Elastic Cloud Compute

EC2 – Elastic Cloud Compute

Terms:

- **EC2 Instances** – Virtual Computing Environments
- **AMI** (Amazon Machine Images) – preconfigured EC2 templates for your instances. ie. can be SQL server, Bastion, NAT instances etc
- **Instance Types** – Processing power of your EC2 configurations. ie. CPU, memory, storage, network capacity are separated as Instance Types
- **Key pairs** – Secure login information for your instances
- **EBS (Amazon Elastic Block Store) –** persistent storage volumes for your EC2
- **Region/AZ (Availability Zone) –** Various physical locations of your resources
- **Elastic IP Address** – static IP address for dynamic cloud computing

Instances Types:

to remember all the types easily here is an acronym: DRMCGFTPX (**think: Doctor Mac Gift Pix, credit to acloudguru for acronym**)

D(ense disk) **R**(AM memory-intensive) **M**(icro) **C**(ompute for processsing) **G**(raphics) **I**(nput/Output) **F**(pga or field programmable arrays) **T**(2 micro) **P**(ics) **X**(treme)

Pricing Model:

- **On – Demand** :pay only what you request and what you use, no up-front fees
- **Spot :** flexible rates, you bid for what price you are willing to pay, somewhat sort of stock market for ec2 instances
- **Reserved :** Pay upfront for a period you want to pay for your instance, and in return you get a significant discount depending on which you are signing up for. (longer period and more instance = more discount)
  - **Payment options:** All Upfront, Partial Upfront, and No Upfront.
  - **Can I move a reserved instance from one region to another? No**

Access to EC2 can be the following:

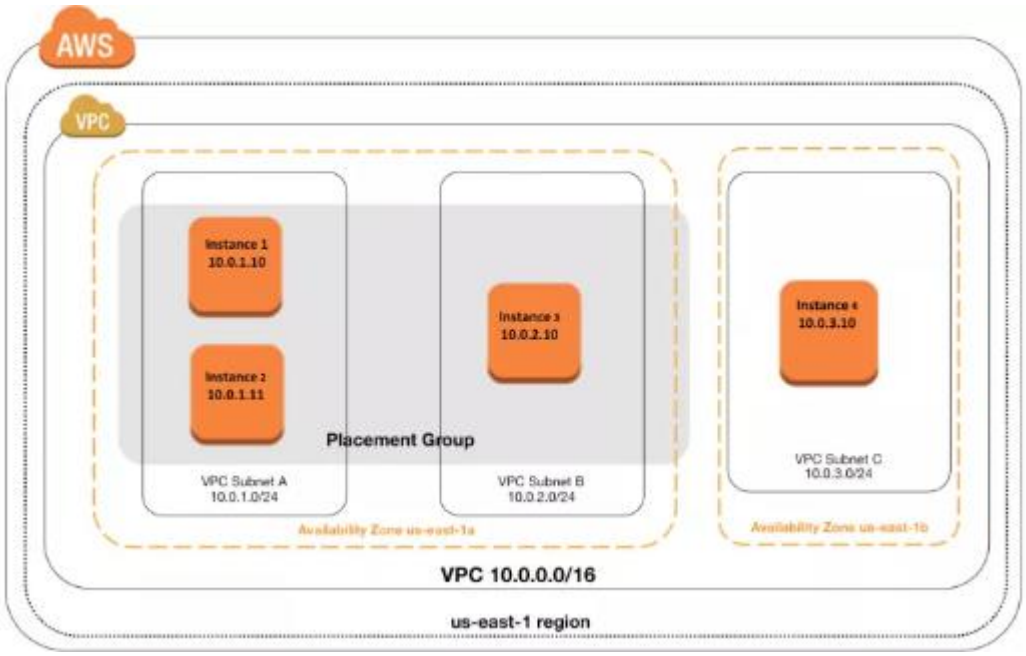
- EC2 Console
- AWS Web API Call
- AWS SDK
- AWS Command Line interface

Amazon Machine Image (AMI)

- Preconfigured Virtual servers
- auto-assign public IP
- Termination protection off by default (you can delete the image by clicking on delete action if termination protection is turned on)

- Root device by default is not encrypted (there are ways to encrypt it)
- EBS backed – default is delete the volume with instance

Placements Groups – AWS feature that enables EC2 instances (usually of similar instance types) to communicate with each other if in the same Availability Zone with high bandwith and low latency. Restrictions: Unique per AWS account, can't move existing EC2 to placement groups, one AZ per group.



Volume Storage: Will be discussed in another entry

Additional Features to be noted

CloudWatch

- Enables performance monitoring
- set on EC3 creation
- Standard config gives you statuse very 5 mins vs Detailed config gives you per minute but with additional charges
- Has Dashboards, Events, Alarms, Logs

AutoScaling

- Enables you to scale your EC2, to increase/decrease in number depending on load and availability
- It checks health status via Elastic Load Balancer

Elastic Load Balancer

- Health check for your instances
- Gives you ability for Cross Zone load balancing
- No Ip given, only DNS

Security Group

- Gives you ability to control inbound and outbound traffic , ie enable port 8080, 1433 (sql), 443 (https), 22 (SSH)
- **Stateful by default**, if you enable inbound port 8080 it is immediately given outbound as well

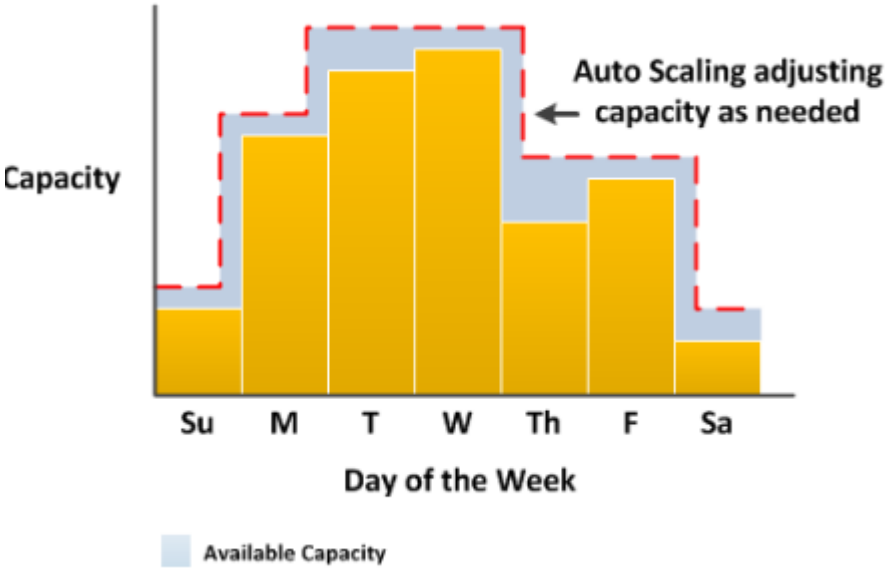
Auto Scaling

Autoscaling – ensures you have correct number of Ec2 to handle load of your application, whether to scale up or scaled down. AutoScaling groups are the cornerstone of any self-healing application on AWS.

Auto scaling ***is not really intended to respond to instantaneous spikes in traffic***, as it will take some time to spin-up the instances that will handle the additional traffic.

