Book of Nutanix Cloud Clusters - Nutanix Cloud Clusters on Azure

Nutanix Cloud Clusters (NC2) on Azure provides on-demand clusters running in target cloud environments using bare metal resources. This allows for true on-demand capacity with the simplicity of the Nutanix platform you know. Once provisioned the cluster appears like any traditional AHV cluster, just running in a cloud provider's datacenter(s).

Supported Configurations

The solution is applicable to the configurations below (list may be incomplete, refer to documentation for a fully supported list):

Core Use Case(s):

- On-Demand / burst capacity
- Backup / DR
- Cloud Native
- Geo Expansion / DC consolidation
- App migration
- Etc.

Management interfaces(s):

- Nutanix Clusters Portal - Provisioning
- Prism Central (PC) - Nutanix Management
- Azure Portal - Azure Management

Supported Environment(s):

- Cloud:
  - AWS
  - Azure

- Bare Metal Instance Types:
  - AN36
  - AN36P

Upgrades:

- Part of AOS

Compatible Features:

- AOS Features
- Azure Services

Key terms / Constructs

The following key items are used throughout this section and defined in the following:

- Nutanix Clusters Portal
  - The Nutanix Clusters Portal is responsible for handling cluster provisioning requests and interacting with Azure and the provisioned hosts. It creates cluster specific details and handles the cluster creation and helps to remediate hardware problems.

- Region
  - A geographic landmass or area where multiple Availability Zones (sites) are located. A region can have two or more AZs. These can include regions like East US (Virginia) or West US 2 (Washington).

- Availability Zone (AZ)
  - An AZ consists of one or more discrete datacenters interconnected by low latency links. Each site has its own redundant power, cooling, network, etc. Comparing these to a traditional colo or datacenter, these would be considered more resilient as a AZ can consist of multiple independent datacenters.

- VNet
  - A logically isolated segment of the Azure cloud for tenants. Provides a mechanism to secure and isolate environment from others. Can be exposed to the internet or other private network segments (other VNets, or VPNs).

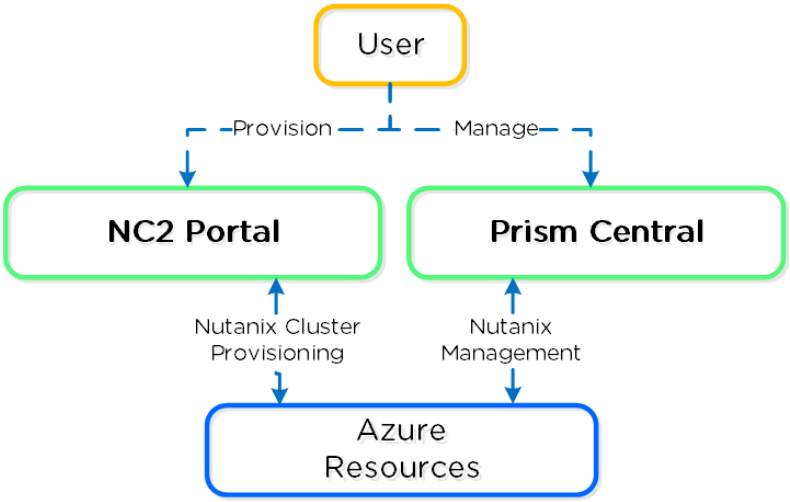
Cluster Architecture

From a high-level the Nutanix Clusters (NC2) Portal is the main interface for provisioning Nutanix Clusters on Azure and interacting with Azure.

