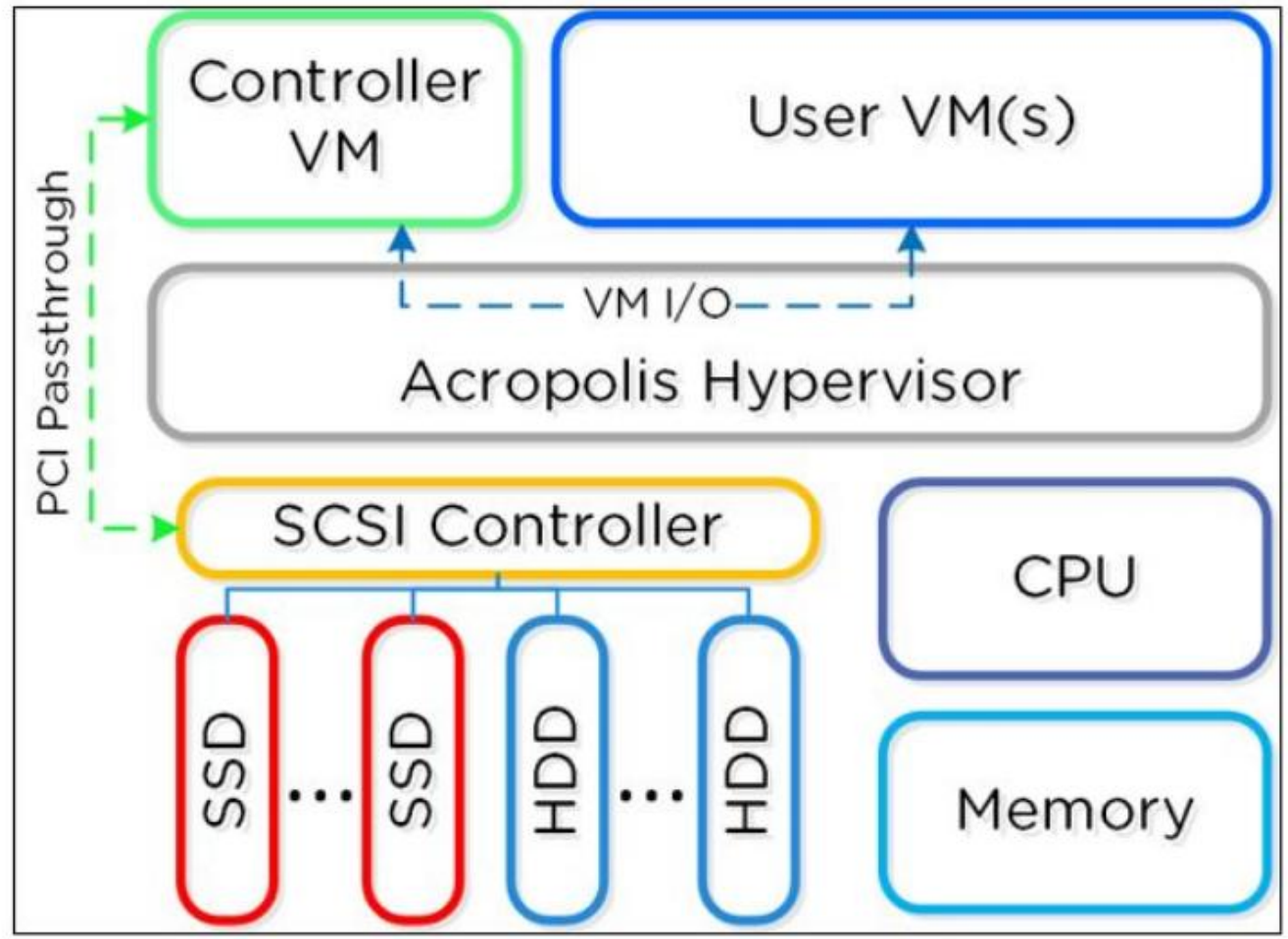
AHV architecture

Overview of AHV

Lenovo

AHV node architecture

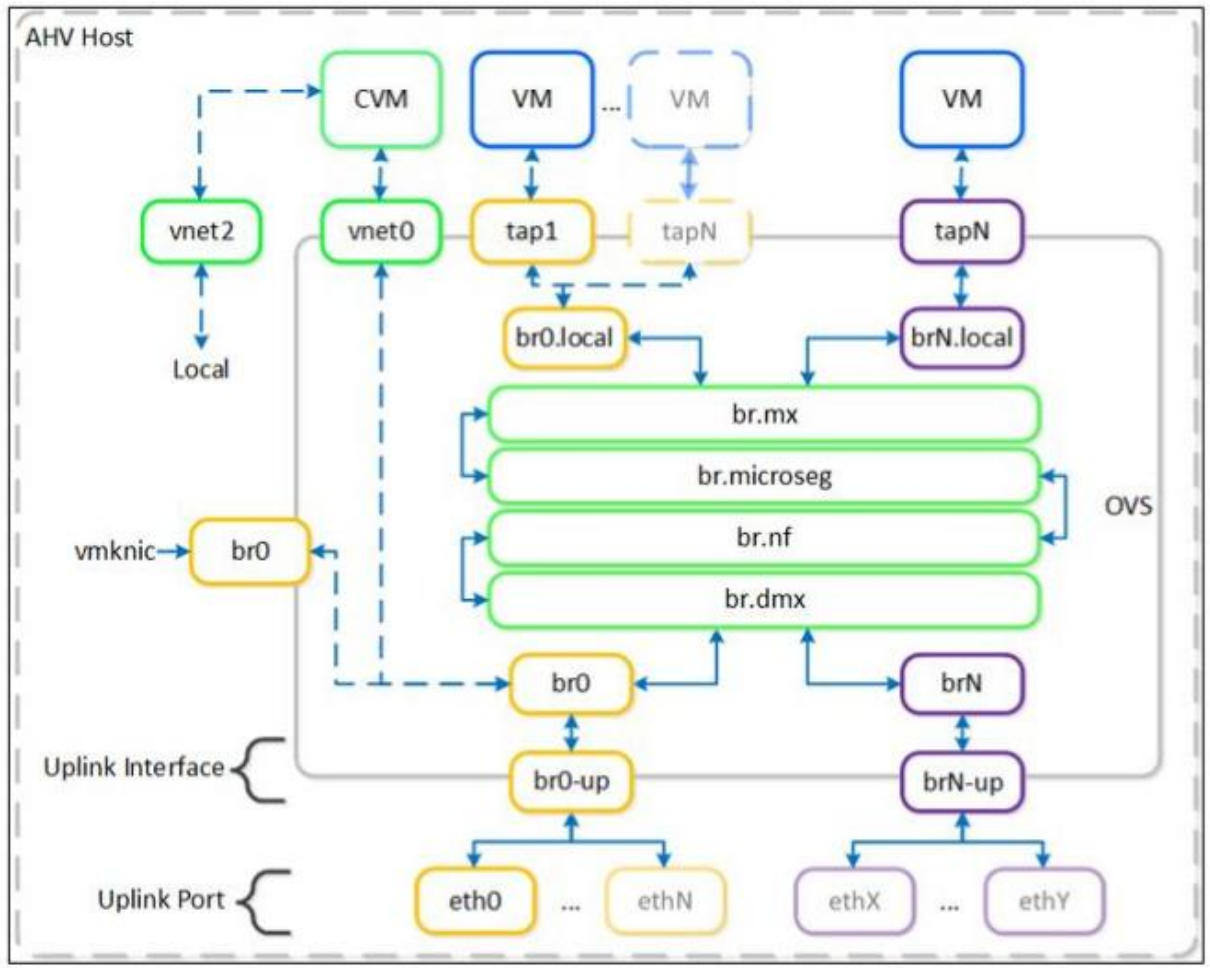
In AHV deployments, the CVM runs as a VM and disks are presented using PCI passthrough. This allows the full PCI controller (and attached devices) to be passed directly through to the CVM and bypass the hypervisor. Full hardware virtualization is used for guest VMs (HVM).
AHV is built on the CentOS KVM foundation and extends its base functionality to include features such as HA and live migration.
AHV is validated as part of the Microsoft Server Virtualization Validation Program and is validated to run Microsoft OS and associated applications.



Networking architecture

AHV leverages Open vSwitch (OVS) for all VM networking. VM networking is configured through Prism / ACLI, and each VM NIC is connected to a tap interface.

The diagram shows a conceptual view of the OVS architecture.



