AMAZON AWS SOLUTIONS ARCHITECT

AMAZON COMPUTE & EC2 OVERVIEW

KEY WORDS AND PHRASES

Component	Description
Instances	Virtual computing environments
Amazon Machine Images (AMIs)	Preconfigured templates for your instances
Key Pairs	Secure login information for your instances
Instance Store Volumes	Storage volumes for temporary data
Amazon Elastic Block Store (Amazon EBS)	Persistent storage volumes for your data
Regions and *Availability Zones*	Multiple physical locations for your resources
Security Groups	A firewall that enables you to specify the protocols, ports, and source IP ranges
Elastic IP Addresses	Static IPv4 addresses for dynamic cloud computing
Tags	Metadata assigned to your Amazon EC2 resources
Virtual Private Clouds (VPCs)	Virtual networks you can create that are logically isolated

INSTANCE TYPES

Instance Type	Use Case	Notable Features
T2, M5, M4	General Purpose. Web Servers, Micro Services, Development	Burstable CPU Credits for T2. Lowest cost to use. Good balance of compute, memory, and network resources. EBS (M5/M4) or Ephemeral storage (T2). Enhanced Networking for M5 and M4.16xlarge
C5, C4	Compute-Intensive Workloads. High performance web servers, scientific modelling, batch processing	EBS by default. *C4 Requires Amazon VPC, Amazon EBS and 64-bit HVM AMIs*
X1e, X1, R4	Optimized for high-performance databases, in-memory databases and other memory intensive enterprise applications	SSD storage and EBS-optimized by default and at no additional cost

INSTANCE TYPES

Instance Type	Use Case	Notable Features
P3, P2, G3, F1	General purpose GPU instances. Machine/Deep learning, high performance computing, 3D Visualizations, 3D Rendering, real-time video processing	Provide Enhanced Networking using Elastic Network Adapter with up to 25 Gbps of aggregate network bandwidth within a Placement Group. Up to 8 NVIDIA Tesla V100 GPUs, each pairing 5,120 CUDA Cores and 640 Tensor Cores. Supports FPGAs
H1, I3, D2	MapReduce-based workloads, distributed file systems such as HDFS and MapR-FS, network file systems, log or data-processing applications	High disk throughput, ENA enabled Enhanced Networking up to 25 Gbps, High Random I/O performance and High Sequential Read throughput, HDD storage

AMI VIRTUALIZATION TYPES

PARAVIRTUAL (PV)

- PV AMIs boot with a special boot loader called PV-GRUB, which starts the boot cycle and then chain loads the kernel specified in the menu.lst file on your image. Paravirtual guests can run on host hardware that does not have explicit support for virtualization, but they cannot take advantage of special hardware extensions such as enhanced networking or GPU processing.

HARDWARE VIRTUAL MACHINE (HVM)

- HVM AMIs are presented with a fully virtualized set of hardware and boot by executing the master boot record of the root block device of your image. This virtualization type provides the ability to run an operating system directly on top of a virtual machine without any modification, as if it were run on the bare-metal hardware. The Amazon EC2 host system emulates some or all of the underlying hardware that is presented to the guest.

ENHANCED NETWORKING

- Enhanced networking uses single root I/O virtualization (SR-IOV) to provide high-performance networking capabilities on supported instance types.
- SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization when compared to traditional virtualized network interfaces.
- Enhanced networking provides higher bandwidth, higher packet per second (PPS) performance, and consistently lower inter-instance latencies.
- There is no additional charge for using enhanced networking.
- Requires configuring the proper drivers on your operating system (Linux/Windows)
- Only supported for instances created in the Amazon VPC

