Domain 7: Infrastructure & Networking

Introduction

The *infrastructure* in Infrastructure as a Service (IaaS) refers to the compute, network, and storage resource pools. Other domains cover compute (workloads) and storage (data) security concerns. This domain covers managing the overall infrastructure footprint and network security. It also includes a small section on the infrastructure security responsibilities of the CSP.

Beginning with cloud infrastructure security, this domain then covers cloud networking fundamentals and Software Defined Networks. We will then discuss elements of security architecture and 'Infrastructure as Code'. The domain finishes with a brief introduction to upcoming approaches for network and infrastructure security: Zero Trust, and Secure Access Service Edge.

Learning Objectives

In this domain, you will learn to:

- Understand the areas and techniques used in securing cloud infrastructures.
- Understand cloud network fundamentals.
- Manage container networking.
- Manage cloud network security and design secure architectures.
- Apply Zero Trust techniques to securing cloud infrastructure and networks.
- Techniques used to manage security for a Secure Access Service Edge (SASE).

7.1 Cloud Infrastructure Security

As discussed in previous domains, cloud infrastructure provisions fundamental resources. It is up to the cloud customers to design and build their architecture on the secure services the cloud provider offers. Since so much of IaaS and PaaS design relies on the cloud customer, it is important to understand the proper use of the infrastructure and how to construct well-architected implementations that help achieve the benefits of the cloud.

7.1.1 Foundational Infrastructure Security Techniques

Three foundational techniques have the most impact on creating and maintaining secure infrastructure:

- **Secure architecture:** Starts with designing the cloud infrastructure with security as a key principle. This includes properly segregating resources and networks, implementing least privilege access, and ensuring secure storage, communications, and service configurations.
- **Secure deployment and configuration:** This involves configuring and/or deploying resources and services, and hardening all cloud infrastructure components, including virtual machines (VMs), containers, storage, and networking. This includes applying security benchmarks and best practices, such as Center for Internet Security (CIS) benchmarks[31], to ensure proper configuration settings.
- **Continuous monitoring and guardrails:** Guardrails are preventative and reactive controls that either block an undesired outcome (e.g., block use of regions, public object storage, or specific cloud services that aren't approved) or auto-remediate/correct a policy violation. This also includes enabling logging and monitoring across all components to detect and respond to security events. Guardrails, such as Amazon Web Service (AWS) Config rules and Service Control Policies, as well as Microsoft Azure Policies, can enforce security policies and prevent deviations from the desired state.

7.1.2 CSP Infrastructure Security Responsibilities

Infrastructure security starts with the CSP, which ensures a secure platform for customers to build or consume on. Under the shared security responsibility model (SSRM), infrastructure security is primarily the CSP's responsibility. CSP Infrastructure Security Responsibilities include:

- **Facilities**: The CSP is responsible for ensuring the physical security of the facilities where the cloud infrastructure is housed. This includes measures like access control, surveillance, and environmental protection.
- **Employees**: The CSP screens, trains, and manages employees with access to cloud infrastructure, helping maintain organization integrity and trustworthiness.
- **Physical network, storage, and compute**: The CSP secures and maintains the underlying physical components of the cloud infrastructure, such as servers, storage devices, and networking equipment.
- **Virtualization layers**: CSPs are responsible for securing the virtualization technology that enables the creation and isolation of virtual machines (VMs) and containers running on the physical infrastructure. Server virtualization is typically done through a hypervisor.
- **Management plane**: The CSP secures and controls access to the web-based interfaces and API endpoints customers use to manage their cloud resources and services.
- **PaaS and SaaS services**: The CSP offers higher-level platforms and software services that handle the security of the underlying infrastructure and applications based on the shared responsibility model.

[31] CIS. (2024) *Foundational Cloud Security with CIS Benchmarks*.

In summary, the CSP secures the physical facilities, hardware, virtualization layer, and management interfaces that comprise the cloud infrastructure. Customers can focus on securing what they consume and deploy on that infrastructure.

7.1.3 Infrastructure Resilience

While cloud providers strive to have highly reliable infrastructure, at the scale of cloud there will be technical failures with varying levels of impact on an ongoing basis. In this section, we'll explore some of the options for mitigating this type of risk.

