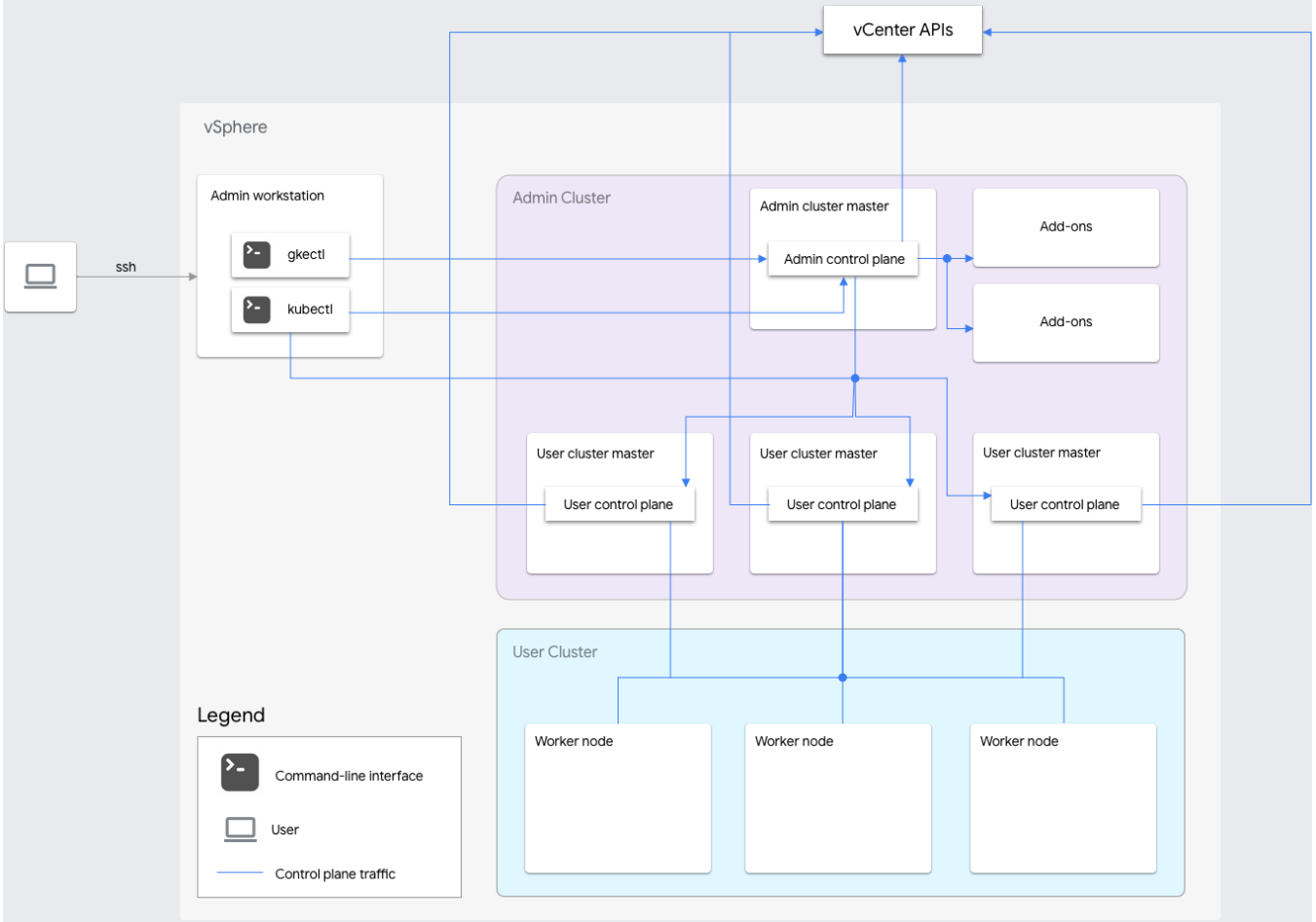
This page describes your high availability (HA) (https://en.wikipedia.org/wiki/High_availability) options in GKE on-prem.

oduct or feature is in a pre-release state and might change or have limited support. For more information, see the pr stages (/products/#product-launch-stages).



(/gke-on-prem/images/architecture_ha_09302019.png)

GKE on-prem architecture with highly-available user clusters (Click to enlarge)

GKE on-prem supports HA user control planes. During cluster creation, you can choose to create three user control planes (/gke-on-prem/docs/concepts/overview#admin_cluster). To do so, specify

`usercluster.masternode.replicas: 3` in the GKE on-prem configuration file you're using to create the user cluster.

To create a HA user cluster, GKE on-prem creates three user control plane VMs (or "master nodes") within the admin cluster. Each control plane VM runs the same Kubernetes control plane components.

During this feature's beta phase, DRS anti-affinity rules are not created for user cluster control plane VMs. You can ex ature in a future release. Until that time, you should enable vSphere High Availability.

GKE on-prem does not currently support HA admin control planes. You can prevent a single point of failure in the admin cluster by enabling vSphere High Availability (https://www.vmware.com/products/vsphere/high-availability.html) for your vSphere cluster.

vSphere High Availability protects GKE on-prem admin clusters from going down in the event of an underlying host failure.

To learn more, see Create a vSphere HA Cluster (https://docs.vmware.com/en/VMware-vSphere/6.5/com.vmware.vsphere.avail.doc/GUID-4BC60283-B638-472F-B1D2-1E4E57EAD213.html)
.

Previous
← Authentication (/gke-on-prem/docs/concepts/authentication)

Next
Logging and monitoring (/gke-on-prem/docs/concepts/logging-and-monitoring) →