AHV - How Nutanix AHV Works

[PDF generated October 31 2024. For all recent updates please see the Nutanix Bible releases notes located at https://nutanixbible.com/release_notes.html. Disclaimer: Downloaded PDFs may not always contain the latest information.]

Storage I/O Path

AHV does not leverage a traditional storage stack like ESXi or Hyper-V. All disk(s) are passed to the VM(s) as raw SCSI block devices. This keeps the I/O path lightweight and optimized.

Note

AOS abstracts kvm, virsh, qemu, libvirt, and iSCSI from the end-user and handles all backend configuration. This allows the user to focus higher up the stack on the VMs via Prism / ACLI. The following is for informational purposes only and it is not recommended to manually mess with virsh, libvirt etc.

Each AHV host runs an iSCSI redirector which regularly checks the health of Stargates throughout the cluster using NOP commands.

In the iscsi_redirector log (located in /var/log/ on the AHV host), you can see each Stargate's health:

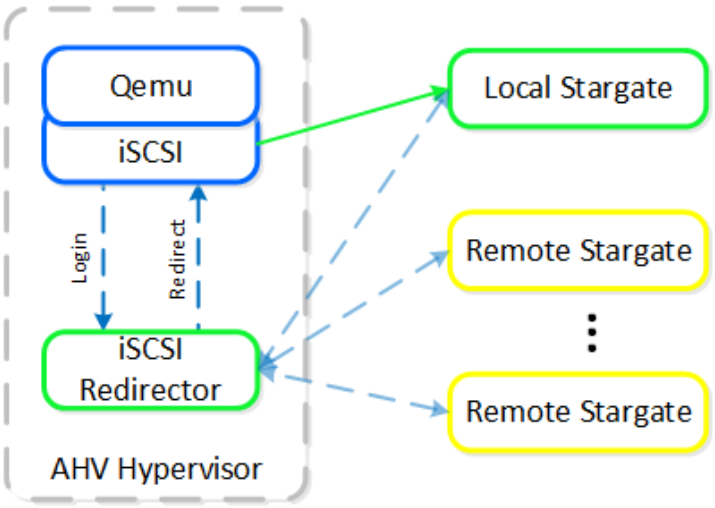
```
2017-08-18 19:25:21,733 - INFO - Portal 192.168.5.254:3261 is up
...
2017-08-18 19:25:25,735 - INFO - Portal 10.3.140.158:3261 is up
2017-08-18 19:25:26,737 - INFO - Portal 10.3.140.153:3261 is up
```

