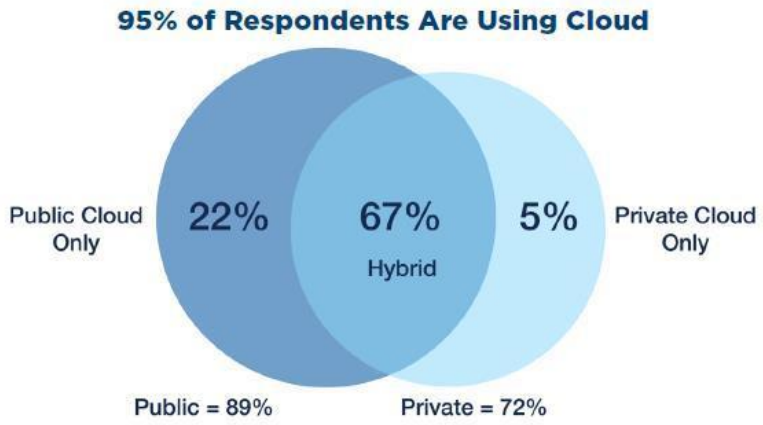
CHECK POINT SECURE CLOUD BLUEPRINT

Agile security architecture for the cloud

Overview

Cloud computing has been widely adopted globally and is expected to grow even further in coming years. Business agility is undoubtedly the main benefit and key driver behind enterprise cloud adoption because IT resources can be acquired and deployed more quickly. Once deployed, these resources can be increased or decreased as needed to meet demand.

According to RightScale's "2017 State of the Cloud Report" (shown in the image below), 95% of the report's respondents indicated that they are using the cloud. Adoption is noticeable across all cloud platforms, both public and private, and what's more organizations utilize multiple vendors while building their hybrid cloud environments.



Source: RightScale 2017 State of the Cloud Report

At the same time, Gartner states that security continues to be the biggest inhibitor of cloud adoption. This is not a surprise considering the fact that cloud services are shared, always connected, and dynamic environments by nature, making them a challenge when it comes to security.

Moving computing resources and data to a public cloud environment means that security responsibilities become shared between you and your cloud provider. While infrastructure protection is delivered by the provider, you want and need the ability to control your own data, keep it private, and protect all of your cloud assets, all while maintaining compliance with regulatory mandates. Check Point CloudGuard IaaS protects applications and data in private and public clouds with advanced threat prevention security while enabling reliable connectivity to public and hybrid cloud environments.

The focus of this document is architectural design and security. It is meant as a blueprint which allows you to achieve the best security controls and visibility aligned with the agility, elasticity, and automated nature of cloud infrastructure.

For detailed operational instructions, separate documents exist per platform to provide practical guidance on creating and deploying the solution.

Cloud Security Architecture Guidance

As stated above, organizations are looking to better utilize their IT resources and align it with the latest and greatest that the cloud has to offer:

- Agility - decrease the time to market interval
- Elasticity - expand and shrink resources on-demand
- Efficiency - only pay for what you use

When designing your cloud-based environment, it is fundamental that the architecture aligns with your and your customers' business use cases all while keeping an uncompromised approach to security.

This document highlights the required principles and best practices to follow in order to build your cloud based environments in a secure manner.

The Top Five Principles

1. **Perimeter Security with Advanced Threat Prevention**
   In recent years, there has been a clear rise in both attack frequency and sophistication of malicious software that is being used. This is happening with correlation to cloud incidents related to vulnerability scanning, web application attacks, and brute force attacks. Many organizations wrongly presume their cloud service providers (CSPs) are responsible for securing their data in the cloud. This is not the case.

   While security is a top priority for CSPs, they typically operate with a paradigm referred to as the Shared Responsibility Model. This actually means that the CSPs assume full ownership (and responsibility) of anything "of" the cloud while leaving anything "in" the cloud as the customer's sole responsibility. The CSPs also provide some elementary security tools that are free of charge for customers but with the latest threats and data breaches, it is obvious that additional advanced threat prevention is required and it is the customer's responsibility to defend their data. It is therefore mandatory for organizations to wrap their environments with the best in-class protections against modern day attacks. This is applied on the environment's perimeter at its main traffic junctions in and out of the environment.

2. **Segmentation**
   Network segmentation is usually done in order to narrow down the network's attack surface and limit a malicious threat's ability to spread across the network freely. Recent cyber-attacks rely heavily on spreading laterally inside a network and infect other machines within that network. This behavior reiterates the need to segment the

network by application or service and place best in-class protections between those network segments.