The provisioning process can be summarized with the following high-level steps:

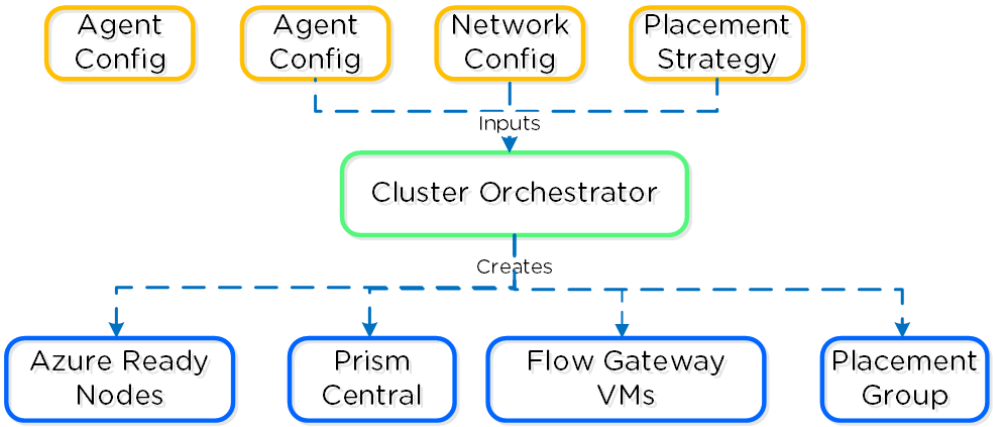
1. Create cluster in NC2 Portal
2. Deployment specific inputs (e.g. Region, AZ, Instance type, VNets/Subnets, etc.)
3. The NC2 Portal creates associated resources
4. Host agent running on AHV checks-in with Nutanix Clusters on Azure
5. Once all hosts as up, cluster is created

The following shows a high-level overview of the NC2 on Azure interaction:



NC2 on Azure - Overview

The following shows a high-level overview of a the inputs taken by the NC2 Portal and some created resources:



Nutanix Clusters on Azure - Cluster Orchestrator Inputs

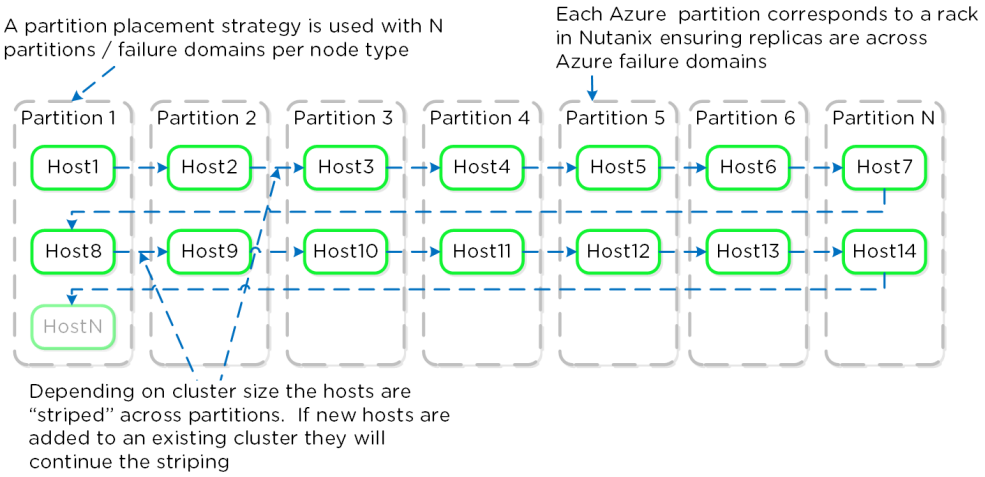
Node Architecture

Given the hosts are bare metal, we have full control over storage and network resources similar to a typical on-premises deployment. We are consuming Ready Nodes as our building blocks. Unlike AWS, Azure-based nodes are not consuming any additional services for the CVM or AHV.

Placement policy

Nutanix Clusters on Azure uses a partition placement policy with 7 partitions by default. Hosts are striped across these partitions which correspond with racks in Nutanix. This ensures you can have 1-2 full "rack" failures and still maintain availability.

The following shows a high-level overview of the partition placement strategy and host striping:



NC2 on Azure - Partition Placement

Storage

Core storage is the exact same as you'd expect on any Nutanix cluster, passing the "local" storage devices to the CVM to be leveraged by Stargate.