Pricing – Autoscaling is FREE

Benefits: you can scale your resource depending on the demand of your application, hence you only pay for what you actually use, instead for example allocating too much memory on your servers but peak usage only occurs on certain days, as per below:



Limitation

- Autoscale can only be in on one region, but can have multiple Availability zones
- Autoscale does rebalancing activities to make sure scaling is balanced among AZs this occurs under the following conditions
  - you issue a request to change AZ for your group
  - explicit call for termination of instance
  - AZ had insufficient capacity recovers
  - Spot market price
- Autoscale launches new instances before termination old ones

Scale Out Occurances:

- Manual Scaling
- Dynamic Scaling
- Scheduled Scaling

Scale In Occurances:

- Manually decrease size
- create scaling policy to decrease size depending on demand
- scheduled scale

Autoscaling cooldowns – ensures that auto scaling does not launch or terminate additional instance before previous scaling takes effect. (does not apply if instance becomes unhealthy). Default cooldown period is **300 seconds**

Cooldown periods are not supported for step scaling policies or scheduled scaling.

Default termination policy

- Check if there are any instances in multiple AZ
  1. if Yes select AZ with most instance
  2. if No select instance with oldest config – TERMINATE
- Are there multiple instance with oldest config
  1. if yes select instance with most closest to billing hour
  2. if No Terminate
- Are there multiple instance closest to billing hour
  1. if yes select at random then terminate
  2. if no terminate

Possible termination policies:

- OldestInstance
- newInstance
- oldestLaunchConfiguration
- closestToNextInstanceHour
- Default

Instance Protection – protects an instance from getting deleted

Lifecycle Hooks – events you want to trigger when autoscaling occurs, ie create notification or perform some lambda function

Standby State – you can put an instance on standby state, meaning it is still part of the scaling group but it does not handle application, this is useful for example if you want to upgrade application and just put the instance on standby while others handle the traffic

RDS

Amazon Relational Database Service

Services:

- RDS Databases covered:
  - Maria
  - Oracle
  - SQL Server
  - Postgre SQL
  - Aurora
  - MySQL
- Dynamo DB – NO SQL databases
  - MemCache
  - Redis
- Redshift – Data Warehousing
- Database Migration Service (DMS)

Types of Processing:

- Online Transaction Processing (OLTP)
  - Standard processing
  - Transactional data ie get one order detail
- Online Analytics process (OLAP)
  - more complex computation
  - query is done on a copy as to not disturb production

Database Backups:

- Automated Backup
  - Retention period – default is 7 days, up to 35 days
  - saved on S3
  - on by default
- DB-Snapshots – manually turned on
- Encryption at rest supported using KMS
- Encrypting existing DB currently not supported
- In RDS, changes to the backup window take effect Immediately

Multi Availability Zone Replication – replicate DB to another AZ for DR not for performance, this enables auto failover from geographic location to another.

Exam NOTES:

- Dynamo DB can scale on the fly, RDS cannot.
- RDS can scale on read not on write (Read Replica)
- You CANNOT RDP or SSH to RDS instance
- At the present time, encrypting an existing DB Instance is not supported. To use Amazon RDS encryption for an existing database, create a new DB Instance with encryption enabled and migrate your data into it. (manually import data)

Read Replica on RDS:

- Replicate RDS on different RDS
- Async replication is done
- Replicas are only read copies, and each EC2 instance can connect to a replica
- Automated backup is required for read replica
- replicas can have replicas (this will have performance implications)
- NO multi AZ
- replicas can be promoted to their own DBs

Dynamo DB features **– No SQL services from AWS**

- consistent latency on scale
- stored on SSD (store sdata in partition)
- 3 geographic locations
- Eventual consistency (1 second), strong consistent read (
- cheaper on read than write
- push button scaling
- DynamoDB is automatically redundant across multiple availability zones.
- terms:
  - **Tables** – similar to RDS
  - **Items** – similar to columns, items is a group of attributes that represent one object in a table, ie a person table, each item would represent one person
    - 400 KB limit on item size
  - **Attributes** – each item are composed attribute ie. FirstName, LastName, Age etc
- **Primary Key**
  - **Partition Key –** simple PK hash function to identify items
  - **Partition key and sort key** – composite pk which parition key as above and sort key which is sorted by sort key value
- API : Control plane (CRUD tables), Data plane (CRUD Data), DynamoDb Streams
- Data Types – Scalar (number, String, binary, boolean, null) Document (JSON document [list/map]), Set Type (String set, number set, binary set)
- **Read Consistency**
  - Eventually Consistent Reads
  - Strongly Consistent Reads
- **Provisioned throughput: 1 read unit = 4kb, 1 write unit 1kb**
- **Batch Operations**: you can get up to 16mb of data, which can be as many as 100 items
- **Conditional Writes** – prevents concurrent overwrite of same row by using **ReturnConsumedCapacity**
- **Scan –** as opposed to query, scan returns all of the data attributes for every item in a table or index, maximum of 1mb
  - Scan uses sequential scanning, you can use parellel scan by using segments for faster retrieval of large amounts of data

Redshift – Data warehousing mechanism from AWS

- Typical database block sizes range from 2 KB to **32 KB**. Amazon Redshift uses a block size of **1 MB**
- **Single node** can be up to **160gb**
- **multinodes** can contain up to **128 nodes**
  - **leader node**– receives query
  - **compute node** – performs query

- Node slices – nodes are partitioned into slices, each slice a portion of memory and disk space is allocated to do the workload

Performance:

- Columnar Data Storage – column based storage (RDS is row based)
- Can handle large data set with fewer I/O
- Advanced Data Compression – 10 times faster than RDS
  - columns are all the same type hence it is faster than RDS
  - no indexes/views
- Massive Parallel Processing (MPP)
- Query optimizer
- Compiled Code – leader node compiles the code already eliminating overhead of interpreting code and distributes across all nodes

ElastiCache **– Caching mechanism of AWS**

Features:

- Automatic detection and recovery from cache node failures.
- Automatic failover (Multi-AZ) of a failed primary cluster to a read replica in Redis replication groups.
- Flexible Availability Zone placement of nodes and clusters.
- Integration with other AWS services such as Amazon EC2, CloudWatch, CloudTrail, and Amazon SNS to provide a secure, high-performance, managed in-memory caching solution.

Engines

- MemCached
  - memory object caching
  - not AZ
- Redis
  - key/value pair storing
  - sets/list
  - master – slave replicatin
  - AZ

AuroraDB – Amazon created SQL, this is always recommended when using AWS cloud computing database

- scaling capability
- **3 Az with 2 copies – total of 6 copies stored by default (highly durable)**
- replication is done immediately and also Free

Simple Storage Service

S3 (Simple Storage Service)

- Object based storage on cloud (meaning you cannot install any programs, you use AWS EBS for installing programs in storage)
- **up to 5 terabytes PER one object**
  - no limit on all objects
- files are stored in "buckets"
- uses universal namespace for buckets, i.e bucket name should be unique throughout whole AWS
- Uses Rest webservice for API calls
- 100 buckets soft limit (you can request more to AWS)
- Static web hosting bucket naming convention for URL: **https://.s3-website-.amazonaws.com**
- Event notification subscriber: SNS, SES, Lambda

Access

- Read after write for **PUTS** (you can immediately access whatever file you upload in s3 buckets)
- Eventual Consistency for **UPDATE** (overwrite **PUTS**) and **DELETE** (it may take a while before the change propagates throughout different regions)

Storage Details/Capabilities

- Key Value mechanism for objects
- Versioning control (AWS keeps tracks of previous version of object, given that you have enabled versioning for your bucket)
- Cross Region Replication – buckets can be replicated accross different regions
- Metadata is also stored (data about your data)
- Access control can be configured for s3
- *Lifecycle Management*: frequently accessed files can be stored in standard bucket, moved to infrequent access if not accessed by any user for a few months , and finally moved to glacier if not accessed by any user for a number of years. Lifecycle management allows you to tweak when to move files from certain bucket to another. below are the types of buckets in detail.