AMAZON MACHINE IMAGES

- AMIs Detail the following information
  - Operating System
  - Initial State of Patching
  - Application or System Software
- Sources of AMIs
  - Published and released by AWS (Include Linux, RHEL, and Amazon Distro, Windows 08/12/12R2/16)
    - Results in the default OS Configuration (think installing from an ISO)
  - AWS Marketplace
    - Pre-Configured by Amazon partners or the community
  - Generated from Existing Instances
    - Created from your already existing instances, and replicated onto additional instances
  - Uploaded Virtual Servers
    - Uploaded from raw disk, or VHD/OVA images

NETWORKING ADDRESS TYPES

Address Type	Description
Public DNS Name	• On launch, a public DNS name is generated for the instance • Persists only while that instance is running and can not be transferred to another instance under any circumstances • Mainly used to SSH into the box from a remote host. Otherwise, Route53 would be setup with a domain name
Public IP	• On launch, a publicly routable IP address is generated for the instance • Persists only while that instance is running and can not be transferred to another instance under any circumstances • IP address is from AWS's CIDR block
Elastic IP	• Address unique on the internet that is reserved independently from AWS and associated with an EC2 instance • Persistent until released by the customer, and is not tied to the lifetime or state of a particular instance

SSH KEY PAIRS

- EC2 uses public-key cryptography to encrypt and decrypt login information
- Key pairs are created through the Management Console, CLI, or API
- Private key is only available for download at the creation of the pair, so be careful not to loose it
- Initial access to the instance is obtained using the EC2-User and the private key when logging on using SSH. After initial logon, other users can be configured or enrolled using LDAP, keyboard authentication, etc.
- For a **Windows** host, a randomly generated password is created and encrypted with the public key. The password is decrypted by the user using their private key, and entered into the RDP authentication prompt. User can then add local users or join to AD

VIRTUAL FIREWALL PROTECTION

- **Security Groups**- Act as a firewall for associated Amazon EC2 instances, controlling both inbound and outbound traffic at the instance level
- **Network access control lists** (ACLs)- Act as a firewall for associated subnets, controlling both inbound and outbound traffic at the subnet level
- When you launch an instance in a VPC, you can associate one or more security groups that you've created.
- Each instance in your VPC could belong to a different set of security groups.
- If you don't specify a security group when you launch an instance, the instance automatically belongs to the default security group for the VPC.
- You can secure your VPC instances using only security groups; however, you can add network ACLs as a second layer of defense.
- You can monitor the accepted and rejected IP traffic going to and from your instances by creating a **Flow Log for a VPC, subnet, or individual network interface.**
  - Flow log data is published to CloudWatch Logs, and can help you diagnose overly restrictive or overly permissive security group and network ACL rules.

SECURITY GROUPS

- You can specify allow rules, but **not deny rules**.
- You can specify separate rules for inbound and outbound traffic.
- When you create a security group, it has **no inbound rules**. Therefore, no inbound traffic originating from another host to your instance is allowed until you add inbound rules to the security group.
- By default, a security group includes an **outbound rule that allows all outbound traffic**. You can remove the rule and add outbound rules that allow specific outbound traffic only. If your security group has no outbound rules, no outbound traffic originating from your instance is allowed.
- Security groups are stateful — **if you send a request from your instance, the response traffic for that request is allowed to flow in regardless of inbound security group rules**. Responses to allowed inbound traffic are allowed to flow out, regardless of outbound rules.
- Instances associated with a security group can't talk to each other unless you add rules allowing it (exception: the default security group has these rules by default).
- Security groups are **associated with network interfaces**. After you launch an instance, you can change the security groups associated with the instance, which changes the security groups associated with the primary network interface (eth0). You can also change the security groups associated with any other network interface.