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

In the following you can see 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
...
```

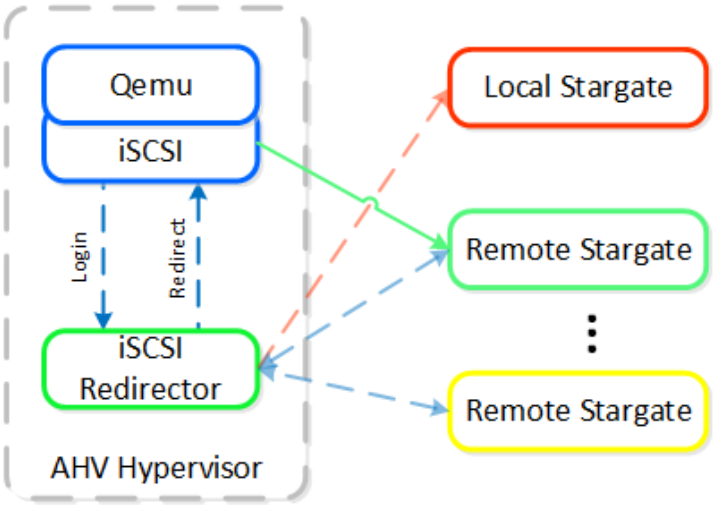
QEMU is configured with the iSCSI redirector as the iSCSI target portal. Upon a login request, the redirector will perform an iSCSI login redirect to a healthy Stargate (preferably the local one).



iSCSI Multi-pathing - Normal State

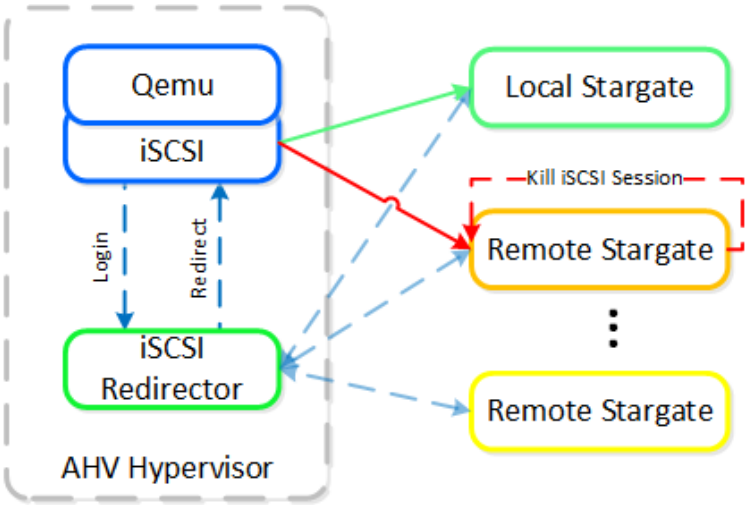
The preferred controller type is virtio-scsi (default for SCSI devices). IDE devices, while possible, are not recommended for most scenarios. In order for virtio to be used with Windows the virtio drivers, Nutanix mobility drivers, or Nutanix guest tools must be installed. Modern Linux distros ship with virtio pre-installed.

In the event where 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.



iSCSI Multi-pathing - Local CVM Down

Once the local CVM's Stargate comes back up (and begins responding to the NOP OUT commands), the remote Stargate will quiesce then kill all connections to remote iSCSI sessions. QEMU will then attempt an iSCSI login again and will be redirected to the local Stargate.



iSCSI Multi-pathing - Local CVM Back Up

Traditional I/O Path

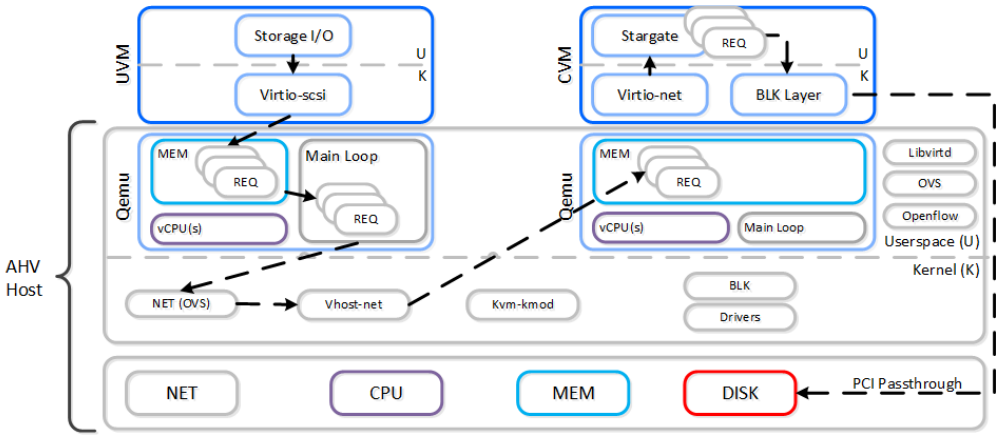
Like every hypervisor and OS there is a mix of user and kernel space components which interact to perform a common activity. Prior to reading the following, it is recommended to read the 'User vs. Kernel Space' section to learn more about how each interact with eachother.

When a VM performs an I/O it will perform the following (some steps have been excluded for clarity):

1. VM's OS perform SCSI command(s) to virtual device(s)

2. Virtio-scsi takes those requests and places them in the guest's memory
3. Requests are handled by the QEMU main loop
4. Libiscsi inspects each request and forwards
5. Network layer forwards requests to local CVM (or externally if local is unavailable)
6. Stargate handles request(s)

The following shows this sample flow:



AHV VirtIO Data Path - Classic

Looking at an AHV host, you can see qemu-kvm has established sessions with a healthy Stargate using the local bridge and IPs. For external communication, the external host and Stargate IPs will be used. NOTE: There will be one session per disk device (look at PID 24845)

