment their compute services with GPUs - AWS, Azure, GCE
- Their intended use case is of course beyond accelerated graphics
- Parallel processing capabilities of GPUs are highly desired by machine and deep learning algorithms
- In research done by Indigo, it was found that while training deep learning neural networks, GPUs can be 250 times faster than CPUs
- Such application of GPU for non-display use cases is popularly referred to as GPGPU (General Purpose GPUs)

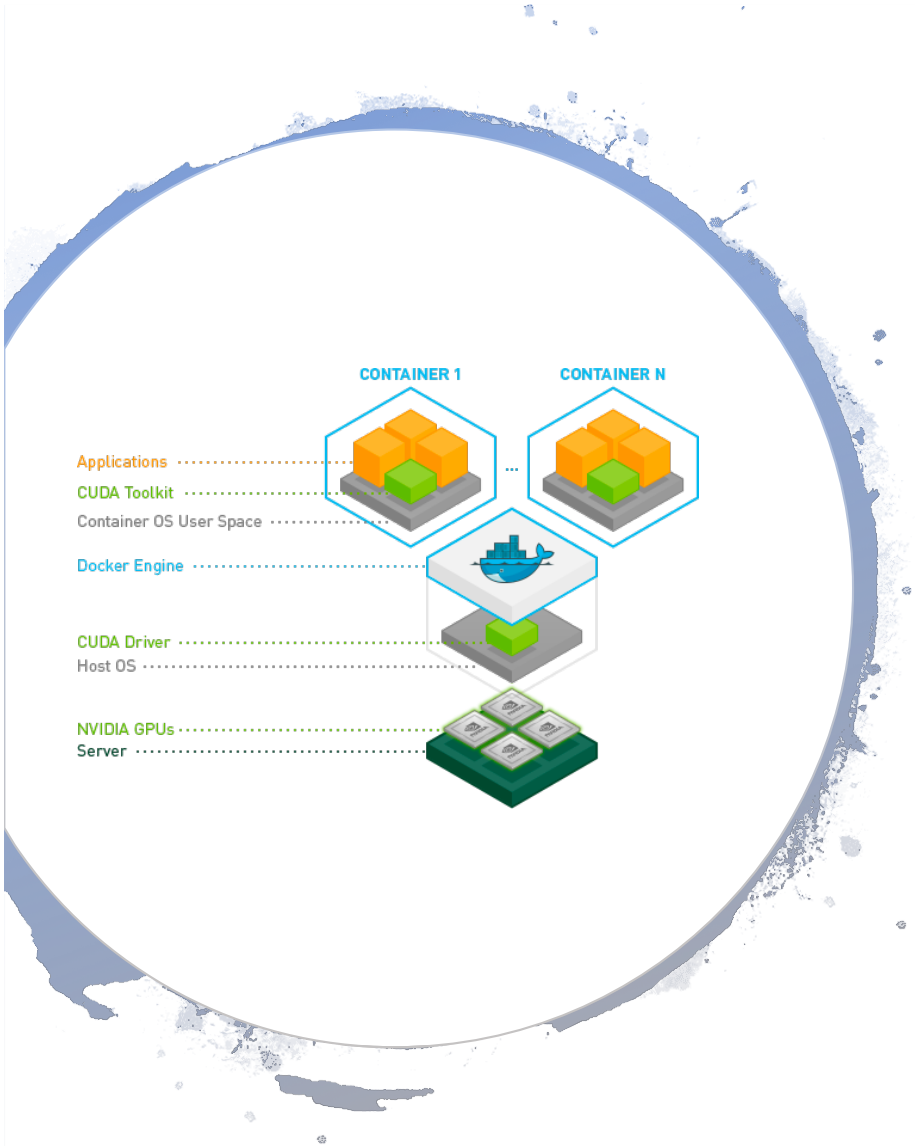
Major Players in GPU industry

- Nvidia and AMD control nearly 100% of the market as of 2018
- Respective market shares are 66% and 33%
- All major public cloud providers today offer Nvidia based GPU instances
- Intel is planning to launch GPUs based on its Xe architecture by mid-2020
- Google has developed a specialized AI accelerator TPU (Tensor Processing Unit) to be used with its Tensorflow framework (these are generally referred to ASICs – App Specific Integrated Circuits)
- AWS Inferentia is a machine learning inference chip scheduled to launch late 2019
- Public clouds also offer FPGA (Field Programmable Gate Arrays) compute instances



Programming GPUs

- Programming models are largely dependent on vendor ecosystem
- For instance, **CUDA** is a parallel computing platform and programming model developed by Nvidia for general computing on its GPUs
- CUDA enables developers to harness the power of GPUs for compute intensive workloads
- NVIDIA CUDA Deep Neural Network library (**cuDNN**) is a GPU-accelerated library of primitives for deep neural networks
- cuDNN accelerates widely used deep learning frameworks including Caffe, Caffe2, Chainer, Keras, MATLAB, MxNet, TensorFlow, and PyTorch.



Containerizing GPU workloads

- CUDA toolkit available as Docker image includes container runtime library and utilities
- Container runtime nvidia-docker has evolved considerably starting as wrapper around Docker in 2016
  - nvidia-docker run -ti --rm nvidia/cuda #docker wrapper
  - docker run --rm --runtime=nvidia -ti nvidia/cuda #runc hook
  - docker run -ti --gpus all nvidia/cuda #docker >=19.03
- K8s v1.16 is validated for docker 18.09 only so --gpus option hasn't reached K8s yet
- For K8s you will also require a device plugin to expose nvidia.com/gpu as a schedulable resource


Demo


Questions?


Thank You!