In the realm of cloud computing, resiliency refers to the ability of an application or system to continue operating seamlessly in the face of various types of disruptions, ranging from minor faults to major outages. The concept of cloud resiliency is layered and can be scaled according to the criticality and budgetary constraints of the services in question.

At the base level, *single-region resiliency* is where most applications begin their journey towards being resilient. In this setup, the application is hosted within a single cloud provider's region, and it employs strategies like auto-scaling and load balancing to handle sudden spikes in traffic, and to be fault-tolerant against individual component failures. Backup and recovery strategies are also put in place to protect data. This foundational level is also the most cost-effective option since it takes advantage of the cloud provider's existing infrastructure and services without the need for significant duplication of resources.

Single-region deployment is, however, vulnerable to regional outages. Although rare, it can have a significant impact on the availability of the application. To mitigate this risk, organizations can step up to *multi-region resiliency*. This involves running parallel deployments of the application across multiple regions within the same cloud provider's network. While this significantly improves fault tolerance and geographic diversity, it also introduces additional costs. These costs come not just from running multiple instances of the application but also from the need to synchronize data across regions. Furthermore, data transfer charges incurred for moving data between regions can quickly become substantial, making this a more expensive option than single-region deployment.

The hardest cloud resiliency is *multi-provider resiliency*. This level is achieved by spreading the application's footprint across multiple cloud providers. The intention is to safeguard the application from a scenario in which an entire cloud provider goes down. Achieving multi-provider resiliency is complex due to the technological differences between cloud providers. Containerization technology can ease some of this complexity by abstracting the application from the underlying infrastructure, but challenges remain. These include managing disparate networking, storage, and security models, as well as orchestrating deployment and operations across environments that are fundamentally different. Costs can escalate rapidly, not just in terms of direct operational expenses but also due to the increased overhead required for design, development, testing, and ongoing maintenance. Despite the costs and complexities, for critical applications that demand the utmost availability—those involved in financial transactions, health services, or global commerce—multi-provider resiliency can be a necessary investment.

7.2 Cloud Network Fundamentals

Cloud networks are software-defined networks. Enforcing strong separation between customer-tenant environments is essential.

The following figure demonstrates the network control logic in a SDN environment, illustrating the encapsulation and unwrapping of packets as they traverse between physical hosts through network control logic.