```
[root@NTNX-BEAST-1 log]# netstat -np | egrep tcp.*qemu
Proto ... Local Address Foreign Address State PID/Program name
tcp ... 192.168.5.1:50410 192.168.5.254:3261 ESTABLISHED 25293/qemu-kvm
tcp ... 192.168.5.1:50434 192.168.5.254:3261 ESTABLISHED 23198/qemu-kvm
tcp ... 192.168.5.1:50464 192.168.5.254:3261 ESTABLISHED 24845/qemu-kvm
tcp ... 192.168.5.1:50465 192.168.5.254:3261 ESTABLISHED 24845/qemu-kvm
...
```

Now in this path there are a few inefficiencies as the main loop is single threaded and libiscsi inspects every SCSI command.

Frodo I/O Path (aka AHV Turbo Mode)

As storage technologies continue to evolve and become more efficient, so must we. Given the fact that we fully control AHV and the Nutanix stack this was an area of opportunity.

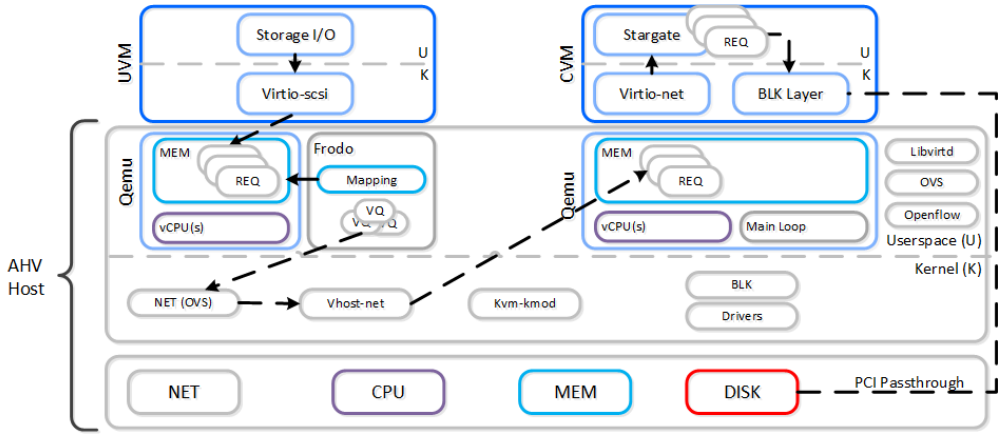
In short Frodo is a heavily optimized I/O path for AHV that allows for higher throughput, lower latency and less CPU overhead.

> **Pro tip**
>
> Frodo is enabled by default on VMs powered on after AOS 5.5.X.

When a VM performs an I/O it will perform the following (some steps have been excluded for clarity):

1. VM's OS perform SCSI command(s) to virtual device(s)
2. Virtio-scsi takes those requests and places them in the guest's memory
3. **Requests are handled by Frodo**
4. **Custom libiscsi appends iscsi header and forwards**
5. Network layer forwards requests to local CVM (or externally if local is unavailable)
6. Stargate handles request(s)

The following shows this sample flow:



AHV VirtIO Data Path - Frodo

The following path does looks similar to the traditional I/O except for a few key differences:

- Qemu main loop is replaced by Frodo (vhost-user-scsi)
- Frodo exposes multiple virtual queues (VQs) to the guest (one per vCPU)
- Leverages multiple threads for multi-vCPU VMs
- Libiscsi is replaced by our own much more lightweight version

To the guest it will notice that it now has multiple queues for the disk device(s), other than that it'll just see the performance improvements. In some cases we've seen a CPU overhead reduction of 25% to perform the I/O and performance increases of up to 3x compared to Qemu! Comparing to another hypervisor we've seen CPU overhead to perform I/Os drop by up to 3x.

Looking at an AHV host, you will see a frodo process for each VM (qemu-kvm process) running:

```
[root@drt-itppc03-1 ~]# ps aux | egrep frodo
... /usr/libexec/qemu-kvm ... -chardev socket,id=frodo0,fd=3 \
 -device vhost-user-scsi-pci,chardev=frodo0,num_queues=16...