Types

- **S3 Standard**
  - 99.99% availability
  - 99.999999999% Durability (eleven 9s)
  - Tiered storage available
  - Lifecycle Management (You can move to archive if files have not been accessed a long time)
  - Versioning
  - Encryption (Encryption at transit: SSL, Encryption at rest: AES-256, different ways of encryption detailed below)
  - Secure Data Access
- **S3 Infrequent Access**
  - Lower costs than standard
  - virtually the same except you have only 99.9% availability
- **Glacier**
  - **40 Terabytes per individual archives**

    - No limit to all
- Archiving option of AWS
- Extremely low cost
- same durability
- takes 3-6 hours before you can access any files

From Amazon comparison of bucket types:

From Amazon comparison of bucket types:

	Standard	Standard - IA	Amazon Glacier
Designed for Durability	99.999999999%	99.999999999%	99.999999999%
Designed for Availability	99.99%	99.9%	N/A
Availability SLA	99.9%	99%	N/A
Minimum Object Size	N/A	128KB*	N/A
Minimum Storage Duration	N/A	30 days	90 days
Retrieval Fee	N/A	per GB retrieved	per GB retrieved**
First Byte Latency	milliseconds	milliseconds	select minutes or hours***
Storage Class	object level	object level	object level
Lifecycle Transitions	yes	yes	yes

Additional Option for bucket types:

Reduced Redundancy Storage: further cheaper version of your buckets,reduces the durabiliy to 99.99% instead of 99.9999999999%, this is for files that can be lost, i.e thumbnails of your pictures that can be regenerated anyways

Important Note: Buckets MUST BE LOWER CAPS

Static website hosting: buckets can be used to host static website (no javascript, no server side scripting, no processing just plain html files). The benefits of using buckets as static website hosting is for temporary web pages, like for example creating a temporary web page for a movie poster that will be displayed for a few days only. Buckets will then handle your auto-scaling, load-balancing should there be a lot of traffic to visiting your site and you wouldn't need to worry about your site going down.

Security:

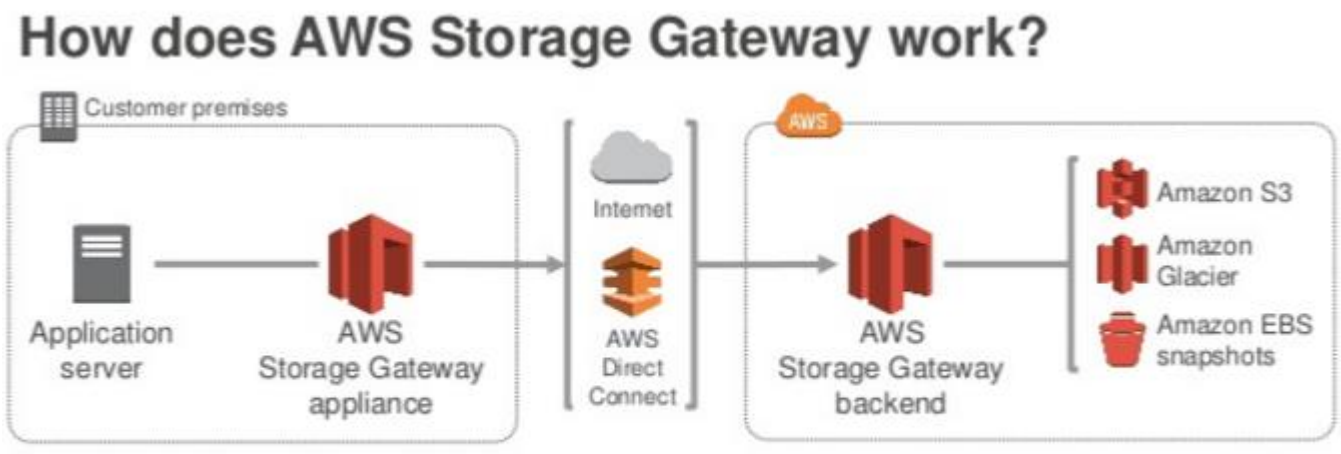
Transit : SSL/TLS

At Rest:

- Server Side Encryption (SSE)
- SSE-S3 – S3 Key – AWS S3 has the master key
- **SSE-C – Customer provided Key – Customer is you, ie, you want to have your own keys, not your clients' keys.**
- Client Side Encryption

Storage Gateway

Amazon Storage Gateway provides on-premise storage infrastructure for clients.



Types:

- Gateway Stored Volume – files are stored on site, asynchronously uploads backup to AWS
- Gateway Cached Volume – frequent accessed files are stored on local, while others are stored in AWS, there is no access to those files if you are connecting via internet and you lose internet connection (best to use Direct Connect in this scenario)
- Gateway Virtual Tape – for replacing physical tapes, enables you to upload data from virtual tapes to AWS.

Amazon S3 Import/Export – enables high speed upload/download of data transfer using Amazon high speed internal network.

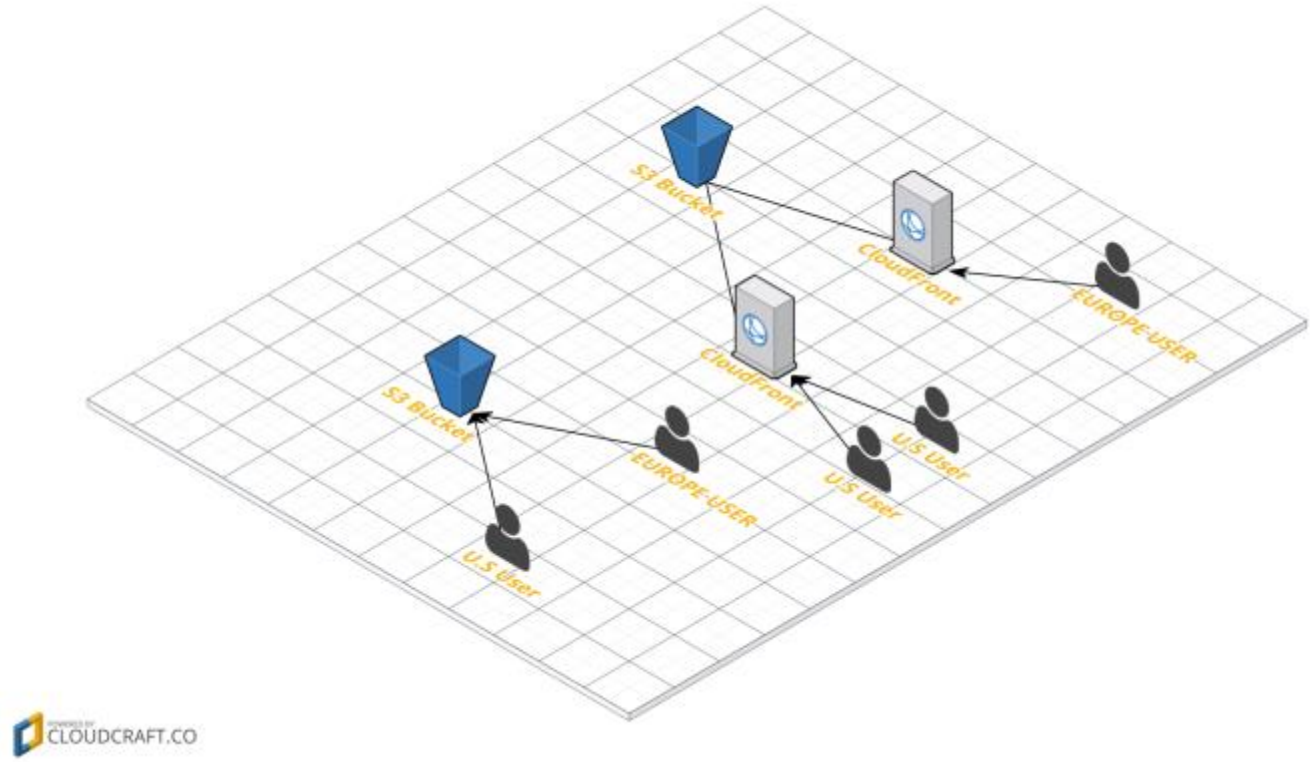
Capabilities:

- Import/Export to Disk
- Import to S3
- Import to Glacier
- Export to S3

Amazon Import/Export Snowball – Import/Export to S3 via Amazon physical device. It is much faster than specified above, but will require you to purchase/rent Amazon Snowball device. It is usually recommended by Amazon to use this when importing or exporting large amount of data.