Security enforcement is done in two layers. The first layer is at the access level where the firewall policy is used to allow certain traffic to flow normally in order to allow normal application operation but also block unwanted traffic between those segments. The second layer of threat prevention is where the firewall inspects traffic which is allowed by the access level but thoroughly inspects it to identify malicious behavior within these flows. This allows applications to securely communicate with each other.

Taking it a step further, cloud's Software Defined Network (SDN) capabilities, also enables us to put those advanced protection inspection points between single hosts (even within the same network segment) and achieve what is also commonly known as "Micro-Segmentation".

Another aspect of segmentation covered by the blueprint is to methodologically enforce traffic restrictions and segmentation to avoid human errors and data leakage from misconfiguration that may expose assets to the public. The method is practiced, for example, by systematically blocking lateral movement through one section of the network while allowing it on an alternate controlled section where it is monitored closely and where security controls are enforced.

3. **Agility**

The ability to operate the business at a high speed and to be truly agile is enabled by the on-demand nature that the cloud has to offer. It is practically impossible to adopt modern, efficient business practices if it takes weeks to provision servers and services, or if your security operation becomes a significant roadblock to the business because each request or approval process is tedious and time consuming.

This blueprint is architected in a way that cultivates agility while ensuring that speed does not come at the loss of control or at an increased operational risk.
This is done by creating scoped delegation of ownership between different stakeholders in the organization. This way, DevOps, application owners, and other groups enjoy enhanced level of authority over their resources and environments. They are thus free to create and manage them. Greater authority comes with enhanced responsibility to own access control within and between their own workloads while leaving threat prevention and advanced security prevention considerations to the Network and Security teams.

4. **Automation, Efficiency, and Elasticity**

Cloud automation is a broad term that refers to the processes and tools an organization uses to reduce the manual efforts associated with provisioning and managing cloud computing workloads. Obviously, this is also very relevant for security operations where the legacy way of manually protecting workloads and resources is no longer relevant in cloud environments.

In a world where business agility is held down because of security operations, the latter is subject to either being bypassed (by using workarounds) or alternatively to opening up the protection in a way that doesn't get in the way of business. When it comes to cloud security operations, automation is vital in order to reduce potential risk and remove the human factor from some organizational processes. The blueprint inherently supports and promotes automation of processes and steps from the ground up, from the environment provisioning phase which is done using a pre-configured template to the day to day policy operation which is done using dynamic, adaptive policies, where no human intervention is required.

5. **Borderless**

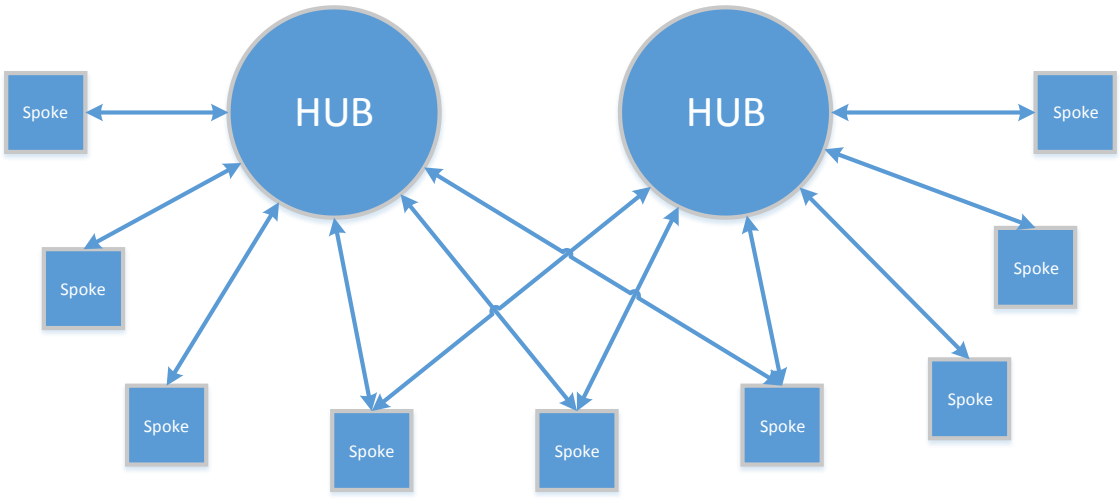
   As previously stated, it is becoming common practice for enterprise customers to launch and run their workloads with multiple cloud vendors, mainly in order to better support their business requirements. The use of multiple cloud computing vendors in a single and heterogeneous environment is also commonly known as a multi-cloud strategy. This is indeed a promising strategy. Using multiple cloud vendors over different geographic locations with a plethora of new and emerging technologies has its security challenges such as:

   a. Enforcing a consistent security policy across all environments
   b. Easily managing the security posture from a unified and central point
   c. Securely connecting various clouds and locations
   d. Allowing applications to easily and securely communicate with each other regardless of their location
   e. Having visibility into the traffic flows in and between locations

   The blueprint answers the above challenges and supports the business urge to follow that strategy.

