One of your most important decisions when creating a [GKE](https://cloudacademy.com/course/introduction-to-google-kubernetes-engine-gke/introduction/) cluster is deciding whether it will be public or private.  Nodes on a public cluster are assigned both a private and a public IP address.  That means you can directly access the nodes from anywhere on the internet.  This can be useful if your cluster is running microservices that need to be accessed by systems running outside of your [GCP](https://cloudacademy.com/library/google/) account.  But it is not as secure.

Nodes on a private cluster only are assigned private IPs.  This means they are isolated from both inbound and outbound communication.  Clients on the internet cannot directly connect to the nodes (but you can still expose them via Services).  And the nodes cannot connect to the internet (unless you configure something like [Cloud NAT](https://cloudacademy.com/course/implementing-a-gcp-virtual-private-cloud-1224/introduction-to-cloud-nat/)).  This provides much better security, at the cost of connectivity.

You should also be aware that both public clusters and private clusters have a public endpoint for the control plane.  You can use this endpoint to make changes to the cluster and do basic management.  For example, this is what the command kubectl connects to.

Private clusters also have a private endpoint as well as a public one.  This is because, on a private cluster, you can choose to disable the public endpoint.  If you do this, the private endpoint is still there to allow your Nodes and Pods to communicate.  It also allows GCP to continue to manage the cluster.  So disabling the public endpoint means you won’t be able to make changes to the cluster outside of GCP.  This is definitely more secure, but also a lot more limiting.

The best configuration depends upon your exact needs.  The most secure would involve a private cluster with the public endpoint disabled.  However, this means that you could not directly connect to the nodes.  And in order to manage the cluster, you are basically limited to using the GCP tools.

Now there is another option that allows you to keep your public endpoint, but control the network range that connections are allowed from.  You can use this option to restrict access by configuring an “authorized network”.  This means only connections from a pre-approved network can connect to the public endpoint.  So you can still remotely administer your cluster, without opening things up too much.

If you choose to disable authorized networks, this will allow you to manage your cluster from anywhere.  Now be aware, this will not mean everyone has open access to your cluster.  Administrators still have to provide authentication credentials.  However, because you are removing a layer of protection, this configuration is a little more risky.

Alright, to summarize, a public cluster means your nodes and control plane are accessible from anywhere on the internet.  You can add authorized networks to limit who can manage your cluster.  But remember this does not affect the connection to your nodes.  A private cluster means your nodes are not directly accessible from the internet.  You can choose to disable the public endpoint, which means you also cannot directly access the control plane from the internet as well.  Or you can choose to leave the public control plane enabled, and then manage the cluster from anywhere.  Or you can leave the public control plane enabled, add an authorized network, and this will allow you to manage your cluster from certain locations on the internet.

Ok, enough theory.  Now let me show how to create a Public and Private cluster.  Once you are logged into the GCP console you need to navigate over to the Google Kubernetes Engine page.  And then click on the “Create” button.  Your first choice is between Standard and Autopilot.  This decision does not limit your public or private options in any way, so you can choose either.

First, I am going to set up an Autopilot cluster, because it is easier.  With Autopilot, you will see the Public / Private option right away.  By default, the cluster is configured to be Public.  So if I wanted, I could get a public cluster by just clicking on the Create button right now.  This would mean that I could directly connect to my nodes and the control plane.  But it’s the least secure option available.

If I want to make things a little more secure, I can enable authorized networks under “Network Options” here.  This will let me specify which networks are allowed to connect to the control plane.  You do this by clicking on “Add Authorized Network”.  And then enter the name and the CIDR range.  And you can add more networks if needed by clicking here.  So if you had four data centers, you might want to add four authorized networks.  Remember this will only limit connections to your control plane.  It still allows anyone on the internet to connect to your nodes.

So those are your public options for an Autopilot cluster.  Next, let me show you the private options.  Now, remember, a private cluster will completely isolate your cluster nodes.   This will prevent your nodes from getting a public IP address.  By default, GKE will still enable the public endpoint here.  So while you can’t directly connect to your nodes, it can still connect to the control plane.  To increase security, you can add an authorized network just like we did for a public cluster.  But if you decide that you want to completely shut off all direct access to the control plane, you can take things one step further and disable the public endpoint by clicking here.  Once you do this, remember the only way to manage the cluster is through GCP.  So this would be your most restrictive, but most secure option.

Ok, now I am going to show you how to do the same thing in a Standard cluster.  Here the public / private option is a little harder to find.  You have to click on the “Networking” tab here.  And just like Autopilot, the default selected option is Public.  You can enable authorized networks if you scroll down to the bottom of the page and click here.  And you add networks in the exact same way as we saw before.

Creating a Private Standard cluster is very similar.  Create the cluster, click on Networking and choose “Private”.  Now this will protect your nodes.  However, the public endpoint is enabled by default, so you can still connect to the control plane.  If you want, you can disable that by clicking here.

You will also notice that you get an error message when setting the cluster to private.  That is because, unlike Autopilot, Standard requires that you specify a control plane IP range.  So make sure you enter that before proceeding.

And if you scroll down, you can add authorized networks just like we did before.  So that is how you create both a public and private GKE cluster.

Now, if you do create a private cluster, then be aware that in order for it to access other Google APIs you will need to make sure that Private Google Access is enabled.  For example, if you want to pull container images from Artifact Registry or send logs to Cloud Logging.  Private Access is normally automatically enabled, except when you are using a Shared VPC.  In that case, you have to enable it manually.  So, if you ever find that you are having issues deploying containers or getting empty logs, you might need to verify that Private Google Access has been enabled for your Private cluster.