Instance Storage

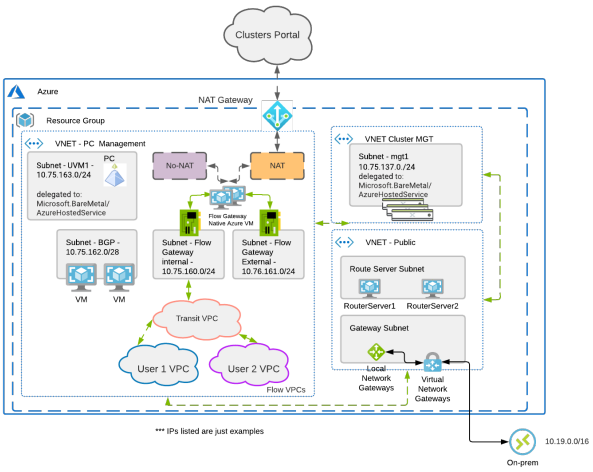
Given that the "local" storage is backed by the local flash, it is fully resilient in the event of a power outage.

Networking

NC2 utilizes Flow Virtual Networking in Azure to create an overlay network to ease administration for Nutanix administrators and reduce networking constraints across Cloud vendors. Flow Virtual Networking is used to abstract the Azure native network by creating overlay virtual networks. On the one hand this abstracts the underlying network in Azure, while at the same time, it allows the network substrate (and its associated features and functionalities) to be consistent with the customer's on-premises Nutanix deployments. You will be able to create new virtual networks (called Virtual Private Clouds or VPCs) within Nutanix, subnets in any address range, including those from the RFC1918 (private) address space and define DHCP, NAT, routing, and security policy right from the familiar Prism Central interface.

Flow Virtual Networking can mask or reduce Cloud constraints by providing an abstraction layer. As an example, Azure only allows for one delegated subnet per VNet. Subnet delegation enables you to designate a specific subnet for an Azure PaaS service of your choice that needs to be injected into your virtual network. NC2 needs a management subnet delegated to the Microsoft.BareMetal/AzureHostedService. Once your subnet is delegated to the BareMetal service the Clusters Portal will be able to use that subnet to deploy your Nutanix Cluster. The AzureHostedService is what the Clusters portal uses to deploy and configure networking on the bare-metal nodes.

Every subnet used for user native VM networking also needs to be delegated to the same service. Since a VNet can only have one delegated subnet, networking configuration would get out of hand with needing to peer VNets among each other to allow communication. With Flow Virtual Networking we can drastically reduce the amount of VNets needed to allow communication of the workloads running on Clusters and Azure. Flow Virtual Networking will allow you to create over 500 subnets while only consuming 1 Azure VNet.



It is recommended to create a new VPC with associated subnets, NAT/Internet Gateways, etc. that fits into your corporate IP scheme. This is important if you ever plan to extend networks between VPCs (VPC peering), or to your existing WAN. I treat this as I would any site on the WAN.

Prism Central (PC) will be deployed onto the Nutanix Cluster after deployment. Prism Central contains the control plane for Flow Virtual Networking. The subnet for PC will be delegated to the Microsoft.BareMetal/AzureHostedService so native Azure networking can be used to distribute IPs for PC. Once PC is deployed, the Flow Gateway will be deployed into the same subnet PC is using. The Flow Gateway allows the User VMs using the Flow VPC(s) to communicate to native Azure services and allows the VMs to have parity with native Azure VMs, such as:

- User defined routes - You can create custom, or user-defined (static), routes in Azure to override Azure's default system routes, or to add additional routes to a subnet's route table. In Azure, you create a route table, then associate the route table to zero or more virtual network subnets.
- Load Balancer Deployment - The ability to front-end services offered by UVMs with Azure-native load balancer.
- Network Security Groups - The ability to write stateful firewall policies.