Check Point's Secure Cloud Blueprint

The blueprint architecture outlined below was designed to meet the above guidelines and assure businesses securely migrate to the cloud. The architectural concept is based on a "hub & spoke" model where the environment is setup as a system of connections arranged like a wire wheel in which all spokes are connected to a central broker (hub) and all traffic to and from the spokes traverses through a broker (hub). The blueprint proposes the usage of two such hubs in the same environment for the sake of traffic separation.



The spokes

A spoke is an isolated network environment which contains a collection of one or more network subnets from which typical workloads can be installed and run from. A common use case is a spoke that contains several virtual servers that comprise either a part of or an entire application stack (web, application and database). Another use-case is a spoke which acts as an extension of an existing on premises network, such as a set of QA servers for testing purposes or a set of data processing servers which utilize the cloud's on-demand provisioning for lower cost and improved agility.

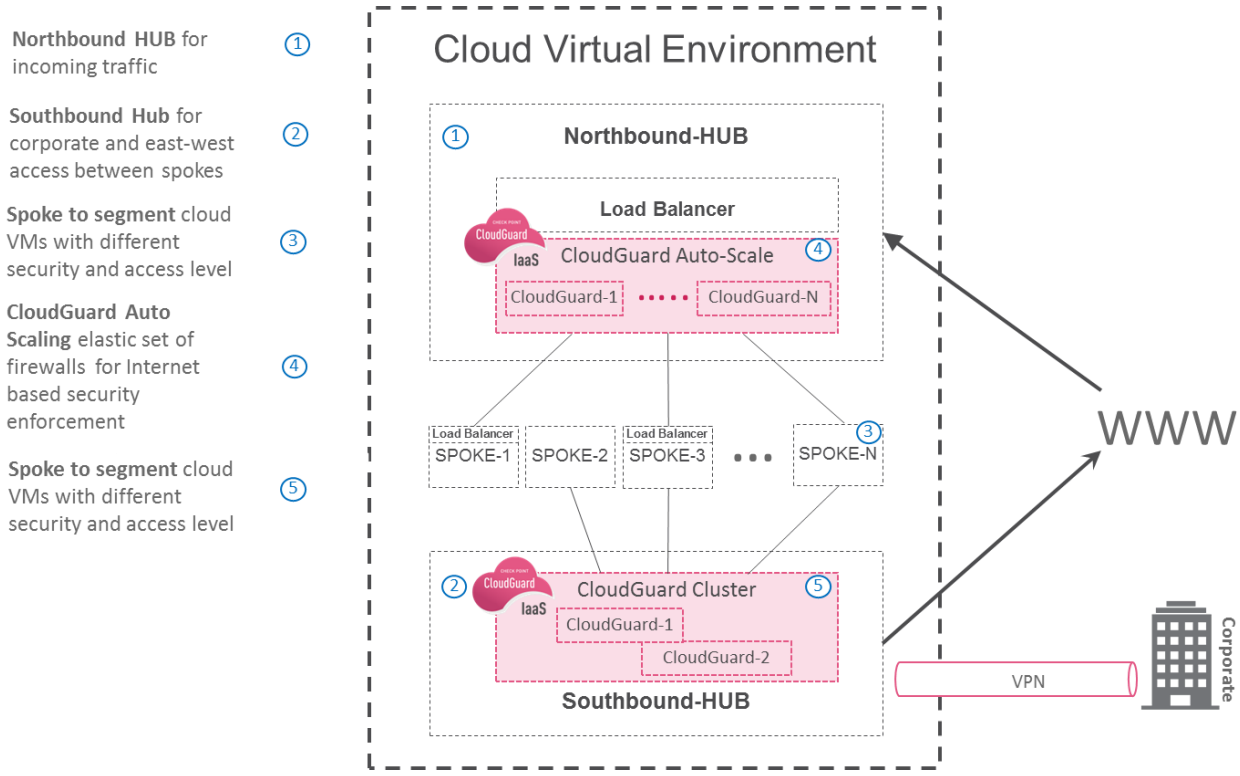
This blueprint was created as a high level design document which is applicable to all leading cloud environment such as AWS, Azure, Google, Oracle Cloud, Alibaba Cloud, and others.