... /usr/libexec/frodo ... 127.0.0.1:3261 -t iqn.2010-06.com.nutanix:vmdisk...
...
```

Pro tip

To take advantage of Frodo's multiple threads / connections, you must have >= 2 vCPU for a VM when it is powered on.

It can be characterized by the following:

- 1 vCPU UVM:
  - 1 Frodo thread / session per disk device
- >= 2 vCPU UVM:
  - 2 Frodo threads / sessions per disk device

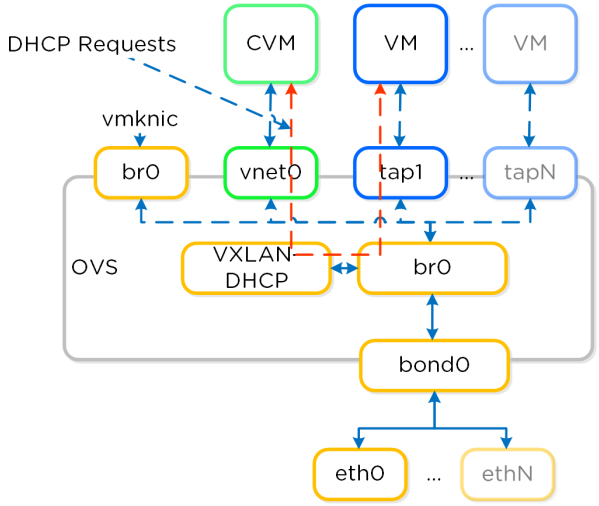
In the following, you can see Frodo has established sessions with a healthy Stargate using the local bridge and IPs. For external communication, the external host and Stargate IPs will be used.