Click the dots for more information.

Networking architecture

Open vSwitch (OVS)

OVS is an open source software switch implemented in the Linux kernel and designed to work in a multiserver virtualization environment. By default, OVS behaves like a layer-2 learning switch that maintains a MAC address table. The hypervisor host and VMs connect to virtual ports on the switch. OVS supports many popular switch features, including VLAN tagging, Link Aggregation Control Protocol (LACP), port mirroring, and quality of service (QoS). Each AHV server maintains an OVS instance, and all OVS instances combine to form a single logical switch. Constructs called “bridges” manage the switch instances residing on the AHV hosts.

Bridge

Bridges act as virtual switches to manage network traffic between physical and virtual network interfaces. The default AHV configuration includes an OVS bridge called br0 and a native Linux bridge called virbr0. The virbr0 Linux bridge carries management and storage communication between the CVM and AHV host. All other storage, host, and VM network traffic flows through the br0 OVS bridge. The AHV host, VMs, and physical interfaces use “ports” for connectivity to the bridge.

Networking architecture

Port

Ports are logical constructs created in a bridge that represent connectivity to the virtual switch. Nutanix uses several port types, including internal, tap, VXLAN, and bond.
An internal port – with the same name as the default bridge (br0) – provides access for the AHV host.
Tap ports act as bridge connections for virtual NICs presented to VMs.
VXLAN ports are used for the IP address management functionality provided by Acropolis.
Bonded ports provide NIC teaming for the physical interfaces of the AHV host.

Bond

Bonded ports aggregate the physical interfaces on the AHV host. By default, a bond named br0-up is created in bridge br0. After the node imaging process, all interfaces are placed within a single bond, which is a requirement for the foundation imaging process. Changes to the default bond, br0-up, often result in the bond being renamed as bond0. Nutanix recommends maintaining the name br0-up, as it allows users to quickly identify the interface as the bridge br0 uplink.
OVS bonds allow for several load-balancing modes, including active-backup, balance-slb, and balance-tcp. LACP can also be activated for a bond. The “bond_mode” setting is not specified during installation and therefore defaults to active-backup, which is the recommended configuration.

Click the dots for more information. →

How AHV storage I/O works

AHV does not leverage a traditional storage stack like ESXi or Hyper-V. All disks are passed to VMs as raw SCSI block devices. This keeps the I/O path lightweight and optimized.
Each AHV host runs an iSCSI redirector which uses NOP commands to regularly check the health of **Stargate** throughout the cluster.
In the iscsi_redirector log (located in /var/log/ on the AHV host), you can check the health of Stargate.

```
2017-08-18 19:25:21,733 - INFO - Portal 192.168.5.254:3261 is up
...
2017-08-18 19:25:25,735 - INFO - Portal 10.3.140.158:3261 is up
2017-08-18 19:25:26,737 - INFO - Portal 10.3.140.153:3261 is up
```