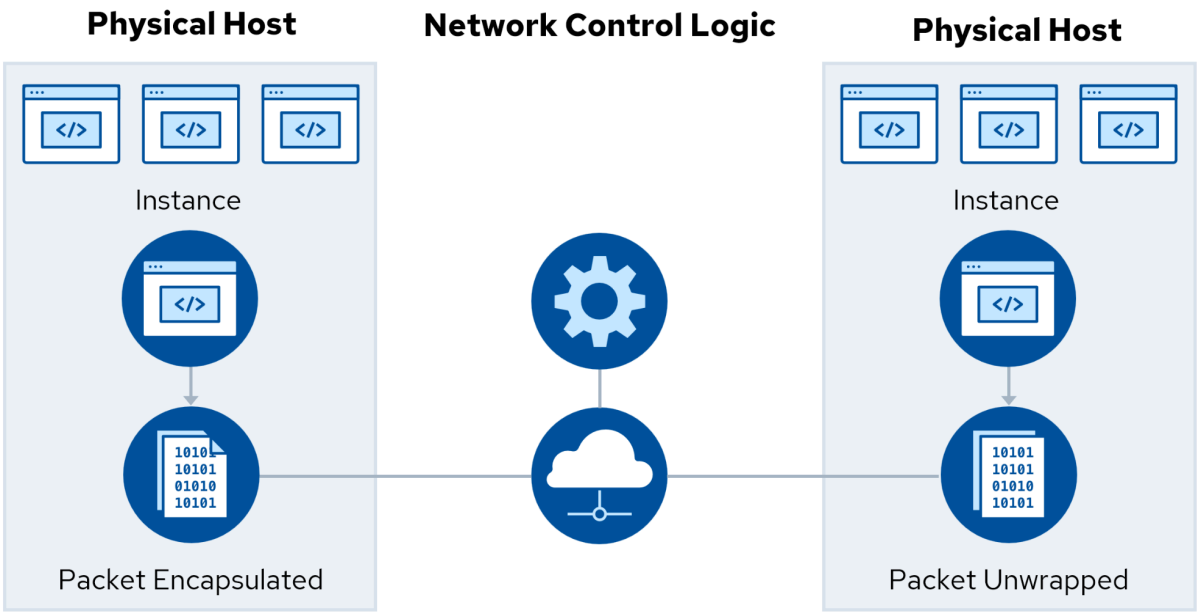

Figure 17: Network Control Logic

7.2.1 Cloud Networks are Software-Defined Networks

Software-defined networking (SDN) decouples the network control plane from the data plane, allowing network components such as routers, switches, and firewalls, to be programmatically configured and controlled through software. The control plane manages routing, network/subnet definitions, and similar while the data plane moves the network traffic between resources and networks. This shift from traditional hardware-based networking to software-defined approaches brings numerous benefits to cloud environments.

One of the primary advantages of SDN is its flexibility and agility. With SDN, network administrators can dynamically configure and manage network resources through software, enabling rapid provisioning and modification of network services. SDN can be used on any network but is the default on all IaaS providers.

IaaS/PaaS providers offer SDN. Customers are not managing individual network components, such as routers, switches, and their access control lists (ACLs). Instead, they define logical groupings and network

connectivity, and let the provider, through their SDN technology, configure individual network components.

7.2.1.1 Common SDN-Based Components

Most cloud networks share a consistent set of foundational components. The following are some common SDN-based components for the cloud:

Virtual Networks/Virtual Private Clouds:

- Some providers use Virtual Networks (or VNet), and others use Virtual Private Cloud (or VPC)
- Group subnets into logically isolated virtual networks
- Enables customers control over network topology, routing, and IP address ranges
- Provides a secure and private networking environment for cloud resources

Subnets (Public and Private):

- Smaller network segments within a VNet/VPC
- Allows for further segmentation and organization of resources
- Enables the application of different security and access control policies
- Private subnets also require a Network Address Translation (NAT) service for egress (outbound) Internet access

Route Tables:

- How network traffic is directed within a VNet/VPC
- Specify the paths for traffic between subnets and external networks
- Enable custom routing configurations for optimal network performance

Cloud Network Security Groups:

- Similar to a stateful firewall but are implemented within the network fabric itself
- Act as virtual firewalls at the network interface, instance, or subnet level
- Control inbound and outbound traffic based on IP addresses, ports, protocols, and other criteria
- Provide granular-level security for individual resources or groups of resources
- Foundation for Micro-Segmentation/Default deny approaches; each network interface is segregated by default from others

Cloud Network ACLs:

- By specifying which packets can pass through network devices and environments, ACLs control both inbound and outbound traffic
- ACLs work lower in the network stack than Security groups and are typically stateless
- Security groups most often apply to resources (e.g., instances), while ACLs apply to subnets/networks
- Implementations of both are different on various CSPs

Load Balancer Service:

- Spreads traffic over multiple VMs running web servers
- Can autoscale to manage fluctuating demand
- Is a service rather than a device
- Enables redundant architectures when spreading applications over multiple availability zones
- A foundation for other security services, such as Web Application Firewalls (WAF) and distributed denial-of-service (DDOS) attack protection

Internet Gateways:

- Serves as the entry and exit points for internet traffic in a VNet/VPC
- Allow resources within the VNet/VPC to communicate with the public internet
- Enable inbound and outbound internet connectivity for cloud resources

Endpoints:

- Some CSP services are accessible by default over public IPs
- Private endpoints are specific network addresses, or URLs, through which cloud services can be accessed using internal addresses and routing
- Private endpoints enable applications to communicate with CSP services without the need to go over the public Internet
- Private endpoints are recommended by CSPs as they can increase security and performance

7.2.2 Cloud Connectivity

Connecting to Resources: Cloud providers allow their resources to be available on the internet. After all, broad network access is one of the essential characteristics of cloud computing. These resources include:

- Management plane
- Workloads (VMs, containers)
- Storage
- Load balancers
- APIs

Access to these can be restricted in several ways, depending on the service.