The hubs

In the diagram below, we utilize two hubs within the environment. This enables flexibility and systematic separation of communication types through the environment. One such hub is designated for incoming traffic from the Internet and the other hub is designated for lateral traffic between spokes, traffic in/out of the corporate network, and outgoing traffic to the Internet or other cloud environments.

Traffic flow inside the environment

By configuring routing and connections between hubs and spokes, we achieve a situation where the hubs are the only way in/out of the environment as well as the only way to traverse inside and between spokes in the environment as spokes are not connected to each other directly but actually only accessible through one of the hubs. This achieves both perimeter and segmentation of the environment.



Perimeter Security

Perimeter enforcement is performed on the hubs (north and south). The recommended security protections to activate on the perimeter are Anti-Virus, Anti-Bot, and IPS.

Segmentation

Segmentation is achieved by placing our resources in different spokes, and enforcing security controls on traffic entering or exiting a spoke. Three main types of spokes exist in the blueprint:

- Internet facing only (e.g. SPOKE-1 in the drawing above) – These spokes are connected to the northbound hubs and thus only accessible through incoming traffic from the Internet. Commonly, those spoke will host front tier servers that are internet facing and that need to be accessible from the Internet. Connectivity from those spokes to corporate resources or to other spokes in the environment is blocked systematically and cannot be enabled by a simple configuration (This is done to avoid human errors and mistakes which expose unwanted resources to the public).
- Private facing only (e.g. SPOKE-2 in the drawing above) – These types of spokes are only connected to the southbound hub and thus are systematically not accessible from the Internet but rather accessible through VPN and/or direct connectivity from the corporate network or from other spokes in the environment (as per the security policy on southbound firewalls). A practical example for such a spoke is hosting database (DB) servers. We do not want them to be directly accessible from the Internet but we want to be able to have secure connectivity.
- Combined (e.g. SPOKE-3 above) – These types of spokes are suitable for servers that are both accessible from the Internet but also require backend access to other spokes or to the corporate network. An example of such use-case is a web server that is exposed to the Internet on one end and needs access to an application server or database server on the other.

Agility

In order to enable and support business agility, the spokes can be created and entirely owned by the organization's different LOBs (lines of business). In fact, anyone within the organization can be a spoke owner as long as this aligns with the organization's policy. Once created, servers, containers, and any other workload within that spoke are controlled and maintained by the spoke owner. This gives freedom to operate inside the spoke whether it's creating, developing, or launching a service or application. It enables the cloud's most desired feature of agility where there is no ticket overhead or dependency for anything occurring inside each spoke.

Automation

As stated above, another important aspect of the blueprint is incorporating IT into cloud operations. The blueprint eases the process of bringing IT to the cloud and helps utilize the cloud's best features such as automation and orchestration. The end game is to allow IT to operate and be a business enabler instead of being a business roadblock.
By using pre-configured virtual firewall deployment templates, IT is capable of securely implementing an entire environment with the "click of a button" and with very little to no hands-on configuration.

This is true for the provisioning of the environment as well as for running daily operations, and supporting an elastic environment out of the box. Take an environment similar to the one in the above diagram as an example. An application owner within the environment adds a new spoke. The Check Point management server (SMS) automatically identifies this new spoke and automatically forms the required secure connections to and from it. This allows full visibility and control of traffic to and from the newly created spoke, and assures it complies with standards and policies as identified by IT.

The same level of dynamicity is achieved with the security posture of the organization where a policy can be pre-approved and then dynamically assigned to workloads (based on resource tags, for example). Changes are instant, allowing business owners to progress at their own pace while making sure that the company's policies and standards are met.