CloudFront

AWS CloudFront – is a cloud global content delivery network (CDN) designed to accelerate delivery of your content to users by caching data to an AWS location (called edge location) closer to your user. i.e if your S3 bucket is located in Europe, it would take a lot longer for U.S customers to access these files. CloudFront enables you to create a cache location of your S3 files:



It is now possible to expedite uploads to S3 by writing directly to an Edge Location.

Updating Cache Data on Cloudfront – While the first 1000 invalidation paths per month are free, additional invalidation paths are $0.005 per request.

CDN Sources:

- S3 Bucket
- Elastic Load Balancer
- Route 53
- EC2
- Non-AWS resources

Capabilities/Features

- RMTP (Real-Time Messaging Protocol) Distribution
- Web Distribution
- Configure TTL (Time To Live) of your objects in cloudfront before they are refreshed again
- Restrict Viewer Access

NOTE: CloudFront PCI DSS Compliance – allows processing, storage and transmission of credit card.

Identity Access Management

IAM (Identity and Access Management)

Amazon service that enables you to do the following:

- Create users
- Manage users and their access
- Create Federated User (Temporary Users)
- Free of charge

IAM User Management

- Create, Delete, List Users
- Manage group memberships, credentials permissions
- **default 100 groups limit, 5000 users limit**

Users

- in this context, Users are individual people that have access to AWS
- Users are global and not region specific

Groups

- Collection of users
- Users can belong to multiple groups (**10 default limit**)
- Groups cannot be nested, i.e Groups cannot be assigned to other groups

Roles

- "An IAM *role* is similar to a user, in that it is an AWS identity with permission policies that determine what the identity can and cannot do in AWS. However, instead of being uniquely associated with one person"
- **250 default roles limit**
- Roles can be assigned to other AWS accounts
  - Power User role – all access except group management
  - Administrator Access role – All account resources except AWS account info

IAM Owner– is the one who created the AWS account

Policies

- JSON format rules that define access
- Policies can be attached to roles/groups/users

Policy Simulator – enables you to test out your policies, AWS provided service free of charge. http://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies_testing-policies.html

Multifactor Authentication **–** additional layer of authentication other than just password, this can be any third party device, virtual authenticator, STS (Security Token Service), SMS authentication.

- When user is trying to login, a code will be sent to his MFA device, he then needs to input the code after providing his password, this ensures another layer of security, and your access is not compromised easily should someone finds out about your password.
- This can be enforced in API calls for developers when calling sensitive API calls to AWS.

Identity Federation – allows third party accounts to login to your AWS, i.e using LDAP, facebook, google, etc, no need to create AWS IAM user

NOTE: Active Directory authentication is possible in AWS through SAML, authentication is done first in AD before being passed to AWS

BEST PRACTICES:

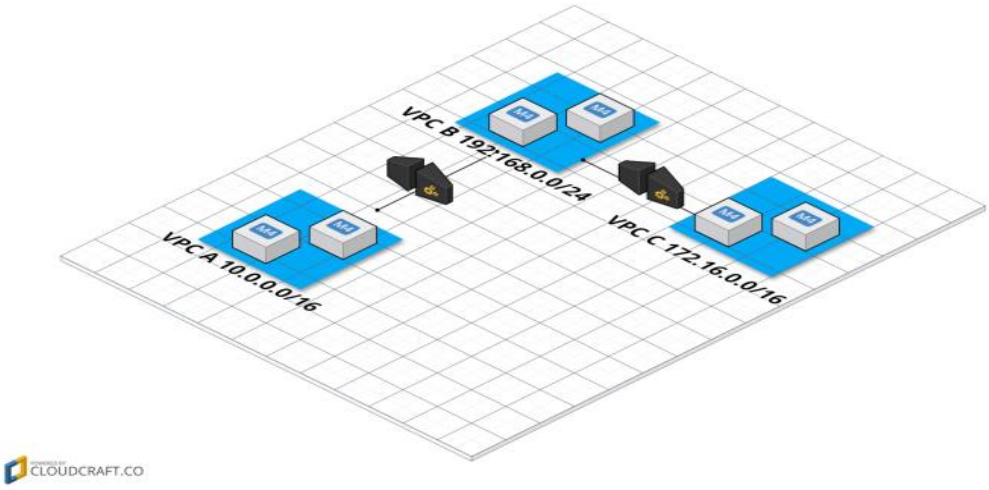
- *Root/privileged users should have MFA*
- Grant only least access privileges
- Each users should have individual IAMs (not sharing accounts)
- Use Groups for managing users
- enforce a strong password policy
- assign IAM roles for your applications
- Rotate credentials
- Track what users are doing with cloudtrail (audit log service for AWS)

VPC Peering

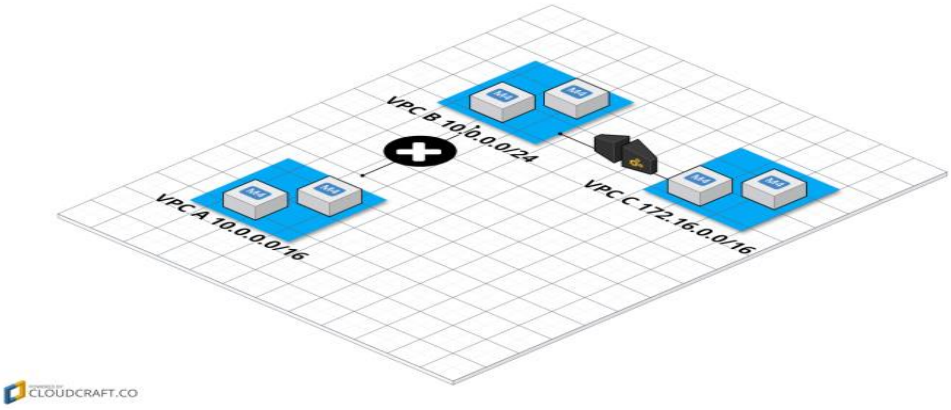
What is **VPC Peering**?

- Connection between two VPCs (single or with other AWS Account) within a single region.
- This is done via private IP address.
- Technology used is existing infrastructure of VPC, it is neither a gateway or a VPN connection.

Example VPC Peering:



Transitive Peering NOT Supported – VPC A cannot access VPC C via VPC B



If VPC B is change to this cidr block as above, it breaks the connection as there is overlapping internal address range (CIDR block).

