# Deployment Plan

Soon after the creation of k8s cluster we need to deploy the nvidia-gpu operator using helm or argocd.
The nvidia GPU-operator consists of below components

1. Gpu-operator validator
2. Gpu-operator
3. Nvidia-driver
4. Driver manager
5. Container toolkit
6. Device plugin
7. Dcgm exporter
8. Node feature discovery
9. Mig manager
10. cuda-validator.

The concept of device plugin in k8s helps in access to hardware resources
The operator manages and automates the management o all nvidia software components

We must install nvidia driver also using containers only, the reason is when we install drivers on bare metal then whenever there is patching for os then some of driver components will update and the driver finds some mismatch with other components, so its better to go with containerisation.

Container toolkit will help containers to access gpu's using the nvidia.com/gpu component in deployment manifests, dcgm exporter helps in exporting the metrics of gpu's helps in monitoring and managing the workload.

When it comes to resource allocation we have requests and limits in k8s, where we can set these based on our workload and our requirements. In Nvidia gpus which are higher than A100, there is one additional functionality which helps in minimising the resources wastage that is MIG multi instance gpu, this helps in creating multiple instances in single gpu based on our workload we can create MIG's with lower capacity for less resource constraint applications and higher capacity for more resource constraint applications, this also helps in performance optimization.

Using the Nvidia-gpu-operator helps in scaling any number of nodes in no time, we just need to take care of adding nodes into the existing cluster, and operators take care scheduling the nvidia containers on these nodes, this helps in easy scaling.