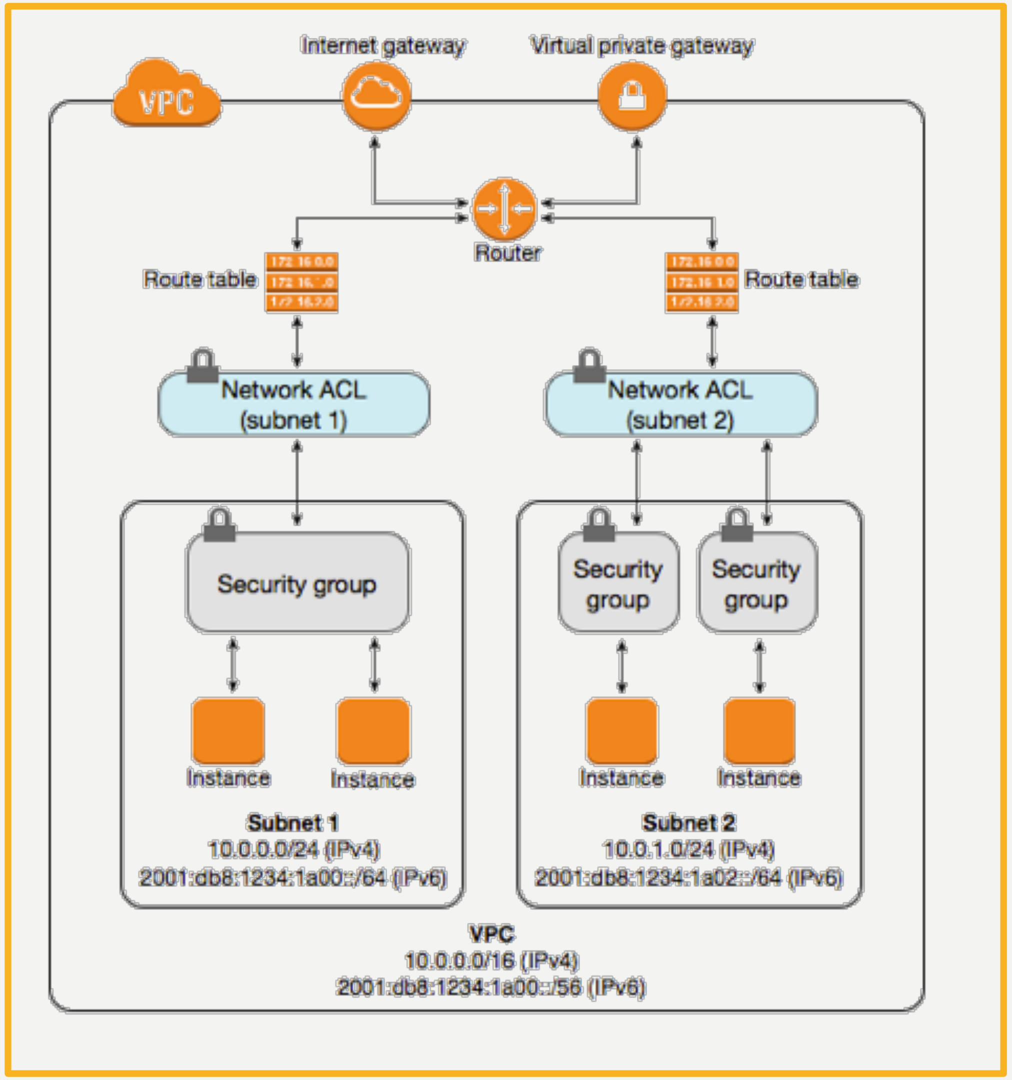
ACCESS SECURITY LISTS (ACL'S)

- Your VPC automatically comes with a modifiable default network ACL. By default, it allows all inbound and outbound IPv4 traffic and, if applicable, IPv6 traffic.
- You can create a custom network ACL and associate it with a subnet. By default, each custom network ACL denies all inbound and outbound traffic until you add rules.
- Each subnet in your VPC must be associated with a network ACL. If you don't explicitly associate a subnet with a network ACL, the subnet is automatically associated with the default network ACL.
- You can associate a network ACL with multiple subnets; however, **a subnet can be associated with only one network ACL at a time**. When you associate a network ACL with a subnet, the previous association is removed.
- A network ACL contains a **numbered list of rules that we evaluate in order, starting with the lowest numbered rule**, to determine whether traffic is allowed in or out of any subnet associated with the network ACL. The highest number that you can use for a rule is 32766. We recommend that you start by creating rules with **rule numbers that are multiples of 100**, so that you can insert new rules where you need to later on.
- A network ACL has **separate inbound and outbound rules**, and each rule can either allow or deny traffic.
- Network ACLs **are stateless; responses to allowed inbound traffic are subject to the rules for outbound traffic** (and vice versa).