Limitations:

- no overlapping CIDR blocks
- no peering connections
- cannot be on different regions

Virtual Private Cloud

VPC – Virtual Private Cloud, is your network configuration for your AWS resources

- public subnet by default means you have one subnet in your route tables whose target is a Internet Gateway enabling access to the public
- on there other hand private subnet for you private resources that shouldn't be available to public

VPC Wizard:

- can opt for vpc with public/private subnet/ and with Hardware VPN accese

VPC Security Groups

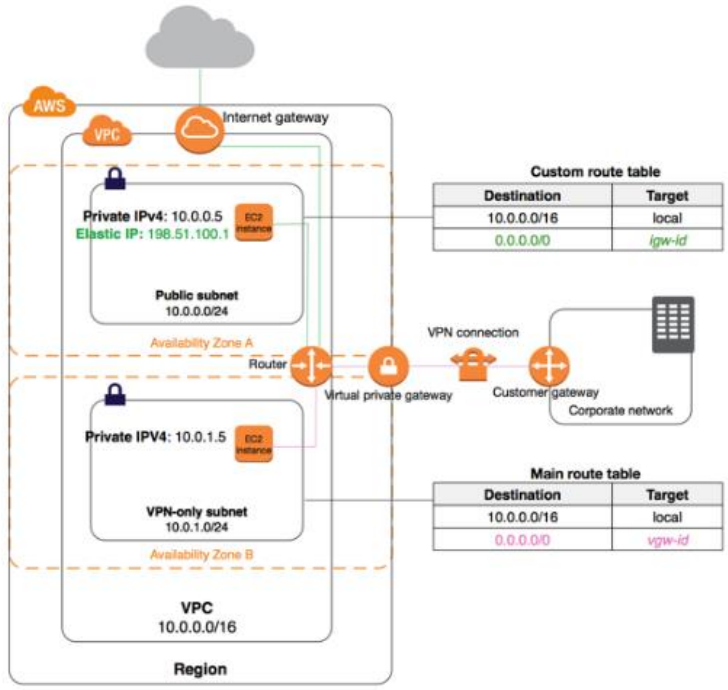
- Configuration of inbound and outbound traffic for your resources
- All outbound traffic by default is allowed when you create a new one
- Security groups is stateful – if you enable an inbound traffic, traffic will also flow outbound regardless of the security group
- You can specify allow rules, but not deny rules.

VPC Network ACL (Access Control List)

- By default allows all inbound and outbound
- ACL is stateless – contrast to Security group
- Associated to subnet
- Evaluated per order

Route Table

A routing table is a set of rules, often viewed in table format, that is used to determine where data packets traveling over an Internet Protocol (IP) network will be directed. All IP-enabled devices, including routers and switches, use routing tables.



NETWORK ACL vs Security Groups

Security groups filters which traffic goes in and out to your instances, ie which ports can be access, in contrast Network ACL operates at a subnet level, ie which IP can SSH to your instance

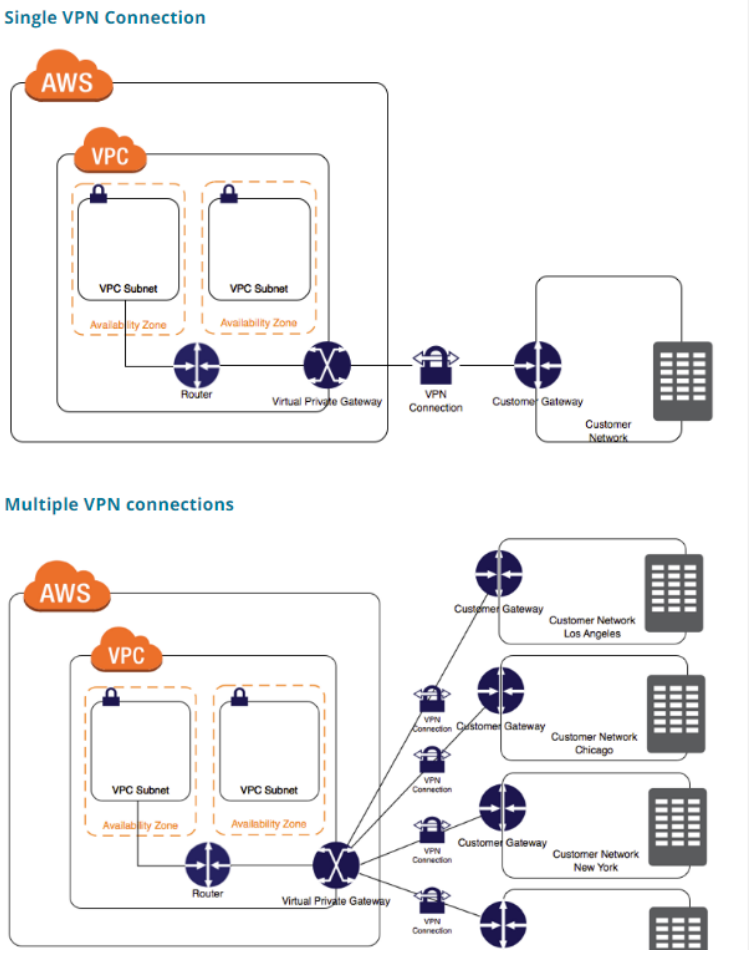
- Network ACL – applies to network level, that it can apply to many instances
- Security Group – applies only to instance level, so if you need it in one instance only may need to apply to Security group only

VPN Connections

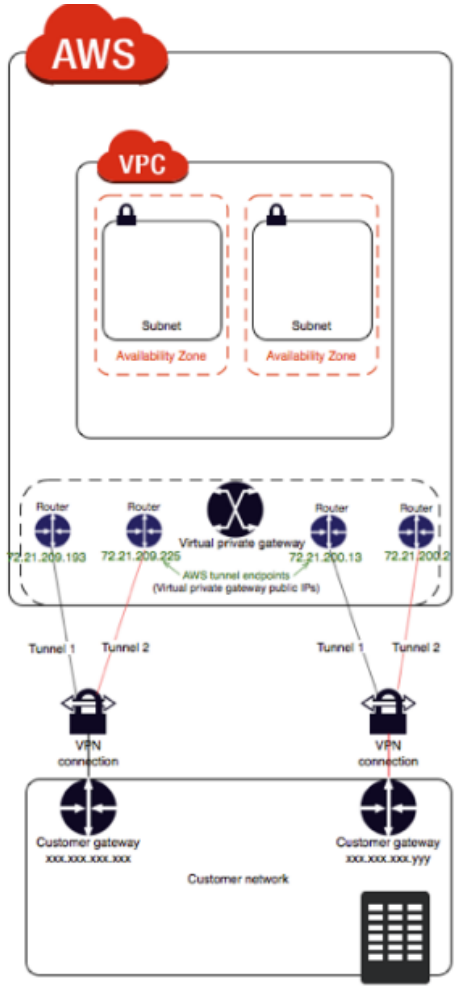
- **By default, instances that you launch into a virtual private cloud (VPC) can't communicate with your own network. You can enable access to your network from your VPC by attaching a virtual private gateway to the VPC**
- AWS Hardware VPN – enables customers to connect between VPC and remote network
- AWS Direct Connect – provides private connection to AWS without internet
- CloudHub – more than one remote network
- Software VPN – can be setup via software VPN

Components

- Virtual Private Gateway – VPN connector on AWS side
- Customer Gateway – customer side

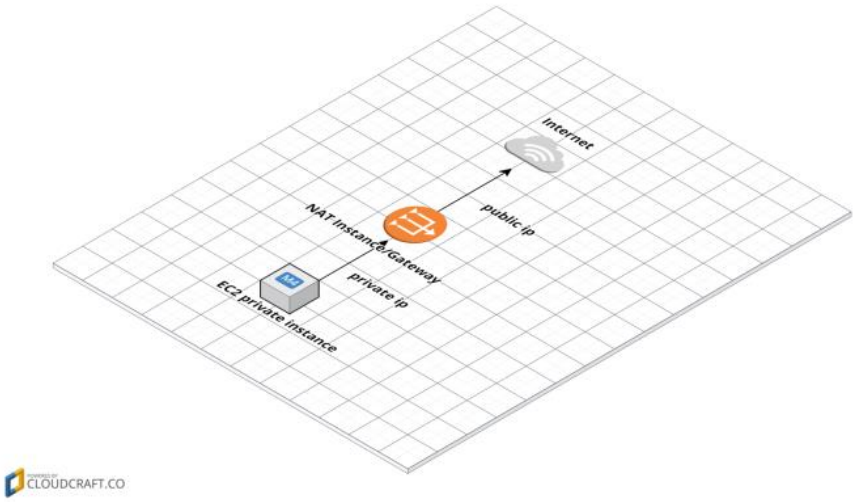


Redundancy:



