CS6023 - GPU Programming

# GPU cluster tutorial

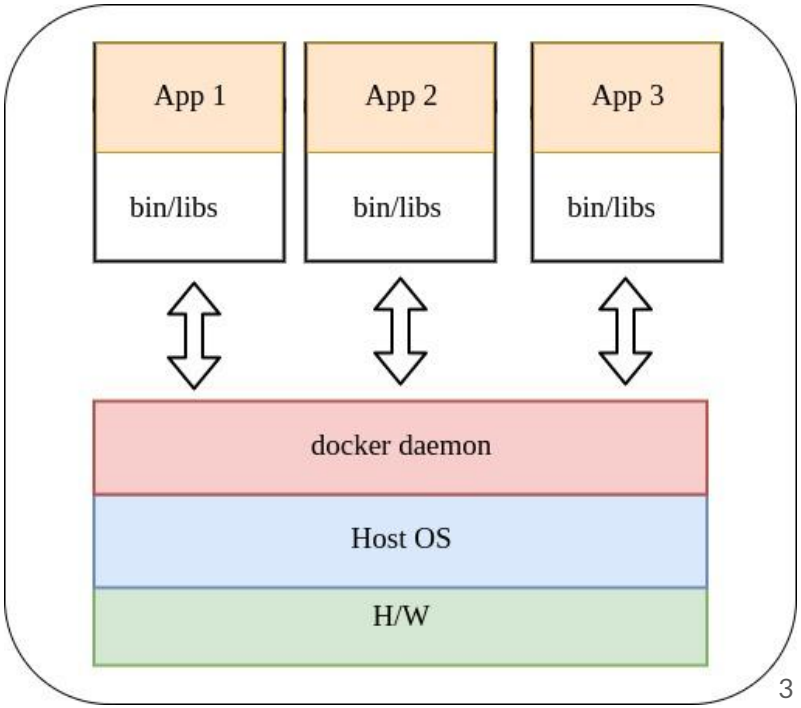
01/02/2019

# Intro

- Overview of cluster manager
- Accessing the cluster
- Job Management

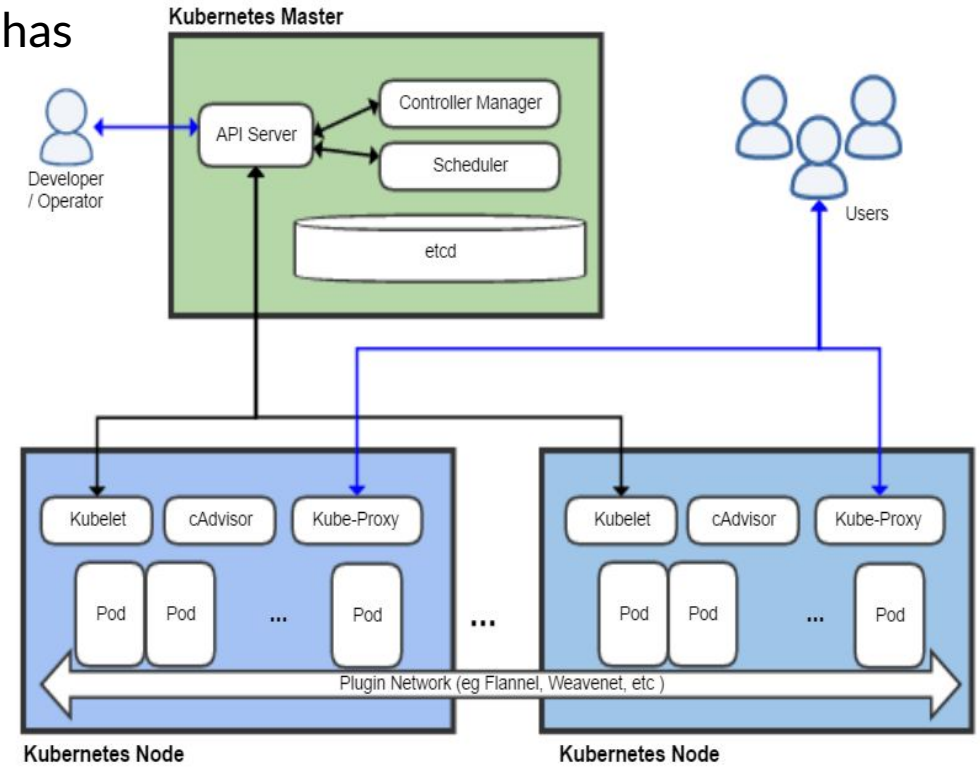
# Overview: Kubernetes

- Kubernetes is a platform for automated management of containers.
- Container is a single entity that comprises of the application and all its dependencies.
- Can be used as cluster manager.