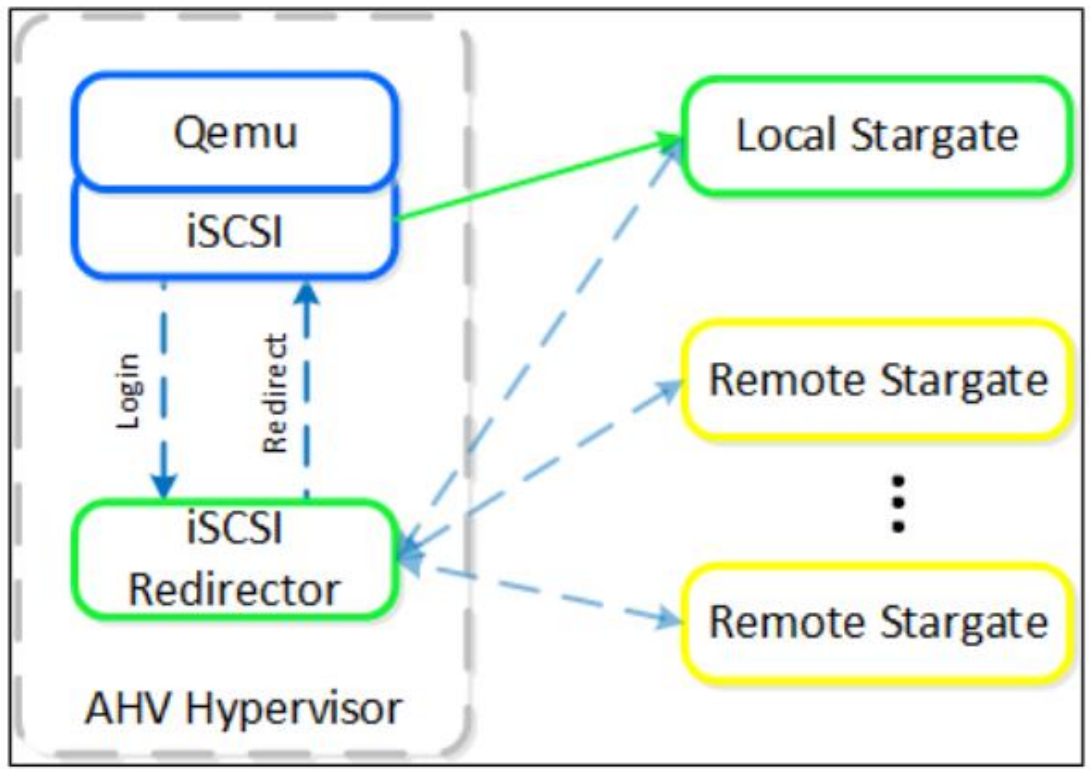
Note: The local Stargate is shown via its 192.168.5.254 internal address.

The iscsi_redirector is listening on 127.0.0.1:3261.

```
[root@NTNX-BEAST-1 ~]# netstat -tnlp | egrep tcp.*3261
Proto ... Local Address   Foreign Address   State     PID/Program name
...
tcp   ... 127.0.0.1:3261  0.0.0.0:*         LISTEN    8044/python
...
```

How AHV storage I/O works

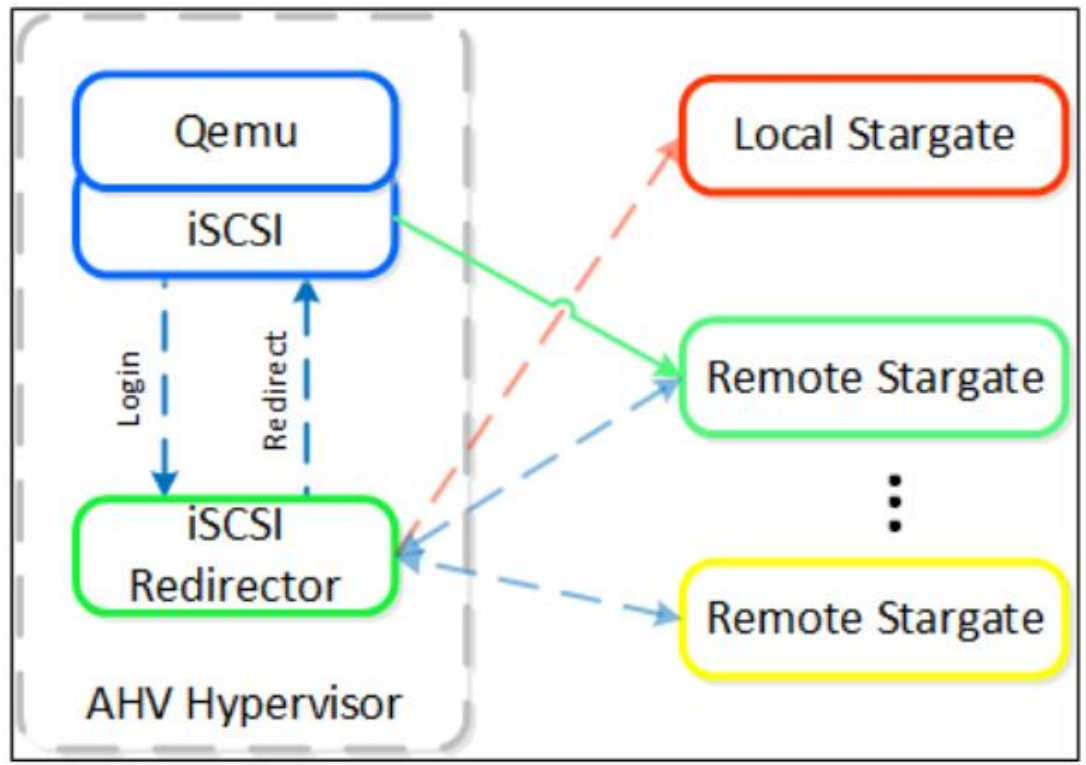
QEMU is configured with the iSCSI redirector as the iSCSI target portal. After a login request, the redirector will perform an iSCSI login redirect to a healthy Stargate (preferably the local one). The diagram shows iSCSI Multi-pathing in Normal State.