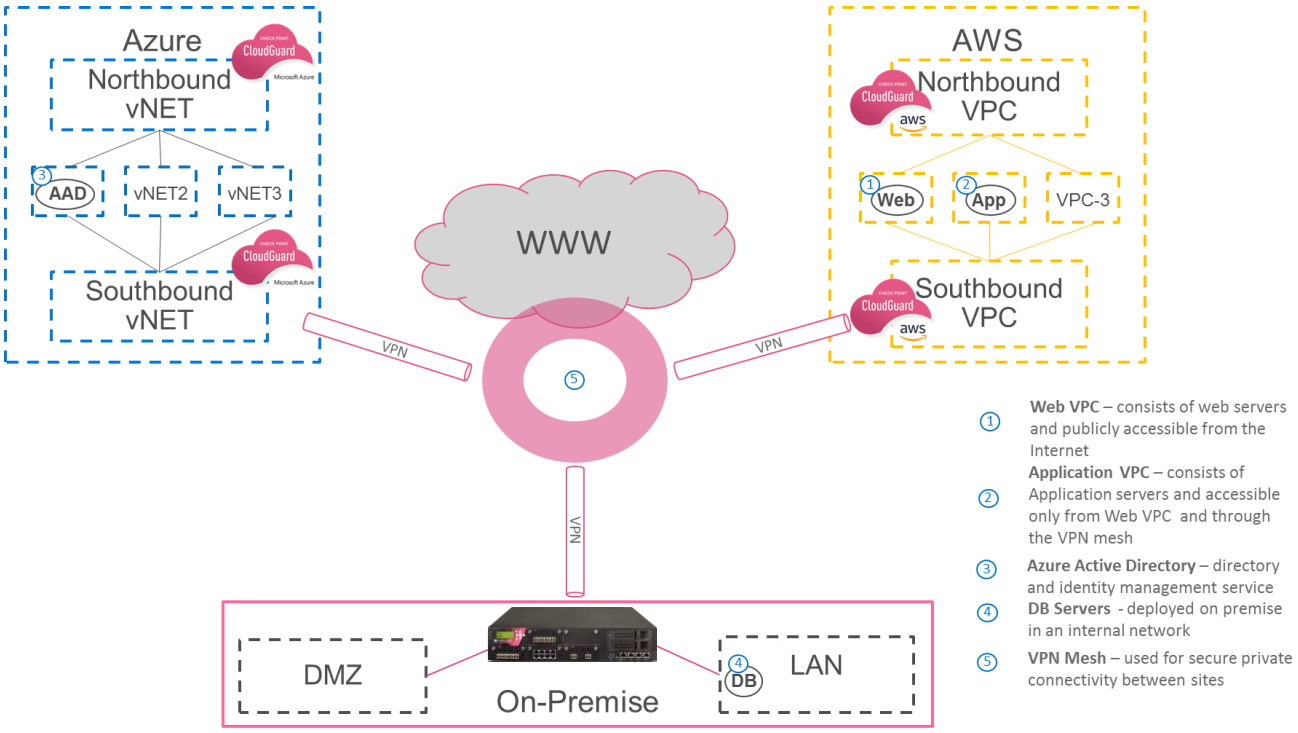
Borderless

This design also inherently supports scaling outside of the regular limits of a single cloud platform and also handles the connectivity between cloud platforms while maintaining the same architecture principles and keeping the same security posture across all the environments.

An example for such a multi-cloud architecture is an online gaming company deploying its services across AWS and Azure. The logic behind it is a "best of breed" approach where each platform is chosen based on team expertise and technological superiority.

For example, the web frontend and application tiers are hosted with AWS across multiple availability zones for redundancy and the Identity and Authentication is provided by the Azure integrated AD service while the DB and storage tiers are hosted in the on premises data center.

This environment should be secure and allow flexibility and agility out of the box. The diagram below illustrates this use case:



As shown in the diagram, implementing CloudGuard IaaS gateways across different platforms allows us to both control traffic in each location specifically but also enforce our security policy across and between locations.