- Route Tables – determine where traffic is determined
- **NAT – Network Address Translation**

- Enables private instances to connect to the internet, but prevents the internet from initiating connection with instances



private IP is replaced with NATs public IP

NAT Key points

- needs to be launched in public subnet
- needs to be assocaited with public/elastic ip address
- disable source/destination flag check – this flag directly conflics with how NAT works as per above
- Security group should allow inbound/outbound
- Route table should configured to have an internet gateway

NAT Gateway vs NAT Instance

NAT Gateway	NAT Instance
AWS managed instance	User created instance, configured to be NAT
10 gbps burst	availability and bandwith depends on the instance type
no Security Group	must have security group
one elastice IP address associated	manually disable source/destination check
specific AZ, with redundancy TCP, UDP, ICMP support ports – 1024- 65535 cannot send through VPC endpoints	

High Availabiltiy NAT Instance design:

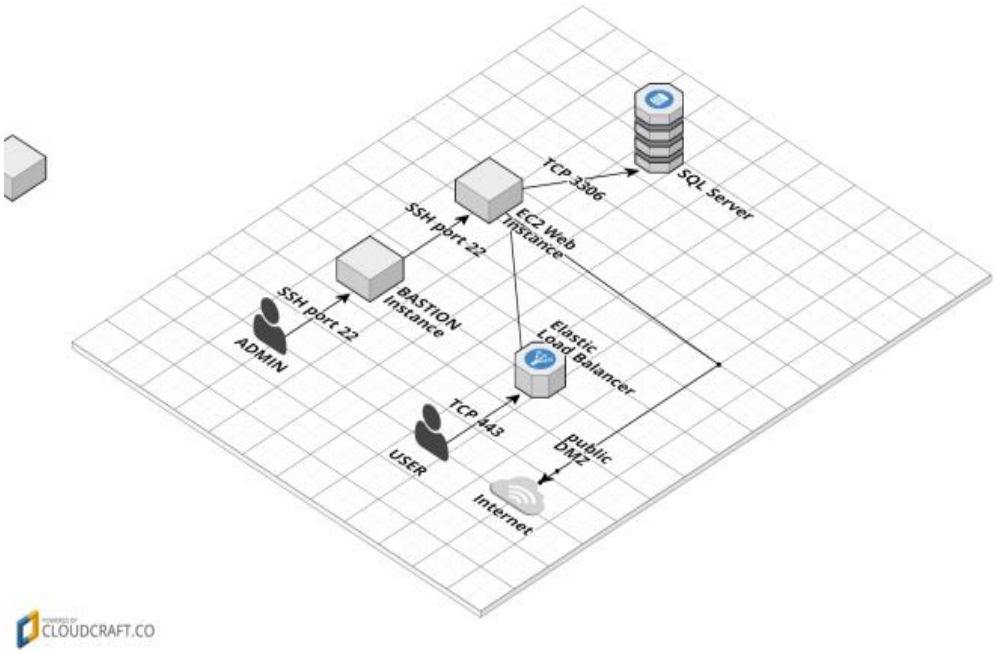
- one NAT instance per AZ
- all private subnet route teables to the same zone NAT instance
- configure AutoScaling for instances
- have it grow if CPU reaches a certain threshold
- create bootstrap scripts for updating NAT instances

BASTION

- structure for fortication to protect things behind it
- in AWS also known as a Jump Server
- used to access instances in private subnet

How it works:

There is no direct access available to connect to your Web Server, you would have to SSH to bastion instance first, users access your server through load balancer.



Route 53 – AWS DNS Server

Supports the following

- NS (name server)
- CNAME (canonical name record)
- SRV (service locator)

Terms

- Domain name – URL for website
- Domain Registrar – company accreditd by ICANN to process domain registration
- Domain Registry – company owns right to sell domains
  - **Route 53 can act as a domain Registry top register domains**
- Domain Reseller – company sessl domain names for registrars
- Top-level Domain – the last part of domain name (.com, .org)
  - generic top level – such as .com, .hockey, .bike
  - geographic – associated with countries (.ph, .au)

Domain Name System Terms

- Alias resource record set – record set you can create with route 53 to route traffic to AWS resource
- Authoritative name server – Name server that response to request from a DNS resolver
- DNS Query – query submitted by devices to DNS Server
- DNS resolver – DNS server managed by internet service provider

- Domain Name System – DNS, worldwide network of servers that enable translation of URLs to IP addresses
- Hosted Zone – container for resource record sets (contains how to route trafic for domain and subdomain
- IP Address- number assigned to device on the internet (laptop, phones etc)
  - IPv4
  - IPv6
- Name servers – Servers in DNS
- private DNS – local version of DNS
- TTL – Time to Live

Simple Queue Service

SQS – Simple Queue Service

Message Size – 256kb

Message Attributes:

- Name
- Type (String/Binary/Number)
- Value – user value

URL : The following is the queue URL for a queue named MyQueueowned by a user with the AWS account number 123456789012.

```
http://sqs.us-east-2.amazonaws.com/123456789012/MyQueue
```

Key Terms:

- **Receipt Handle** – each time you recieve a message you get a different receipt handle, this is used to handle when you request to delete a message
- **Message Deduplication Id** – the token used for deduplication of sent messages, any messages sent with same aren't delivered during 5 minute deduplication interval
- **Message Group Id** – group identifier
- **Sequence number**
- **Inflight Message –** message received but not yet deleted
- **Dead Letter Queue –** Queue for others queues to target for messages that can't be processed, this allow messages to be isolated and help determine why it wasn't procesed.

Two Types of Queues:

- Standard
  - high throughput
  - guaranteed at least once delivery (can be more than once)
- FIFO
  - ordered messages
  - first in first out delivery

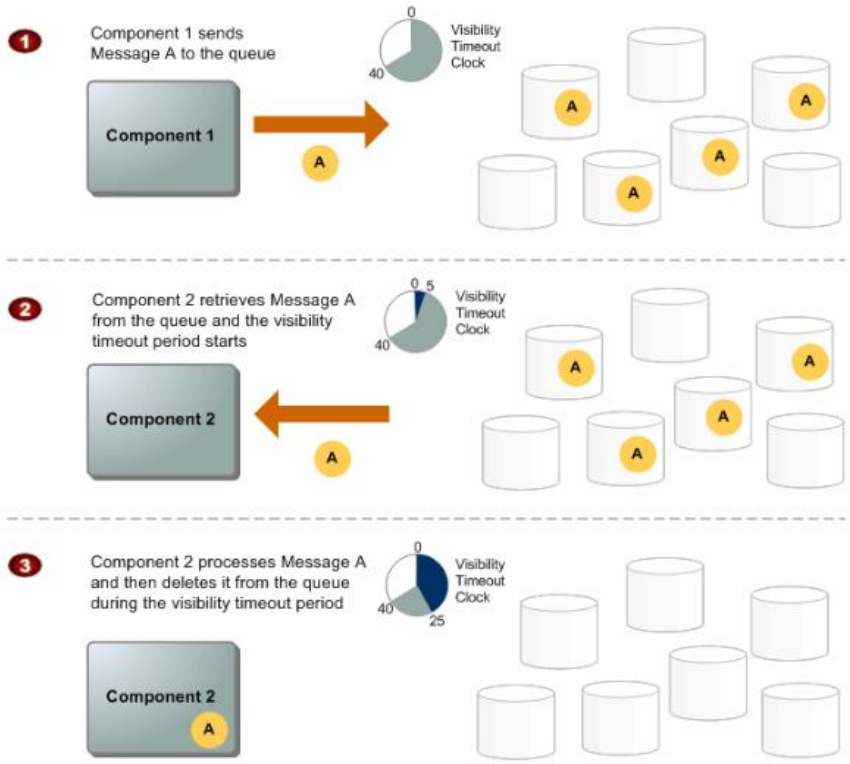
Visibility Timeout (default 30 seconds, Maximum: 12 hours.)