SECURITY GROUPS & ACL'S

Security Group	Network ACL
Operates at the **instance** level (first layer of defense)	Operates at the **subnet** level (second layer of defense)
Supports **allow** rules only	Supports **allow** rules and **deny** rules
Is ***stateful***: Return traffic is automatically allowed, regardless of any rules	Is ***stateless***: Return traffic must be explicitly allowed by rules
We **evaluate all rules** before deciding whether to allow traffic	We process rules in **number order** when deciding whether to allow traffic
Applies to an instance **only if someone specifies the security group** when launching the instance, or associates the security group with the instance later on	Automatically applies to **all instances in the subnets** it's associated with (backup layer of defense, so you don't have to rely on someone specifying the security group)

SECURITY GROUPS & ACL'S



1. Traffic from an Internet gateway is routed to the appropriate subnet using the routes in the routing table.
2. The rules of the network ACL associated with the subnet control which traffic is allowed to the subnet.
3. The rules of the security group associated with an instance control which traffic is allowed to the instance.

NETWORKING LIMITS

* as of documentation available on March 7, 2018

Resource	Default	Comments
VPCs per region	5	• The limit for internet gateways per region is directly correlated to this one. Increasing this limit increases the limit on internet gateways per region by the same amount. • The number of VPCs in the region multiplied by the number of security groups per VPC cannot exceed 5000.
Subnets per VPC	200	n/a
Internet gateways per region	5	• This limit is directly correlated with the limit on VPCs per region. • To increase this limit, increase the limit on VPCs per region. • Only one internet gateway can be attached to a VPC at a time.
NAT gateways per Availability Zone	5	• A NAT gateway in the pending, active, or deleting state counts against your limit.
Network ACLs per VPC	200	• You can associate one network ACL to one or more subnets in a VPC. This limit is not the same as the number of rules per network ACL.
Rules per network ACL	20	• This is the one-way limit for a single network ACL, where the limit for ingress rules is 20, and the limit for egress rules is 20. This limit includes both IPv4 and IPv6 rules, and includes the default deny rules (rule number 32767 for IPv4 and 32768 for IPv6, or an asterisk * in the Amazon VPC console).
Security groups per VPC (per region)	500	• The number of VPCs in the region multiplied by the number of security groups per VPC cannot exceed 5000.
Inbound or outbound rules per security group	50	• You can have **50 inbound and 50 outbound rules per security group (giving a total of 100 combined inbound and outbound rules)**. To increase or decrease this limit, contact AWS Support — a limit change applies to both inbound and outbound rules. However, the limit for inbound or outbound rules per security group multiplied by the limit for security groups per network interface cannot exceed 250. For example, if you want to increase the limit to 100, we decrease your number of security groups per network interface to 2.

BOOTSTRAPPING

- Script virtual hardware management on instantiation
- Runs as soon as the instance starts up and is executed as part of the launch process in the “UserData” field
    - Linux: Bash Script
    - Windows: Batch or PowerShell
- Sample Use Cases:
    - Applying patches and updates
    - Enrolling in directory service
    - Installing application software
    - Copy information from S3/EBS to ephemeral storage
    - Installing Chef/Puppet or running Ansible
- UserData is stored with the instance and not encrypted. This is NOT a spot for credentials or secrets