The Flow Gateway VM is responsible for all VM traffic going north and south bound from the cluster. During deployment you can pick different sizes for the Flow Gateway VM based on how much bandwidth you need. It’s important to realize that CVM replication between other CVMs and on-prem do not flow through the Flow Gateway VM so you don’t have to size for that traffic.

Flow Virtual Networking Gateway VM High Availability

When you initially deploy your first cluster, you’re able to choose how many FGW VMs you want to create (2 - 4). Prior to AOS 6.7, If you deploy only one gateway VM, the NC2 portal redeploys a new FGW VM with an identical configuration when it detects that the original VM is down. Because this process invokes various Azure APIs, it can take about 5 minutes before the new FGW VM is ready to forward traffic, which can affect the north-south traffic flow.

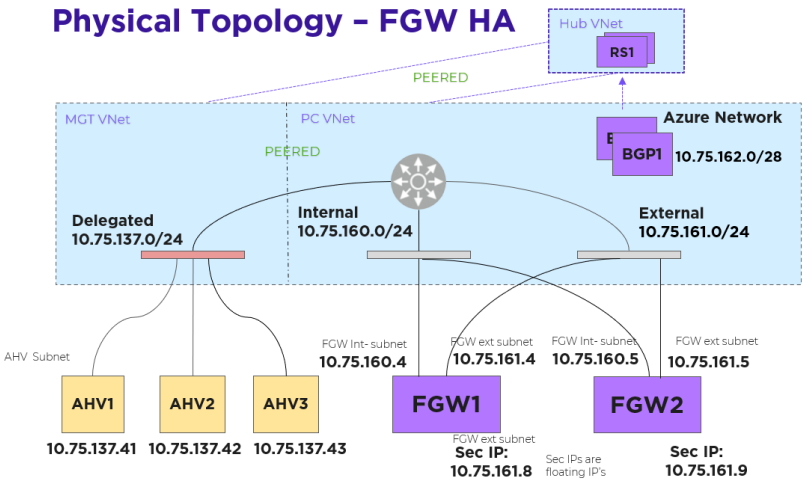
To reduce this downtime, NC2 on Azure has moved to an active-active configuration. This setup provides a flexible scale-out configuration when you need more traffic throughput.

The following workflow describes what happens when you turn off a FGW VM gracefully for planned events like updates.

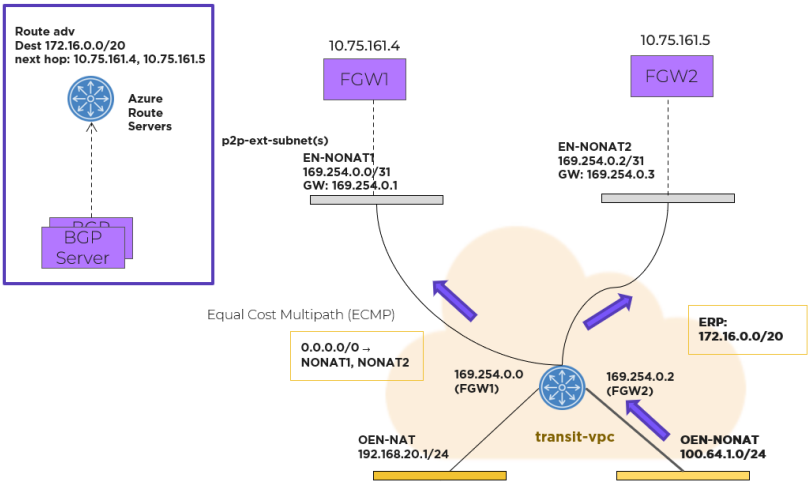
1. The NC2 portal disables the FGW VM.
2. Prism Central removes the VM from the traffic path
3. The NC2 portal deletes the original VM and creates a new FGW VM with an identical configuration.
4. The NC2 portal registers the new instance with Prism Central.
5. Prism Central adds the new instance back to the traffic path.