Click the dots for more information.

How AHV storage I/O works

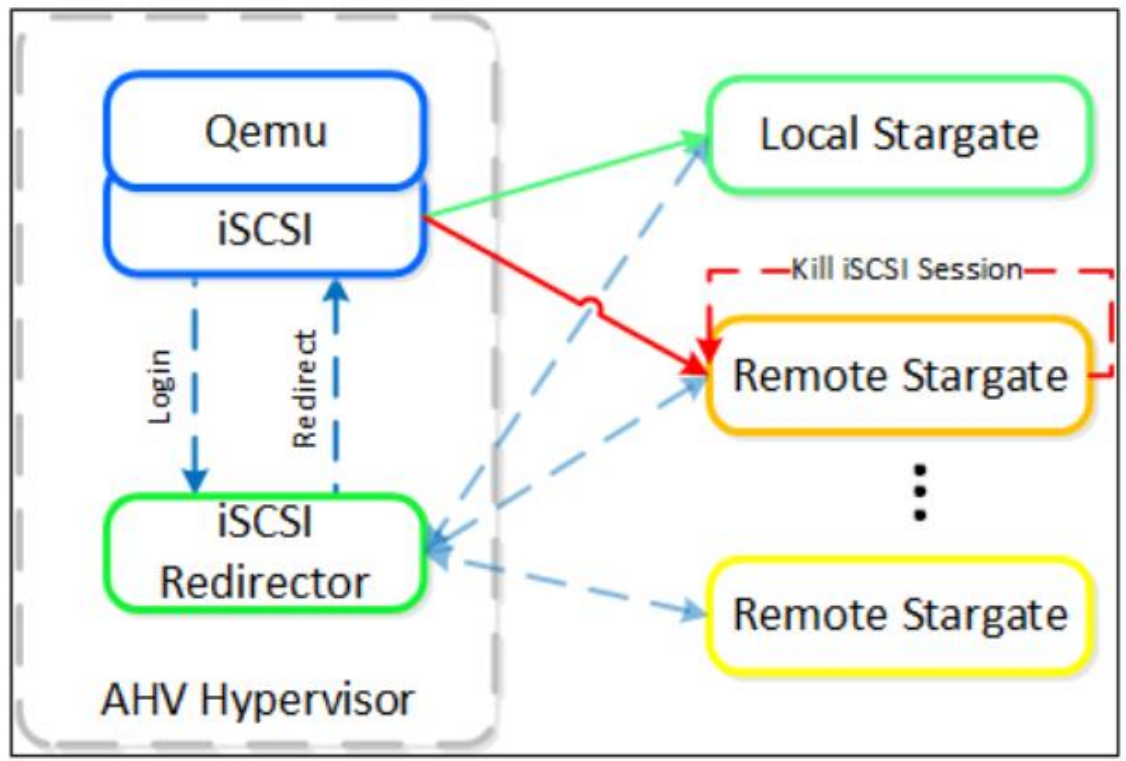
In the event that the active Stargate goes down (thus failing to respond to the NOP OUT command), the iSCSI redirector will mark the local Stargate as unhealthy. When QEMU retries the iSCSI login, the redirector will redirect the login to another healthy Stargate. The diagram shows an overview of iSCSI Multi-pathing when the local CVM goes down.



Click the dots for more information.

How AHV storage I/O works

When the local CVM's Stargate comes back up (and begins responding to the NOP OUT commands), the remote Stargate will aquiesce and then kill all connections to remote iSCSI sessions. QEMU will then attempt an iSCSI login again and will be redirected to the local Stargate. The diagram shows an overview of iSCSI Multi-pathing in the local CVM backup state.