When a consumer receives and processes a message from a queue, it remains in the queue AWS does not automatically delete the message, consumer must request a delete. This is because queue service are in a distributed system there is no guarantee the message has been received already (ie connection broken, component fail etc)

A message once received, to prevent other consumer to reprocess the message, there is a visibility timeout, which a message is not visible to other components. (not a guarantee for standard queues)

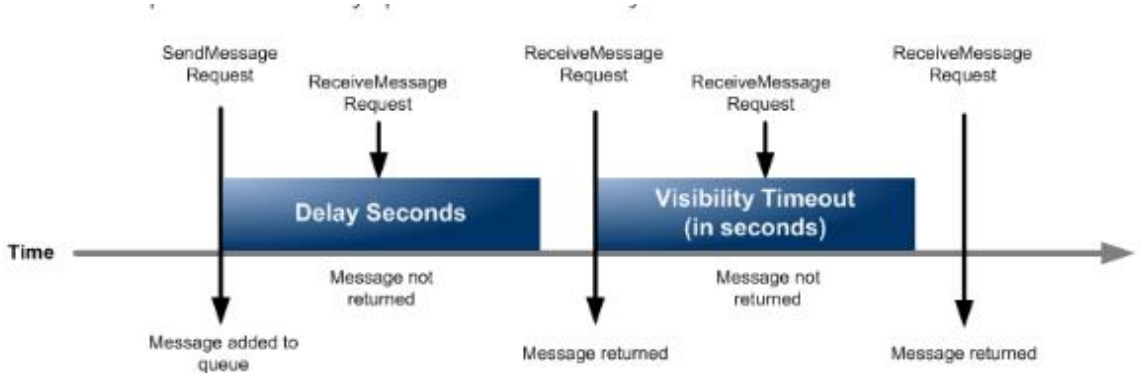
Increasing the visibility timeout will decrease cost over time.

You can turn off visibility timeout
How it works:



Delay Queues

Works similar like visibility timeout, except message is invisible when first added to queue:



Long Polling

Long polling reduces number of empty responses by allowing SQS to wait until a message is available before sending a response. it returns a message as soon as any messages becomes available.

By default SQS uses short polling, querying only a subset of the servers to determine whether any messages are available. Short Polling occurs when WaitTimeSeconds parameter of a message is set to 0.

Short polling may fail to retrieve messages sometimes, but if no messages can be retrieved after multiple attempts, permissions are the more likely cause.

Billing

1 million messages free per month under free tier

Simple Notification Service

SNS – Simple Notification Service, allows delivery or sending of message to subscribing endpoints.

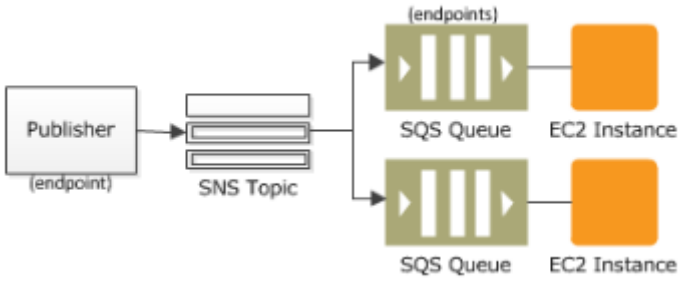
Two Types of clients:

- Publisher – publishes the message to a topic
- Subscriber/Consumer – consumes the message
  - Lambda
  - SQS
  - HTTP/S
  - Email
  - SMS

Publisher -> Amazon SNS Topic -> Subscriber

Common scenario: Fanout

Publisher receives for example an order for some product, it is sent to Topic, multiple SQS queues consumes the message and is allocated to seperate instances for parallel processing



Other uses:

- Application Alerts
- Push Email/Text Messaging
- Mobile Push Notification

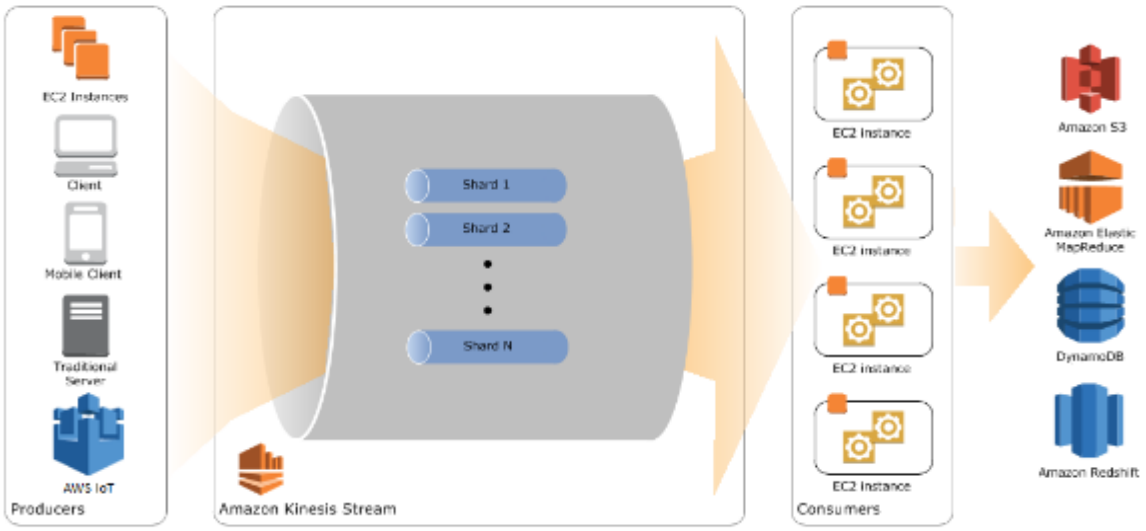
Kinesis

Kinesis enables to collect and process streams of data records in real time.