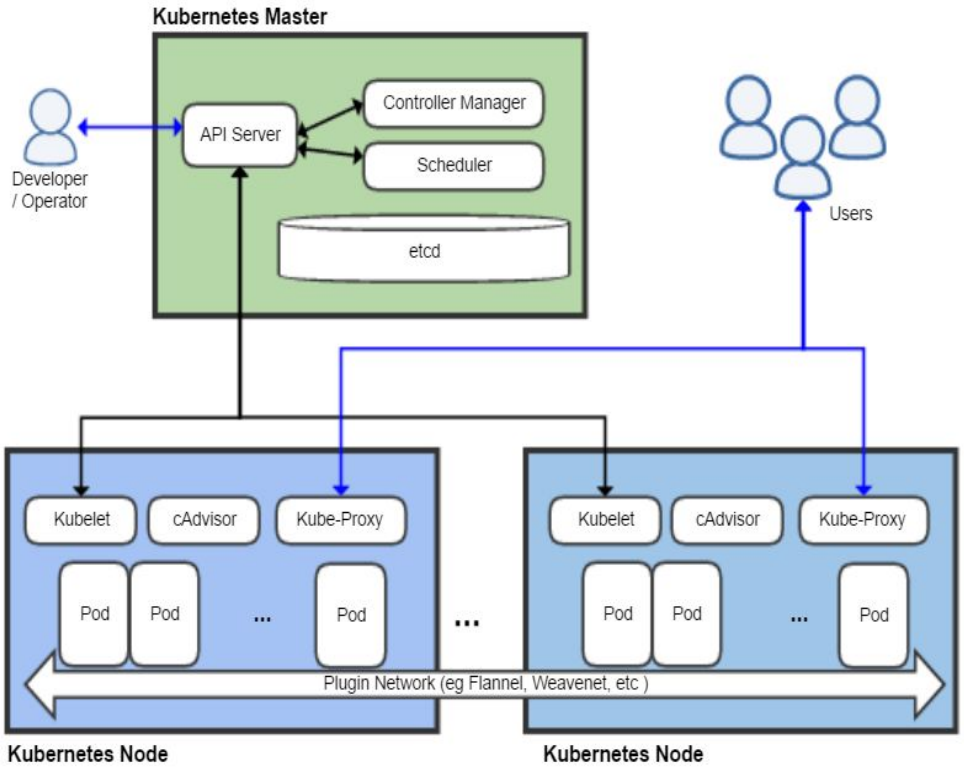
# Overview: Kubernetes

- A typical Kubernetes cluster has the following 3 components
  - Master
  - Node
  - Pod



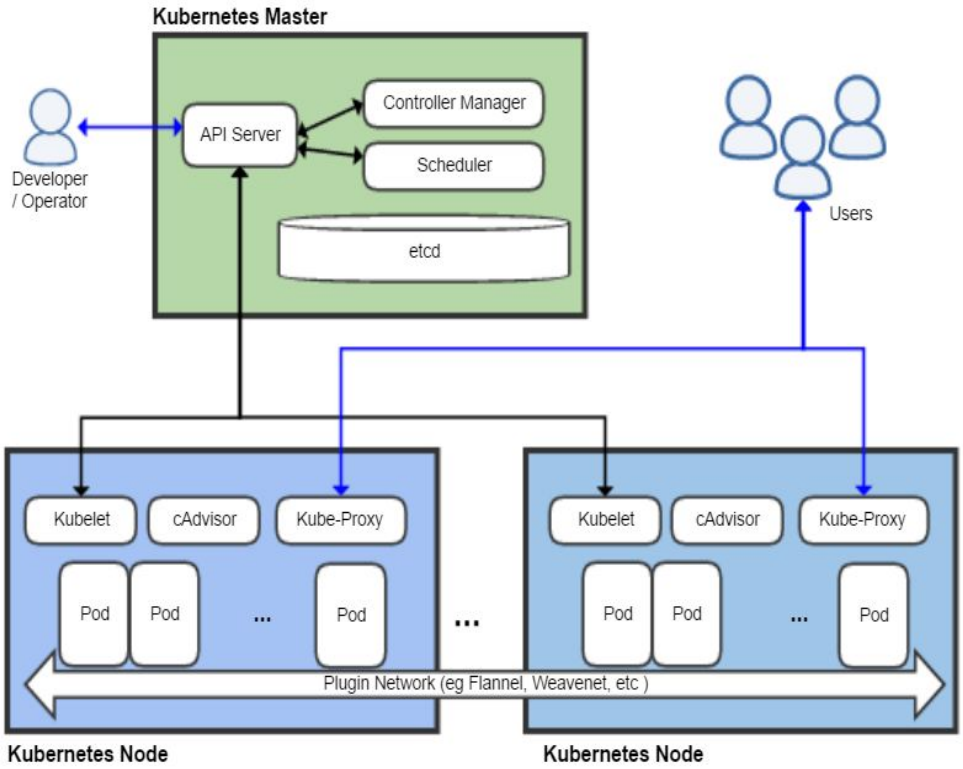
# Overview: Kubernetes

- Master:
  - Maintains the cluster.
  - Schedules jobs on the cluster.