Cloud providers support a variety of private networking options. These include traditional VPNs and third-party connectivity services. These can be used to connect a cloud deployment or VPC to a local data center, or even multiple data centers to multiple cloud deployments. When improperly architected, this can be a big risk and expose a lot of resources to the public internet.

While private data center networks can have some intrinsic security benefits, their configuration also has some security risks. In a typical private network setup, there will be an IP address range that straddles the boundary between the data center and the cloud provider's resources. While both ends can implement

security controls, such as security groups and ACLs, these are not necessarily effective across the entire IP range, which may leave security holes.

The number one security control is the ability to control traffic flows between these networks. This will reduce the blast radius of cloud attacks.

7.3 Cloud Network Security & Secure Architectures

As you gain experience developing cloud-based network environments, you will also learn about recommended reference architectures. Regardless of which mixing-and-matching you do with an on-premises setup, it is important to have a solid idea of what each one of these CSP services does and why it has been built for all the major hyperscalers and CSPs (e.g., AWS, Azure, GCP, IBM, Oracle).

7.3.1 Preventative Security Measures

CSP Firewalls: CSP firewalls, such as Amazon Network Firewall or Azure Firewall, are built into the cloud platform. They offer the advantage of not requiring additional instances or servers to maintain, simplifying management and reducing operational overhead. However, they may have limitations in terms of customization and advanced features compared to virtual appliances.

Virtual Appliances: Virtual firewall appliances provide greater flexibility and control over firewall rules and configurations. They can be deployed in a load-balanced configuration to ensure high availability. However, this approach adds complexity and requires ongoing maintenance of the virtual machines or instances running the firewall software. Virtual appliances are commonly available for next generation firewalls (NGFWs) and intrusion detection system/intrusion prevention system (IDS/IPS) products.

WAFs: WAFs specifically protect web-facing applications from common exploits like SQL injection, cross-site scripting (XSS), and other Open Worldwide Application Security Project (OWASP) Top 10 vulnerabilities. Depending on the CSP, and your requirements, they can be deployed as a cloud-native service or as a virtual appliance.[32]

7.3.2 Detective Security Measures

Flow Logs and DNS Logs: Flow logs and domain name system (DNS) logs provide valuable visibility into network traffic patterns and help detect anomalous activities. Flow logs capture information about the source, destination, protocol, and other attributes of network flows, while DNS logs record domain name resolution requests and responses. These logs can help identify potential security breaches, unauthorized access attempts, and data exfiltration. However, handling these logs at scale is challenging[33].

[32] Additional material on security monitoring and logs is provided in Domain 10: *Application Security*.
[33] Additional material on security monitoring and logs is provided in Domain 6: *Security monitoring*.

7.4 Infrastructure as Code

One of the biggest changes the cloud brings is the capability to programmatically create complete infrastructure architectures. Many concepts are used for this, including programmable infrastructure, infrastructure automation, and declarative infrastructure. We'll use Infrastructure-as-Code (IaC). The enabler for this is the API of the cloud management plane.

Infrastructure-as-Code (IaC) is defined in the NIST Special Publication (SP) 800-172 as "The process of managing and provisioning an organization's IT infrastructure using machine-readable configuration files, rather than employing physical hardware configuration or interactive configuration tools"[34]. IaC is the dominant model for deploying cloud resources and is supported by every major provider. IaC is discussed in multiple areas of this guidance, including the sections on application and workload security.

Key IaC concepts are:

- Architectures can be described by code and defined in a machine-readable format, from low-level network design to high-level application components.
- Through the management plane API infrastructure and services are deployed and configured.
- Typically deployed using continuous implementation/continuous delivery (CI/CD) automated pipelines[35].
- Security scanning for misconfigurations can occur in the pipeline.
- Full version control and change tracking.

Amongst the benefits of IaC are:

1. **Automated Compliance Checks:** IaC facilitates automatic validation against security standards and regulations, ensuring compliance every time infrastructure is provisioned or modified.