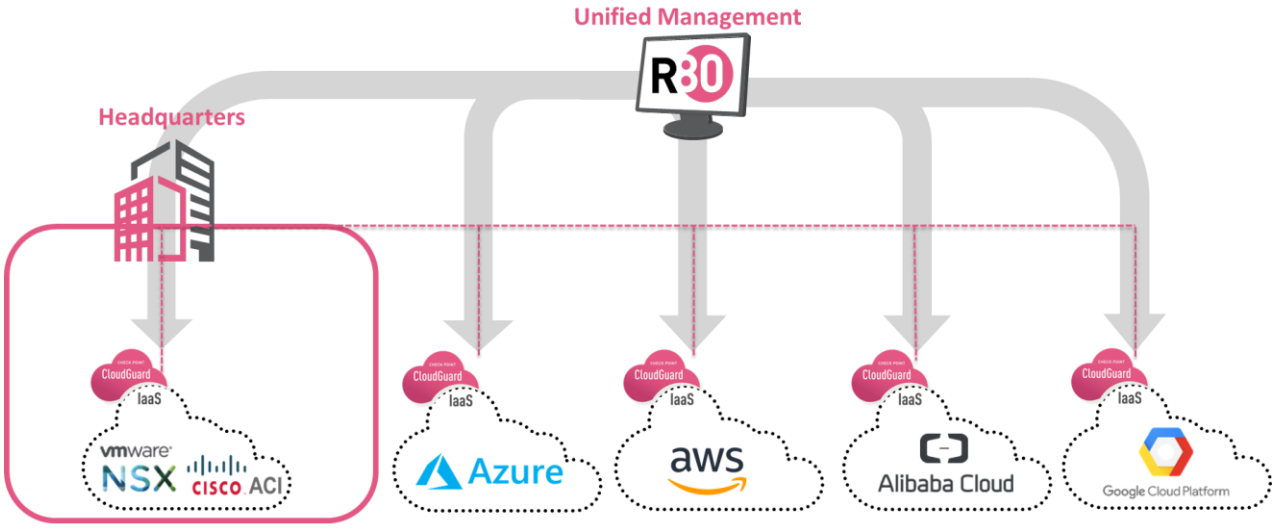
Additional considerations

Unified management

Operating security in a multi-cloud environment can be challenging as it involves managing and controlling resources in a variety of locations that use different management tools. Trying to maintain a unified policy in such conditions is obviously tedious and inefficient, notwithstanding troubleshooting connectivity issues or security events in the environment.

R80.10 management server (SMS) is an integrated security management solution which includes policy, logging, monitoring, event correlation, and reporting – all in a single system using a unified security policy which enables administrators to easily identify security risks across the environment and maintain the security policy.

A unified policy enables organizations to translate their security definitions into a simple set of rules, which then streamlines policy administration and enforcement throughout the organization.



Management server location is completely flexible and can be located anywhere across the environment. It is also possible to setup the management server in high availability mode across platforms (e.g. primary management located on premises with a backup server located in a location designated for disaster recovery).

Redundancy and Resiliency

As a rule of thumb, the blueprint is created with built-in resiliency for local failure events. This is implemented in several layers of the environment.

1. Data Center level redundancy – The environment is built into multiple (2 or more) zones where each represents a separated Data Center (e.g. separate network, electricity, air conditioning, and usually even a separated building).

2. Software level redundancy – The gateways are deployed with N+1* redundancy across the environment. This translates into two separate solutions based on the location and role of the gateways.
    a. On the Northbound Hub, where http/https based connections are entering from the Internet, the gateways are implemented in an elastic manner where the number of gateways is dynamic and is based on the load that flows through the gateways. This type of scaling, also known as horizontal scaling**, is whenever the load increases, additional gateways are added and are put into production within a few minutes (five to seven minutes for gateway initialization) and the load is balanced between gateways. When the load reduces, unneeded gateways are removed from the pool and the environment is kept efficient from a cost and performance point of view.
    b. On the Southbound Hub, the gateways are deployed as an active-standby cluster. The scaling mechanism here is known as vertical scaling***. Whenever there is a need to support an increase of the load through that hub, more resources are added to the gateways respectively.

* N+1 redundancy is a form of resilience that ensures system availability in the event of component failure. Components (N) have at least one independent backup component (+1).

** Horizontal scaling means that you scale by adding more machines into your pool of resources

*** Vertical scaling means that you scale by adding more power (CPU, RAM) to an existing machine

Failover

In case of failure of a gateway in the Northbound Hub, the connections that existed on that gateway are not preserved and new connections are rebalanced to a healthy gateway in the environment. The connections being http/https based are stateless by nature. The user experience is a mere refresh of a browser in a matter of seconds.

In the Southbound Hub, where connections are more diverse, complex, and usually stateful, the connections are constantly synchronized between cluster members. A failover of the active gateway will cause the standby member to regain all active connections and become the active member.

Connection Affinity

Connection affinity in the Northbound Hub is based on the client's IP address and port number such that whenever a connection is initiated from the Internet, the load balancer choses the target gateway to send the connection to and keeps the same target as long as the session is kept alive.

Affinity on the Southbound Hub is based on the active member so that all traffic is always directed to the active member.

Recommended Sizing

Sizing the solution is normally based on performance requirements inside the environment and the required security level. The typical recommended environment will consist of the following sizing:

1. In the Northbound Hub, a minimum of 2 gateways (N+1) each with 4 virtual CPU cores and 8GB of RAM is recommended. As stated above, scaling is performed horizontally, thus new additional gateways are automatically added to the environment upon growth of the load hitting the environment.
2. In the Southbound Hub, the recommendation is that the cluster be comprised of 2 gateways each with 8 virtual CPU cores and 8GB of RAM. Growth is vertical, which means that more resources (CPU/RAM) are added to each gateway.