INSTANCE METADATA

- Accessed by CURL'ing: **http://169.254.169.254/latest/meta-data/**
  - Remember this as ***169.254.0.0/16*** is an APIPA block (i.e. reserved for when your device cant acquire a DHCP address)
- Information about the instance that you are currently running
- You can also use instance metadata to access *user data* that you specified when launching your instance.
- Although you can only access instance metadata and user data from within the instance itself, the data is not protected by cryptographic methods.

```
[ec2-user ~]$ curl http://169.254.169.254/latest/meta-data/
        ami-id
        ami-launch-index
        ami-manifest-path
        block-device-mapping/
        hostname
        iam/
        instance-action
        instance-id
        instance-type
        local-hostname
        local-ipv4
        mac
        metrics/
        network/
        placement/
        profile
        public-hostname
        public-ipv4
        public-keys/
        reservation-id
        security-groups
        services/
```

INSTANCE TAGS

- Tags are unique identifiers that can be assigned to an instance or AWS resource for auditing, labeling, or other identification purpose
- Maximum of 10 tags per resource
- Simple key/value pairs

RESIZING AWS INSTANCE

- As your needs change, you might find that your instance is over-utilized (the instance type is too small) or under-utilized (the instance type is too large). If this is the case, you can change the size of your instance.
- For example, if your t2.micro instance is too small for its workload, you can change it to an m3.medium instance.
- When you resize an instance, you must select an instance type that is compatible with the configuration of the instance.
- If the instance type that you want is not compatible with the instance configuration you have, then you must migrate your application to a new instance with the instance type that you need.
- If the root device for your instance is an EBS volume, you can change the size of the instance simply by changing its instance type, which is known as *resizing* it. If the root device for your instance is an instance store volume, you must migrate your application to a new instance with the instance type that you need.
- You also need to check for compatibility issues, and ensure that the virtualization type, network, platform, etc. are all compatible

TERMINATION PROTECTION

- Termination protection allows the instance from being terminated from the AWS Management Console, CLI, or API
- This does not stop an instance from being terminated by an OS Shutdown command, auto scaling termination, or spot instance termination

INSTANCE PURCHASING OPTIONS

Instance Type	Description
On Demand Instances	Pay, by the second, for the instances that you launch.
Reserved Instances	Purchase, at a significant discount, instances that are always available, for a term from one to three years.
Scheduled Instances	Purchase instances that are always available on the specified recurring schedule, for a one-year term.
Spot Instances	Request unused EC2 instances, which can lower your Amazon EC2 costs significantly.
Dedicated Hosts	Pay for a physical host that is fully dedicated to running your instances, and bring your existing per-socket, per-core, or per-VM software licenses to reduce costs.
Dedicated Instances	Pay, by the hour, for instances that run on single-tenant hardware.

TENANCY OPTIONS

Tenancy Option	Description
Shared Tenancy	• Default model for all AWS instances. • Houses instances from different customers
Dedicated Instances	• Runs on hardware that is dedicated to a single customer/account. • Non-Dedicated instances in the account will run on shared tenancy and will be isolated at the hardware level from the dedicated instances in the account.
Dedicated Hosts	• Physical server with EC2 instance capacity fully dedicated to a single customer's use. • Complete control over which host runs an instance at launch • Differs from Dedicated Instance in that a Dedicated Instance can launch on any hardware that has been dedicated to the account

STORAGE TYPES

- Instance Stores/Ephemeral Storage
  - Provides temporary block storage for the instance
  - Used for buffers, cache, scratch data, temporary content, replicated data
  - Some provide Hard Disk Drive (HDD) and others have Solid State Drives (SSD)
  - Data is lost under the following circumstances:
    - Underlying disk drive fails
    - Instance stops (will persist a reboot)
    - Instance termination

ELASTIC BLOCK STORAGE (EBS)

- Provides block level storage volumes for use with EC2 instances.
- EBS volumes are highly available and reliable storage volumes that can be attached to any running instance that is in the same Availability Zone.
- EBS volumes which are attached to an EC2 instance are exposed as storage volumes that persist independently from the life of the instance. With Amazon EBS, you pay only for what you use.
- Recommended when data must be quickly accessible and requires long-term persistence.
- EBS volumes are particularly well-suited for use as the primary storage for file systems, databases, or for any applications that require fine granular updates and access to raw, unformatted, block-level storage.
- Amazon EBS is well suited to both database-style applications that rely on random reads and writes, and to throughput-intensive applications that perform long, continuous reads and writes.