2. **Consistent Security Posture:** By codifying infrastructure setup, IaC can guarantee that every element, from servers to databases, is consistently configured according to security best practices. This eliminates human error associated with manual setups, detects and eliminates configuration drifts, and maintains a uniform security level across all resources.

3. **Rapid Rollback:** In the same way that IaC supports swift and automated roll-out, it also supports rapid rollback. This is very helpful in fixing update problems and reduces change management complexities.

IaC is a component of many automated deployment pipelines and contributes to *shift left* security functions.[36] Most cloud providers have native IaC tools, and various open-source products for IaC exist. IaC has such significant benefits for reproducibility and security that many organizations mandate its use in place of manual configuration.

[34] NIST. (2024) Information Technology Laboratory - Computer Security Resource Center Glossary.
[35] Additional material on CI/CD is provided in Domain 10: *Application Security*.
[36] Additional material on the shift left security function is provided in Domain 10: *Application Security*.

7.5 Zero Trust for Cloud Infrastructure & Networks

Zero Trust (ZT) is a comprehensive cybersecurity strategy. A more elaborate discussion on ZT is in *Domain 12: Related Technology and Strategies*. In this domain, we focus on the implication of ZT for Infrastructure and Networking.

7.5.1 Software-Defined Perimeter & ZT Network Access

Two key technology approaches that enable ZT network security are Zero Trust Network Access (ZTNA) and Software-Defined Perimeter (SDP). These approaches are not mutually exclusive and elements of each can be combined into a tailored ZT security implementation.

7.5.1.1 Software Defined Perimeter (SDP)

- Establishes a secure, "dark" network that is invisible to unauthorized users and devices
- Implements a blackout approach, where the network is inaccessible by default
- Users and devices must authenticate and be authorized before they can access the SDP-protected resources.
- SDP leverages identity-centric controls and micro-segmentation to limit lateral movement

According to the SDP Architecture Guide v2, SDP works in the following manner:

- The SDP Client software on the initiating host opens a connection to the SDP controller. Initiating host devices, such as laptops, tablets, and smartphones, are user-facing, meaning the SDP Client software is run on the devices themselves. The network can be outside the control of the enterprise operating the SDP.
    - The SDP Controller is the policy decision point (PDP) where, based on a variety of attributes, an access decision is taken. It instructs the SDP gateway on a separate channel.
    - The SDP Gateway provides authorized users and devices with temporary access to protected processes and services. The gateway can also enact monitoring, logging, and reporting on these connections.

In this way, the protected resources are completely invisible to unauthorized users and devices.

7.5.1.2 Zero Trust Network Access (ZTNA)

- Replaces traditional virtual private networks (VPNs) with a more granular, application-specific access control model
- Users are verified and authorized based on identity, device, location, and other contextual factors
- Access is provided to specific applications or resources rather than granting broad network access
- ZTNA solutions can be cloud-hosted (ZTNA-as-a-Service (ZTNAaaS)) or on-premises

7.6 Secure Access Service Edge

SASE is an emerging cybersecurity concept that combines network security functions with Wide Area Network (WAN) and proxy capabilities to deliver a comprehensive, cloud-native service. It is designed to address the challenges of securing endpoint devices and access to applications and data in a cloud-first, mobile-first world, where users and resources are increasingly distributed outside of traditional network perimeters.

Historically, traffic control points, such as firewalls, were built on the perimeter of data centers. Beyond protecting the servers in the data center, this aimed to protect users and their devices from malicious websites, but also to prevent data exfiltration. Common terms in this category include Firewalls, Next-Generation Firewalls, Proxies, Secure Web Gateways, Data Leakage Prevention (DLP) tools, and Cloud Access Security Brokers (CASB). They can filter traffic on IP addresses and ports, Web URLs, content inspection, user attributes, user behavior, threat intelligence, and more.

SASE can be an important element in a Zero Trust Architecture (ZTA). A core element of SASE is to bring this filtering to the edge, that is, near the user's device. This works through a combination of centrally controlled endpoint agents and a global network of points of presence where security functions are executed. This avoids the need for 'backhauling' or 'hairpinning', where a user on one continent accesses a cloud service on the same continent, but the traffic is routed through a data center on another continent.