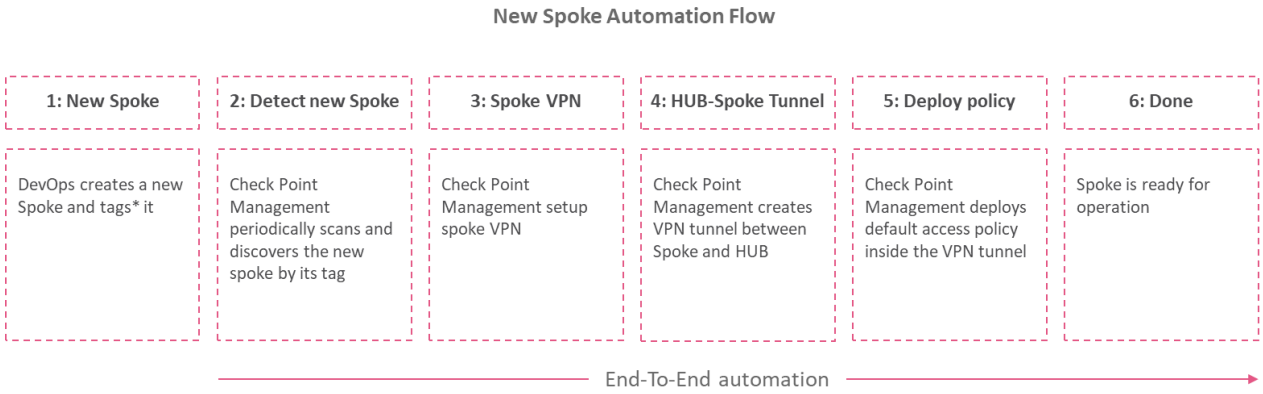
Disaster Recovery

Setting up a resilient architecture for a doomsday event is common practice with most organizations. Having the flexibility gained through this blueprint, a DR site can easily be created and maintained as another site within the architecture. Furthermore, the infrastructure can be narrowed down to have no redundancy but sufficient to support the load during peak time.

Practical implementations of the Blueprint

To apply the Secure Cloud Blueprint design concepts outlined in this document in a practical setting, we've added support for automation into our construct. The way this automation works is once the Blueprint framework is deployed, any new spoke introduced into the environment (via tagging) is automatically configured and added to existing VPN connectivity, allowing it to communicate with networks outside the spoke (e.g. Neighboring spokes, Corporate, Internet). Traffic to and from the spoke is then forced through the hub where it is inspected by the security policy before being forwarded on.

The automation flow is described below:



* Tag= "x-chkp-vpn = <Check Point management host name>/<vpn-community name>"

This type of built-in Automation is currently supported on AWS

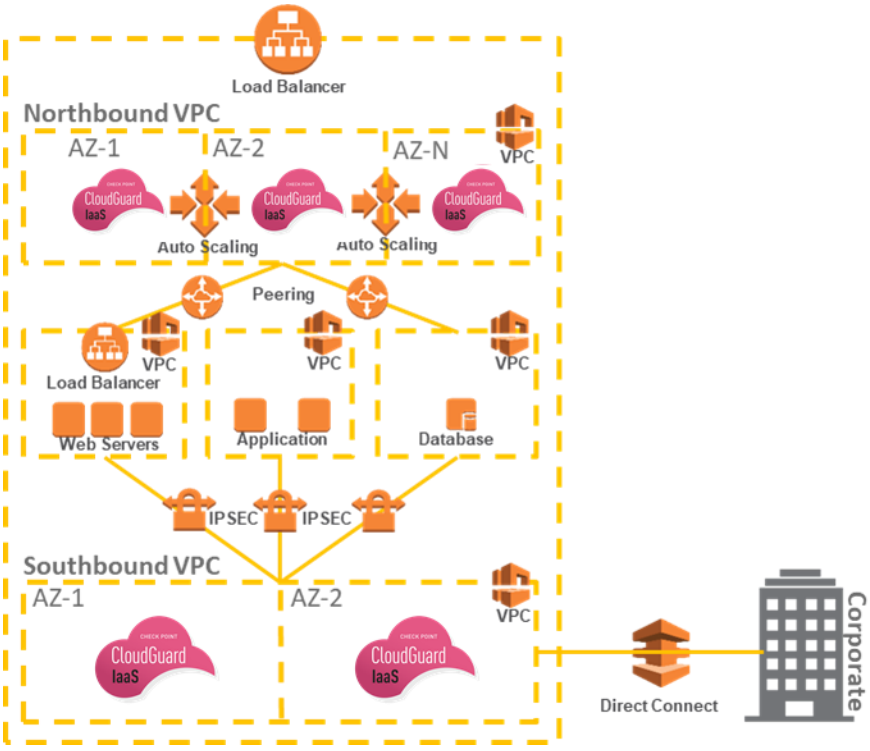
On AWS

From a terminology standpoint, the HUBs and Spokes are constructed by Virtual Private Clouds (VPCs) on the AWS platform. Setting up the blueprint on AWS is done using predefined CloudFormation templates.