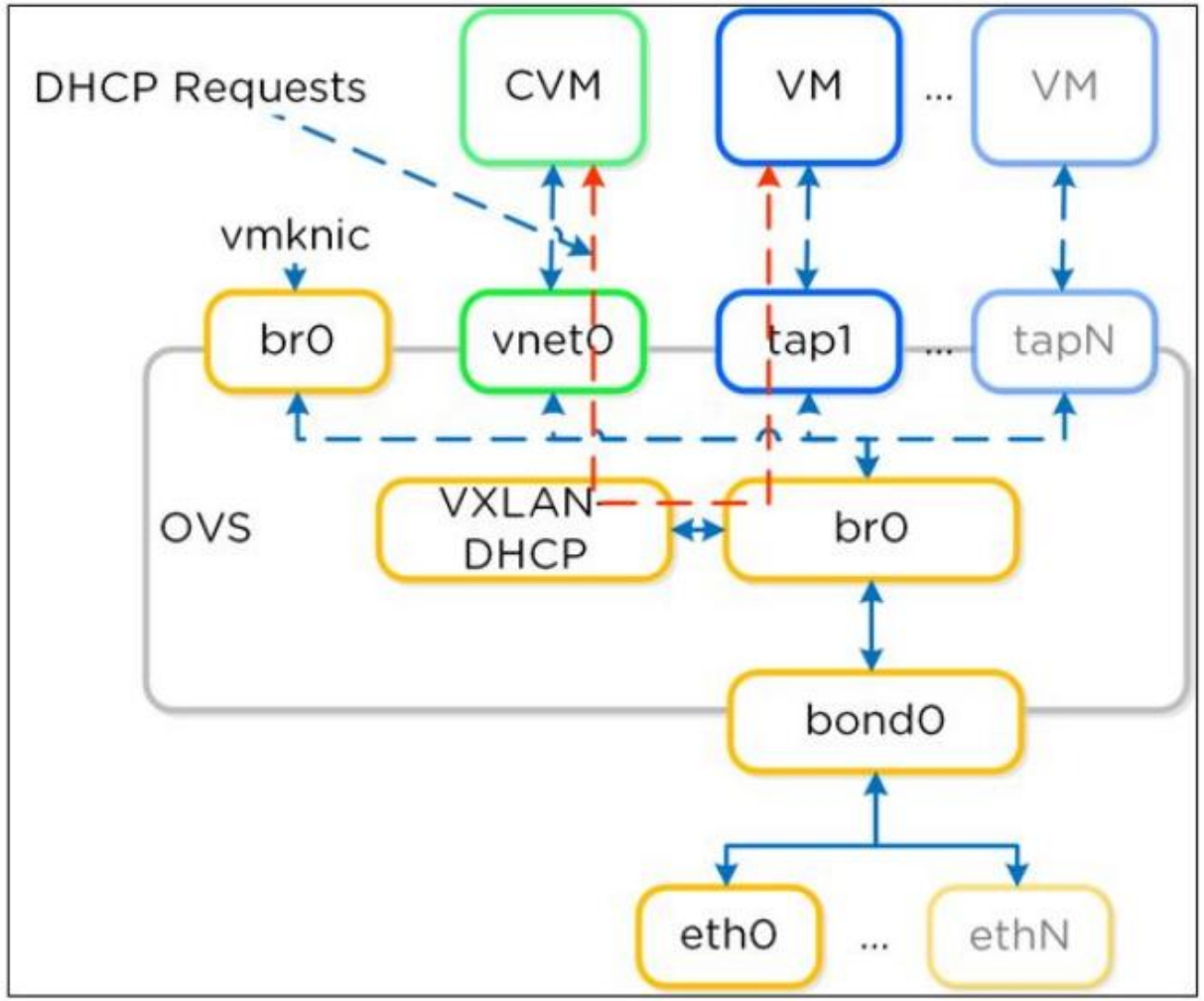


Click the dots for more information.

AHV networking

The Acropolis IP address management (IPAM) solution gives users the ability to establish a DHCP scope and assign addresses to VMs. This leverages VXLAN and OpenFlow rules to intercept the DHCP request and respond with a DHCP response.

The diagram shows an example of a DHCP request using the Nutanix IPAM solution where the Acropolis Master is running locally.



AHV remote networking

If the Acropolis Master is running remotely, the same VXLAN tunnel will be leveraged to handle the request over the network.