```
[root@NTNX-BEAST-1 log]# netstat -np | egrep tcp.*frodo
Proto ... Local Address Foreign Address State PID/Program name
tcp ... 192.168.5.1:39568 192.168.5.254:3261 ESTABLISHED 42957/frodo
tcp ... 192.168.5.1:39538 192.168.5.254:3261 ESTABLISHED 42957/frodo
tcp ... 192.168.5.1:39580 192.168.5.254:3261 ESTABLISHED 42957/frodo
tcp ... 192.168.5.1:39592 192.168.5.254:3261 ESTABLISHED 42957/frodo
...
```

IP Address Management

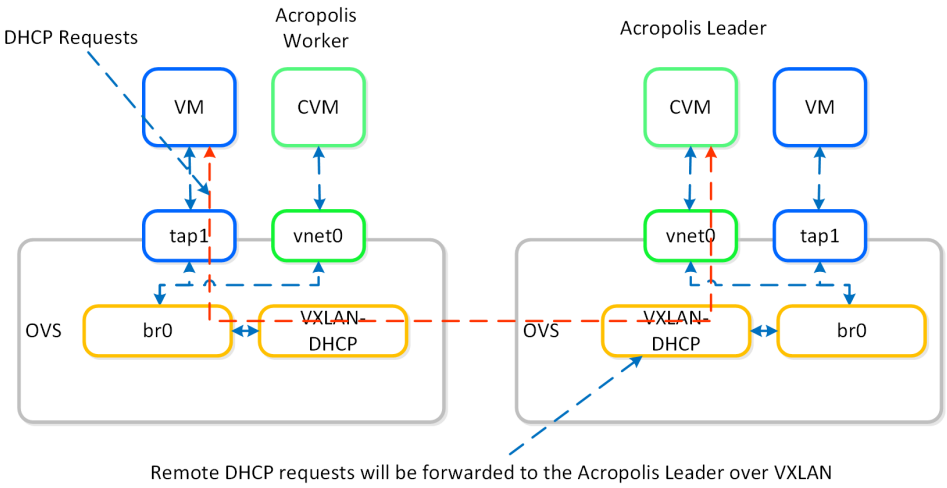
The Acropolis IP address management (IPAM) solution provides the ability to establish a DHCP scope and assign addresses to VMs. This leverages VXLAN and OpenFlow rules to intercept the DHCP request and respond with a DHCP response.

Here we show an example DHCP request using the Nutanix IPAM solution where the Acropolis Leader is running locally:



IPAM - Local Acropolis Leader

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



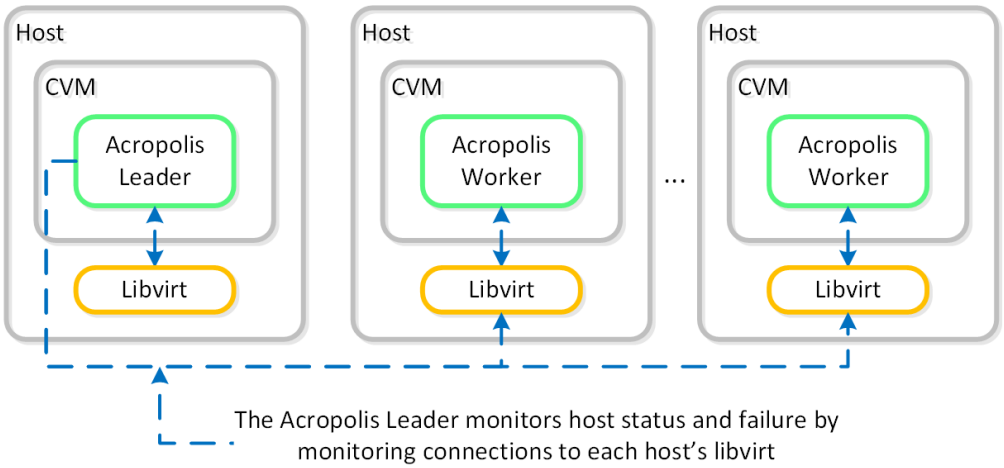
IPAM - Remote Acropolis Leader

Traditional DHCP / IPAM solutions can also be leveraged in an ‘unmanaged’ network scenario.

VM High Availability (HA)

AHV VM HA is a feature built to ensure VM availability in the event of a host or block outage. In the event of a host failure the VMs previously running on that host will be restarted on other healthy nodes throughout the cluster. The Acropolis Leader is responsible for restarting the VM(s) on the healthy host(s).

The Acropolis Leader tracks host health by monitoring its connections to the libvirt on all cluster hosts:



HA - Host Monitoring

Once the libvirt connection goes down, the countdown to the HA restart is initiated. Should libvirt connection fail to be re-established within the timeout, Acropolis will restart VMs that were running on the disconnected host. When this occurs, VMs should be restarted within 120 seconds.

In the event the Acropolis Leader becomes partitioned, isolated or fails a new Acropolis Leader will be elected on the healthy portion of the cluster. If a cluster becomes partitioned (e.g X nodes can't talk to the other Y nodes) the side with quorum will remain up and VM(s) will be restarted on those hosts.

There are two main modes for VM HA:

- Default