EBS SECURITY AND ENCRYPTION

- You can launch your EBS volumes as encrypted volumes.
- Amazon EBS encryption offers you a simple encryption solution for your EBS volumes without the need for you to build, manage, and secure your own key management infrastructure.
- When you create an encrypted EBS volume and attach it to a supported instance type, data stored at rest on the volume, disk I/O, and snapshots created from the volume are all encrypted.
- The encryption occurs on the servers that hosts EC2 instances, providing encryption of data-in-transit from EC2 instances to EBS storage.
- Amazon EBS encryption uses AWS Key Management Service (AWS KMS) master keys when creating encrypted volumes and any snapshots created from your encrypted volumes.
- The first time you create an encrypted EBS volume in a region, a default master key is created for you automatically.

EC2 VOLUME TYPES

* as of documentation available on March 7, 2018

	Solid State Drives		Magnetic Drives	
Volume Type	General Purpose SSD (gp2)*	Provisioned IOPS SSD (io1)	Throughput Optimized HDD (st1)	Cold HDD (sc1)
Description	General purpose SSD volume that balances price and performance for a wide variety of workloads	Highest-performance SSD volume for mission-critical low-latency or high-throughput workloads	Low cost HDD volume designed for frequently accessed, throughput-intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads
API Name	gp2	io1	st1	sc1
Volume Size	1 GiB - 16 TiB	4 GiB - 16 TiB	500 GiB - 16 TiB	500 GiB - 16 TiB
Max. IOPS/Volume**	10,000	32,000	500	250
Max. Throughput/Volume	160 MiB/s	500 MiB/s***	500 MiB/s	250 MiB/s
Max. IOPS/Instance	80,000	80,000	80,000	80,000
Max. Throughput/Instance†	1,750 MiB/s	1,750 MiB/s	1,750 MiB/s	1,750 MiB/s
Dominant Performance Attribute	IOPS	IOPS	MiB/s	MiB/s
Use Cases	● Recommended for most workloads ● System boot volumes ● Virtual desktops ● Low-latency interactive apps ● Development and test environments	● Critical business applications that require sustained IOPS performance, or more than ***10,000 IOPS or 160 MiB/s of throughput per volume*** ● Large database workloads, such as: ○ MongoDB ○ Cassandra ○ Microsoft SQL Server ○ MySQL	● Streaming workloads requiring consistent, fast throughput at a low price ● Big data ● Data warehouses ● Log processing ● Cannot be a boot volume	● Throughput-oriented storage for large volumes of data that is infrequently accessed ● Scenarios where the lowest storage cost is important ● Cannot be a boot volume

EBS SNAPSHOTS

- Data for the snapshot is stored using S3 Technology
- The action of taking the snapshot is free, only the storage cost is incurred
- When you request a snapshot, the point-in-time snapshot is created and the volume may continue to be used, but the snapshot may remain in pending status until all of the modified blocks have been transferred to S3
- Snapshots are held in AWS-Controlled storage and not in your account's S3 Buckets
- Snapshots are constrained to the region where they were created, meaning you can use them to create new volumes only in the same region
- If you need to restore a snapshot in a different region, you can copy the snapshot to another region
- When restoring data from a snapshot, the volume is created immediately, but the data is loaded lazily, meaning that the volume can be accessed upon creation, and if the data being requested has not been restored, it will be restored upon first request.

QUESTIONS?