High Availability

https://campus.barracuda.com/doc/92766210/

High Availability aims to ensure that the services of Barracuda CloudGen WAN are available even if one unit is unavailable due to maintenance or a hardware defect. Although the Barracuda CloudGen WAN gateway is automatically deployed in a redundant high availability cluster, you must manually configure High Availability on the site appliances.

Barracuda CloudGen WAN Gateway

High Availability is automatically available after deployment without further configuration since the gateway operates with redundant virtual machines in the cloud. The virtual machines work in a failover and failback configuration, which means that if a virtual machine becomes unavailable, another virtual machine takes over (failover), and as soon as the virtual machine is available again, it resumes working (failback) so that the load is balanced on all available virtual machines. Failover and failback are fully automated; there is no configuration necessary.

Barracuda CloudGen WAN Site Appliance

For redundancy and reliability, you can set up two Barracuda CloudGen WAN appliances in a high availability cluster. The high availability cluster works for both virtual and hardware appliances in the same way. During normal operations, the primary unit is active while the secondary unit waits in standby mode. The secondary unit has the same configurations as the primary unit and only becomes available when the primary unit is down. The failover is reversed when the primary unit can resume operations. Failover and failback are executed automatically, but y ou can also manually execute a failover. For more information, see How to Trigger a Failover of the Site Appliance. The passive appliance always accesses the Internet (e.g., for firmware or configuration updates) through the active appliance.

- Both units must use the same platform. You cannot mix virtual and physical appliances.
- Both units must be the same model.
- It is recommended to use the same firmware version on both models; however, using different firmware versions is possible. For more information, see Migration Notes 8.1.2.
- It is recommended to upgrade both appliances to the newest version.
- Latency on the HA sync connection must not exceed 80 ms.
- The subnet 169.254.128.0/24 is reserved for communication between the two appliances of the high availability cluster . Do not use this subnet anywhere else.

Creating a High Availability Cluster

When installing two appliances in a high availability cluster, ensure that the cabling is done exactly the same on both units. For example, if port 4 on the primary box is connected to ISP 1, the secondary box must also connect port 4 with ISP 1. If you install cabling incorrectly, HA failover does not work properly. Port 1 is always reserved on both virtual and hardware appliances for the direct connection to the other appliance of the high availability cluster. Since on a virtual appliance you cannot connect the two ports directly, connect port 1 of both appliances to a virtual switch where only port 1 of the other appliance is connected to. For an example of correct cabling, see the following diagram:
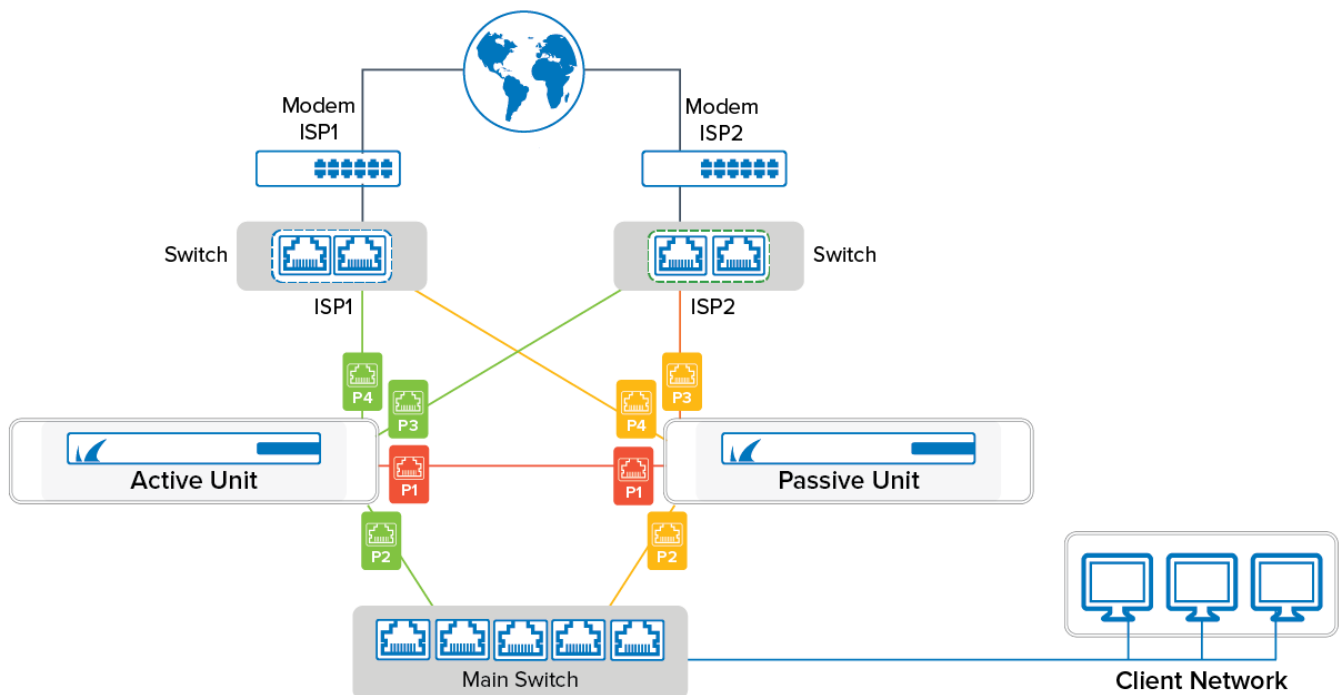
To configure High Availability, simply select two appliances during the deployment while creating the site configuration or when adding a new site. Both appliances must be the same model and have the same firmware version.

In order to downgrade a high availability cluster to a single unit, you must delete the entire site configuration in **SITES** and create a new site configuration.

For more information, see How to Create a Site Configuration in Barracuda CloudGen WAN and How to Enable Interface Monitoring in a High Availability Cluster.

Recommendations for a High Availability Cluster

Reliable High Availability depends on the correct configuration of the surrounding switches and routers. Especially important is the ARP cache time or ARP timeout, which must be set to a value between 30 and 60 seconds. When the primary appliance fails over to the secondary appliance, the MAC addresses change. The MAC address is immediately sent out via gratuitous or unsolicited ARP requests, updating the MAC address table or ARP cache of the connected switches and routers. If the lifetime of the ARP timeout of the switch is set to be longer, for example 300 seconds, the secondary unit would not be reachable for up to 5 minutes because the ARP cache would not be updated for that time frame.

Further Information

- [How to Access the Passive Unit of the Site Appliances](#)
- [How to Trigger a Failover of the Site Appliance](#)
- [How to Create a Site Configuration in Barracuda CloudGen WAN](#)
- [How to Enable Interface Monitoring in a High Availability Cluster](#)

Figures

1. ha_conf_01.png
2. ha_821.png