For ungraceful or unplanned failures, the NC2 portal and Prism Central both have their own detection mechanisms based on keepalives. They take similar actions to those for the graceful or planned cases.

Network Address Translation (NAT): UVMs that want to communicate with AHV/CVM/PC and Azure resources will flow though the external network card on the Flow Gateway VM. The NAT provided uses native Azure address to ensure routing to all resources. User defined routes in Azure can be used to talk directly to Azure resources if using a NAT is not preferred. This allows for fresh installs to communicate with Azure right away but also gives customers options for more advanced configurations.



Each FGW instance has two NICs: one on the internal subnet that exchanges traffic with AHV and another on the external subnet that exchanges traffic with the Azure network. Each FGW instance registers with Prism Central and is added to the traffic path. A point-to-point external subnet is created for each FGW and the transit VPC is attached to it, with the FGW instance hosting the corresponding logical-router gateway port. In the following diagram, EN-NONAT1 and EN-NONAT2 are the point-to-point external subnets.



Flow Virtual Networking Gateway Using the NoNAT Path

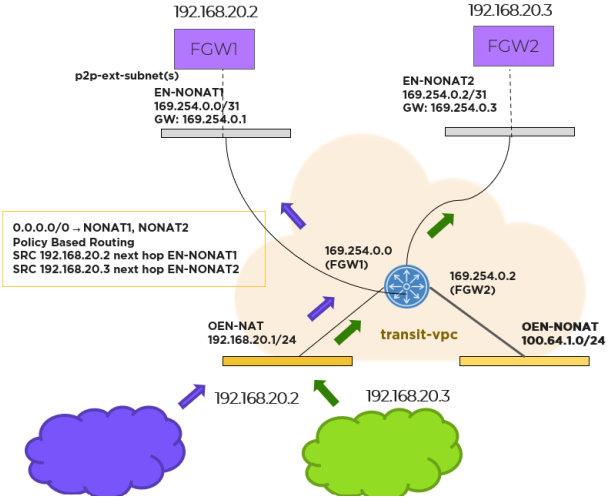
For northbound traffic, the transit VPC has an equal-cost multi-path (ECMP) default route, with all the point-to-point external subnets as possible next hops. In this case, the transit VPC distributes traffic across multiple external subnets hosted on different FGWs.

For southbound traffic using more than one FGW, a Border Gateway Protocol (BGP) gateway is deployed as an Azure native VM instance in the Prism Central Virtual Network (VNet). Azure Route Servers deploy in the same Prism Central VNet. With an Azure Route Server, you can exchange routing information directly through BGP between any network virtual appliance that supports BGP and the Azure VNet without the need to manually configure or maintain route tables.

The BGP gateway peers with the Azure Route Servers. The BGP gateway advertises the externally routable IP addresses to the Azure Route Server with each active FGW external IP address as the next hop. Externally routable IP addresses compose the address range that you've created and want advertised to the rest of the network in your Flow Virtual Networking user VPCs. Once the externally routable IP address is set in Prism Central at the user VPC, the Azure network distributes southbound packets across all the FGW instances.

Prism Central determines which FGW instance should host a given NAT IP address and then configures each NAT IP address as a secondary IP address on each FGW. Packets sourced from those IP addresses can be forwarded through the corresponding FGW only. No NAT traffic distributes across all FGWs.

NAT traffic originating from the Azure network and destined to a floating IP address goes to the FGW VM that owns the IP address because Azure knows which NIC currently has it.