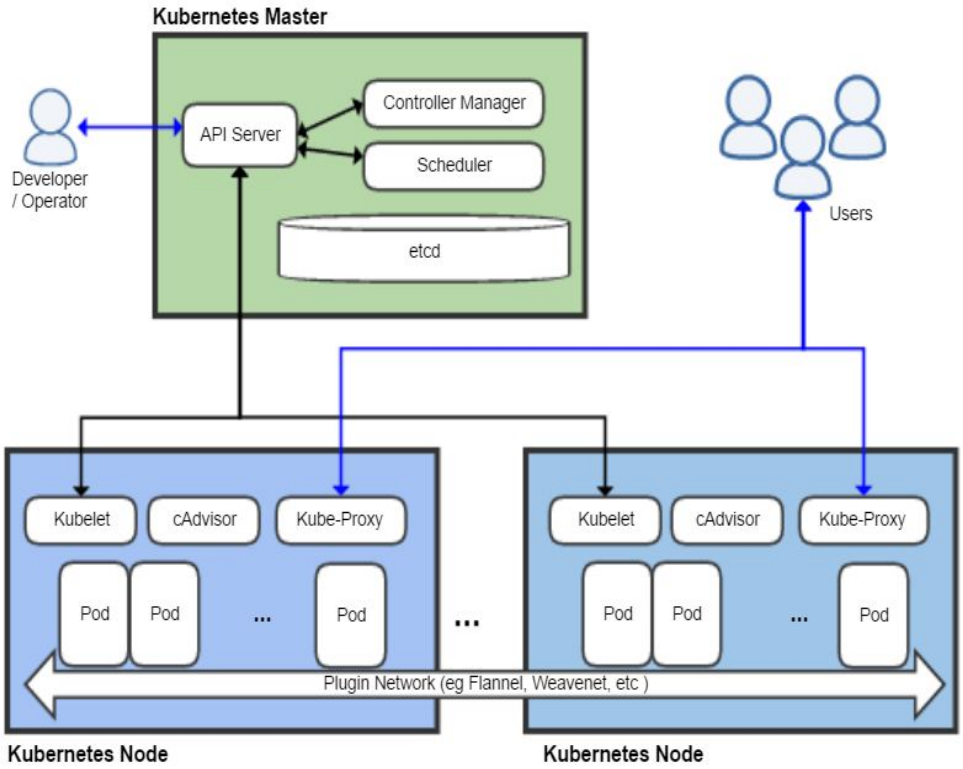
# Overview: Kubernetes

- Node:
  - Where the jobs actually run.
  - Interacts with the master through a service called **kubelet**



# Overview: Kubernetes

- Pod:
  - Jobs are referred to as containers.
  - A pod is a collection of containers.
  - A Pod runs on a Node.



# Overview

Master:

- Core i7 Machine
- 16 GB RAM
- 4 Cores


Slave:

- Xeon Server
- 256 GB RAM
- 16 Cores
- 4 K40 cards

# Basic Information

1. Every user has to create an account on the cluster.

   Fill the form: https://goo.gl/gch3Q7

2. You have **time limits** on the cluster.
   - 1 hr per day
   - 7 hrs per assignment

# Accessing the cluster

1. **Linux / macOS**

   ```
   >ssh -X <username>@10.21.230.1
   ```

2. **Windows**

   Try **PuTTY** & **WinSCP**

You will now have access to the master node of the cluster.

Default password is your username. You will be prompted to change it after first login.

# Job Management

1. Write your source code in the master itself (or) copy your files from your computer to the master using **scp.**
2. Compile the source code in the master using **nvcc.**
3. Run your executable by submitting on the slave by submitting a job (aka pod). (How?)

# Basic commands

- **gsub** - to submit a job

```
user3@gpumaster-machine:~$ gsub job.sh

pod "user3-pod" created
```

- **gstat** - to view the status of the job

```
user3@gpumaster-machine:~$ gstat

NAME           READY          STATUS         RESTARTS   AGE

user3-pod      0/1            Completed      0          3h
```

# Basic commands (contd.)

- **`gdel`** - to delete a job

```
user3@gpumaster-machine:~$ gdel

pod "user3-pod" deleted
```

- **`gtime`** - to view the remaining quota

```
user3@gpumaster-machine:~$ gtime

Today’s remaining quota : 0h 55m 32s

Total remaining quota : 5h 42m 29s
```

# Notes

- A user can run only one job/pod at a time.
- If you want to access any file/directory in the job, use the full path: `/home/<username>/<your_dir_or_file>`
- ‘`gstat -o`’ will show the output of your program if not redirected to a file
- If (and only if) ‘`gdel`’ doesn’t kill the pod, use ‘`-f`’ flag with ‘`gdel`’ to force kill the pod.
- A template ‘job.sh’ and CUDA program will be put to every user’s home, as an example.

# Viewing all running pods

- A web interface is setup to view the status of all running pods on the cluster.
- Available at: 10.21.230.1:6277
- Can be viewed from any web browser.

# Thank you. Questions?