  - This mode requires no configuration and is included by default when installing an AHV-based Nutanix cluster. When an AHV host becomes unavailable, the VMs that were running on the failed AHV host restart on the remaining hosts, depending on the available resources. Not all of the failed VMs restart if the remaining hosts do not have sufficient resources.

- Guarantee
  - This nondefault configuration reserves space throughout the AHV hosts in the cluster to guarantee that all failed VMs can restart on other hosts in the AHV cluster during a host failure. To enable Guarantee mode, select the Enable HA check box, as shown in the figure below. A message then appears displaying the amount of memory reserved and how many AHV host failures can be tolerated.

Resource Reservations

When using the Guarantee mode for VM HA, the system will reserve host resources for VMs. The amount of resources which are reserved is summarized by the following:

- If all containers are RF2 (FT1)
  - One "host" worth of resources

- If any containers are RF3 (FT2)
  - Two "hosts" worth of resources

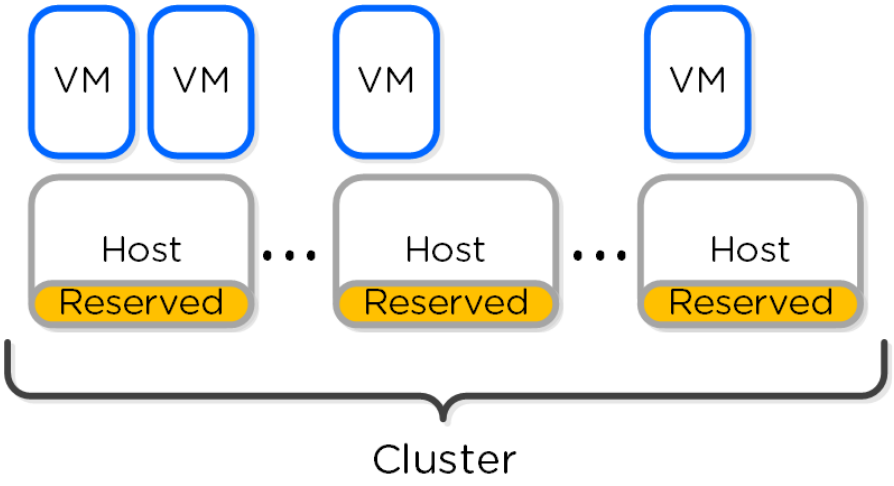
When hosts have uneven memory capacities the system will use the largest host's memory capacity when determining how much to reserve per host.

Post 5.0 Resource Reservations

Prior to 5.0, we supported both host and segment based reserevations. With 5.0 and later we now only support a segment based reservation which is automatically implemented when the Guarantee HA mode is selected.

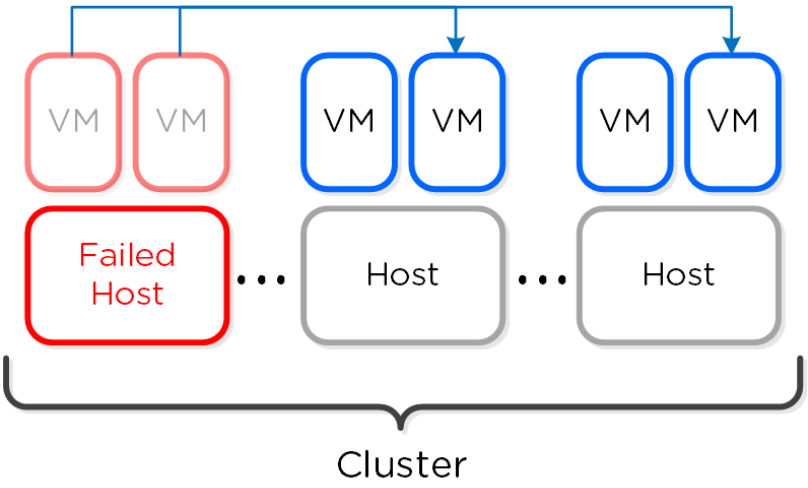
Reserve segments distributes the resource reservation across all hosts in a cluster. In this scenario, each host will share a portion of the reservation for HA. This ensures the overall cluster has enough failover capacity to restart VM(s) in the event of a host failure.

The figure shows an example scenario with reserved segments:



HA - Reserved Segment

In the event of a host failure VM(s) will be restarted throughout the cluster on the remaining healthy hosts:



HA - Reserved Segment - Fail Over

Reserved segment(s) calculation

The system will automatically calculate the total number of reserved segments and per host reservation.

Finding reservations reduces to a well known set of problems called Knapsack. The optimal solution is NP-hard (exponential), but heuristic solutions can come close to optimal for the common case. We implement one such algorithm called MTHM. Nutanix will continue improving its placement algorithms.