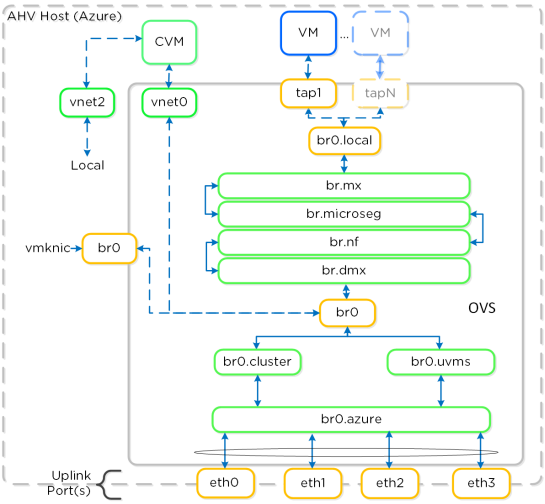


Prism Central uses policy-based routing to support forwarding based on the source IP address matching the floating IP address. The custom forwarding policy-based routing rules built into Flow Virtual Networking are used to auto install routes in the transit VPC.

Host Networking

The hosts running on baremetal in Azure are traditional AHV hosts, and thus leverage the same OVS based network stack.

The following shows a high-level overview of a Azure AHV host's OVS stack:



NC2 on Azure - Host Networking

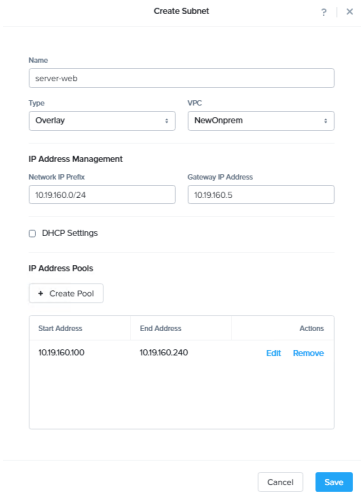
Nutanix's Open vSwitch implementation is very similar to the on-premises implementation. The above diagrams shows an internal architecture of the AHV that is deployed onto the bare-metal. Br0 bridge will split traffic between br0.cluster (AHV/CVM IPs) and br0.uvms(User VMs IPs).

For AHV/CVM traffic via br0.cluster, it will be a simple pass-through to br0.azure bridge, with no modification to data packets. The top of rack switching is providing the security for br0.cluster traffic. For UVM IPs traffic will flow via br0.uvms, OVS rules would be installed for vlan-id translation and pass-through traffic to br0.azure.

br0.azure will have OVS bond br0.azure-up which will form a bonded interface with bare-metal attached physical nics. Thus, br0.azure hides the bonded interface from br0.uvms and br0.cluster.

Creating a Subnet

Subnets you create will have its own built in IPAM and you will have the option to stretch your network from on-prem into Azure. If outside applications need to talk directly your UVM inside the subnet you also have the option to assign floating IPs from a pool of IPs from Azure that will come from the external network of the Flow Gateway.



NC2 on Azure - IPAM with Azure

For a successful deployment, Nutanix Clusters needs outbound access to the NC2 portal, either using an NAT gateway or an on-prem VPN with outbound access. Your Nutanix cluster can sit in a private subnet that can only be accessed from your VPN, limiting exposure to your environment.

WAN / L3 Networking

In most cases deployments will not be just in Azure and will need to communicate with the external world (Other VNets, Internet or WAN).

For connecting VNets (in the same or different regions), you can use VPC peering which allows you to tunnel between VPCs. NOTE: you will need to ensure you follow WAN IP scheme best practices and there are no CIDR range overlaps between VNets / subnets.

For network expansion to on-premises / WAN, either a VNet gateway (tunnel) or Express Route can be leveraged.

Usage and Configuration

The following sections cover how to configure and leverage NC2 on Azure.

The high-level process can be characterized into the following high-level steps:

1. Setup up an active Azure subscription.
2. Create a My Nutanix account & subscribe to NC2.
3. Register Azure resource providers.
4. Create an app registration in Azure AD with “Contributor” access to the new subscription
5. Configure DNS.
6. Create a resource group or re-use an existing resource group.
7. Create required VNets and required subnets.
8. Configure two NAT gateways.
9. Establish the VNet peering required for the Nutanix cluster.
10. Add your Azure account to the NC2 console.
11. Create a Nutanix Cluster in Azure by using the NC2 console.

More to come!