In the case of AWS, the spoke VPCs that are to be accessible to the Public Internet are connected to the Northbound VPC using VPC peering capabilities. Since VPC Peering in AWS is non transitive, meaning you cannot traverse to a third VPC using the connection between the first two, we utilize peering to make sure that traffic originating from the Internet only travel one-hop within the environment (systematically). For Internet-bound connections originating from the Spoke VPCs, connections are routed through the Southbound VPC. Spoke VPCs are connected to the Southbound VPC via IPsec tunnels over route based VPN with AWS's VGWs at each Spoke. This infrastructure is also used for inter-VPC connectivity (east-west within the environment), access to on premise environments and access to Internet.

In this method, the security gateways in the Southbound VPC propagate routes to the VGWs to steer outgoing traffic from the spokes through the southbound gateways for inspection, routing and VPN services. The southbound VPC also creates VPN with the on premises environment using a separate VPN community that may be either domain based or route based.

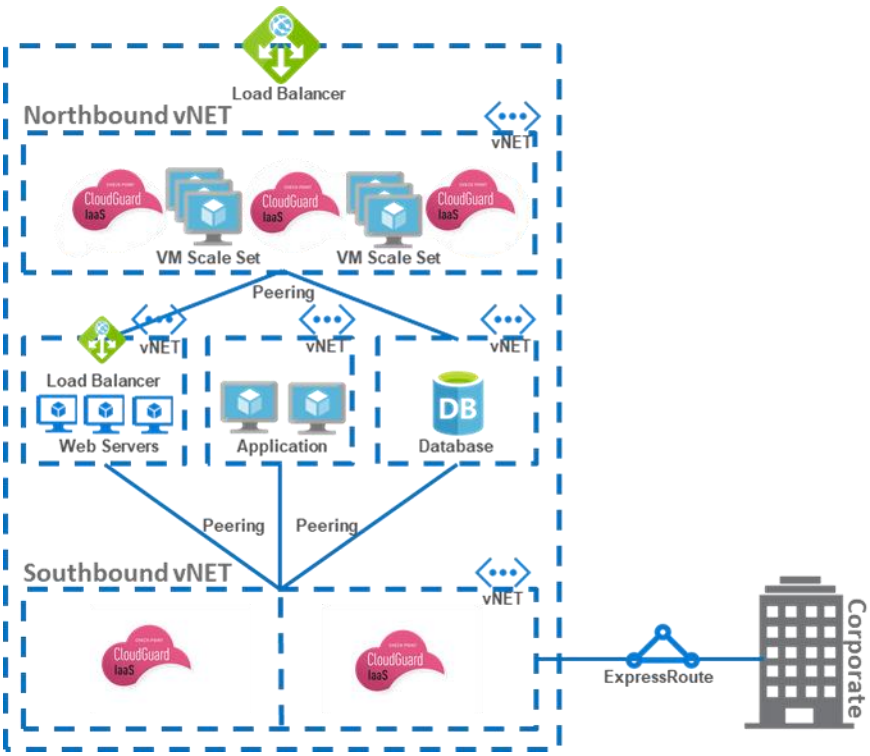
In the Northbound VPC, the CloudFormation template deploys an Auto Scaling group and an internal load balancer to serve the published applications in the spoke VPCs. ELB, NLB or ALB may be deployed as an external load balancer to the environment. In the Southbound VPC the CloudFormation template deploys 2 security gateways in separate AZs to serve the Spoke VPCs in high availability mode.



On Azure

From a terminology standpoint, the HUBs and Spokes are constructed by Virtual Networks (vNETs) on the Azure platform. Setting up the blueprint on Azure is done using predefined solution templates.

For Internet-bound connections originating from the Spoke vNETs, connections are routed through the Southbound vNET. Connection from Southbound vNET to on premises is via ExpressRoute or VPN over the Internet (domain based or route based). All other connections are based on peering between vNETs of the environment. Peering in Azure could be configured as transitive; however from security perspective that capability should be avoided. This is done using the peering configuration settings and by forcing traffic using UDRs to the CloudGuard IaaS gateways of the vNET hubs (either Southbound or Northbound).



Summary

The above design principles allow maximum operational flexibility where each tenant (e.g. department, customer, application owner, etc.) deployed in a dedicated spoke environment is free to create its own network and compute resources within that environment while maintaining unconditional control and security of anything traveling to and from these environments. This is done to comply with the company's security policy and allow the same level of control and visibility of the company's network by security teams; similar to what is done in a legacy network environment.