What you can you do:

- accelerated log and data feed intake and processing -ie large amount of applicaiton logs, market data feeds, web clickstream data, social media
- real-time metrics and reporting
- real-time data analytics
- Complex stream processing

High level architecture



Terminologies:

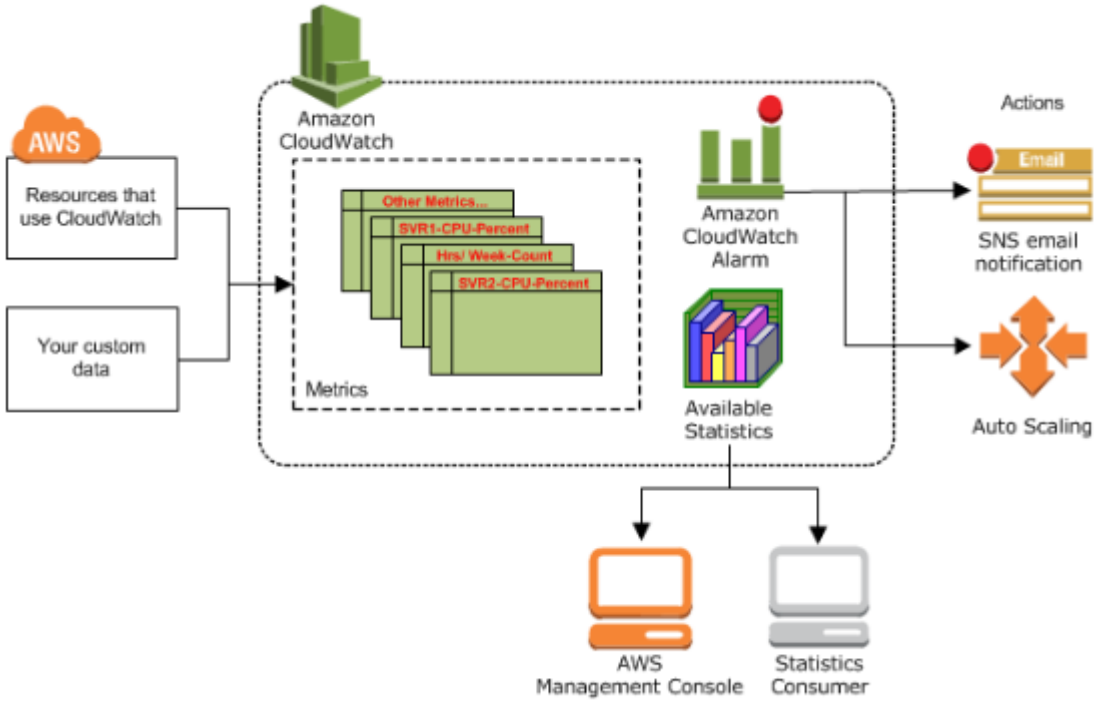
- **Stream** – ordered sequence of data

- **Data records** – unit of data stored in kinesis (contains sequence number, partition key, data blob) up to **1mb**
- **Retention period –** configurable in hourly increments from 24 to 168 hours (1 to 7 days), default **24 hrs**
- **Producers** – those who create the stream
- **Consumers** – stream consumers
- **Stream** application
- **Shard** – group of data records in a stream
- **partition** key – segregates data records
- sequence number

Cloudwatch

Cloudwatch – enables you to monitor AWS resources and applications you run in real time, Sends notification

How it works:



Concepts:

- Namespaces- container of cloudwatch
- Metrics – represents time ordered set of data points published to cloudwatch
  - **Metrics can't be deleted, automatically expire after 15 months**
  - Time stamps
  - Metrics retention
    - period 60 seconds – available for 15 days
    - period of 300 seconds – 63 days
    - period of 3600 seconds – 455 days

    - Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 2 weeks.
- Dimension – key value that identifies metric
    - Comibination of dimension are considered seperate metric
- Statistics
    - Minimum
    - Maximum
    - Sum
    - Average
    - Sample count – count of data points
    - pNN.NN – value of specified percentile
- Units – unit of measure – Bytes, seconds, counts, percent

Common Available metrics:

- RDS
    - CPU
    - Connections
    - Memory
    - Read/Write Throughput/latency
- EC2
    - CPU
    - Disk
    - Network
    - Autoscaling size, instances
- S3
    - Bucket size
    - number of objects
    - requests/put/delete/get
- SQS
    - Number of message
    - message size
    - delayed messages

Memory Utilization requires a custom Cloudwatch metric

CloudWatch **CANNOT** see the following:

- web server visible metrics such as number failed transaction requests -Too detailed for EC2 – Amazon don't even want to know whether you have or haven't even installed a web server.
- operating system visible metrics such as memory utilization –
  Too detailed for EC2 – Amazon don't want to interact with your operating system.

Limits:

Actions 5/alarm. This limit cannot be changed.

Alarms 10/month/customer for free. 5000 per region per account.

Additional Notes p1

Amazon Kinesis – real time processing of streaming data at massive scale ie website clickstream, application logs, social media feeds

Uses:

- Used to consume big data – can analyze from big amount of data, for example twitter can scan all the tweets for negative/positive comments
- Processing large amount of data

Exam tips

- Business Intelligence – think Redshift
- Consuming big data – think Kineses
- Big Data Processing – think Elastic Map Reduce

EC2 – EBS Backed vs Instance Store

- EBS back volumes are persistent, Instance store are not persistent (Ephemeral)
- EBS can be dettached/reattached
- Instance store cannot be detached
- EBS volume can be stopped, data will persist
- Instance store cannot be stopped, if an EC2 has EBS root volume is stopped with Instance stores as additional volumes, all data with instance stores will be lost

Exam Tips

- EBS Backed – Store Data Long Term
- Instance Store – should be used for long term

OpsWorks

- Orchestration Service that uses Chef
- Chef – converts infra to code, to maintain consistent state, consists of **recipes, cook books, chef**

Elastic Transcode

- Convert media to different formats, ie one video can be played to mp4, to iphone or ipad, laptop etc
- pay based on minutes per transcode, and resolution

SWF – Actors

- Workflow starter
- Deciders
- Activity Worker

- Domain – group of related workflows

EC2 – get public IP address

- curl http://169.254.169,254/latest/meta-data/
- META DATA for public ip, USER DATA is for user data

Consolidated Billing

- Paying Account – linked to different accounts (dev/production/backoffice)
- paying account will get all the monthly bill, does not access the resources of different accounts
- Soft limit is 20
- Benefits:
  - One bill per AWS account,
  - easy to track all the costs,
  - Volume pricing discount for all accounts (AWS gives you discount the more you use their services, for example for S3 since there are multiple users, you use more gb, and hence take advantage of the discount provided, instead of paying individual)

Tagging and Resource Group

- Tags- are key value pair for metadata, ie. EC2 can have tags
- Resource Group make it easy for you to group resources using your tags, ie group your resources based on region or by dev or whatever metadata you specify
- you can export as csv for resource group so you can generate reports
- **Tag Editor** – enables you to edit all your resource and add them tags

Amazon Lex is a service for building conversational interfaces using voice and text. Polly is a service that turns text into lifelike speech.

Exam Notes part 2

CloudTrail

- You can use AWS CloudTrail to get a history of AWS API calls and related events for your account. This includes calls made by using the AWS Management Console, AWS SDKs, command line tools, and higher-level AWS services.

Cloudfront

- Amazon CloudFront can handle data transfer rate 1,000 Mbps and 1000 requests per second.

S3

- S3 Standard – IA offers the high durability, throughput, and low latency of Amazon S3 Standard, with a low per GB storage price and per GB retrieval fee.
- Only difference of IA with standard is 99.99 availibility!

- IA has minimum of 128KB bytes, S3 standard has 0 bytes minimum
- S3 does support website redirects.
- Using IPv6 support for Amazon S3, applications can connect to Amazon S3 without needing any IPv6 to IPv4 translation software or systems.
- Using an encryption client library, such as the Amazon S3 Encryption Client, you retain control of the keys and complete the encryption and decryption of objects client-side using an encryption library of your choice. Some customers prefer full end-to-end control of the encryption and decryption of objects; that way, only encrypted objects are transmitted over the Internet to Amazon S3.
- CRR replicates every object-level upload that you directly make to your source bucket. The metadata and ACLs associated with the object are also part of the replication.

Glacier

- Because Amazon S3 maintains the mapping between your user-defined object name and Amazon Glacier's system-defined identifier, Amazon S3 objects that are stored using the Amazon Glacier option are only accessible through the Amazon S3 APIs or the Amazon S3 Management Console.

ELB

- ELB supports ipv6
- **Elastic Load Balancing** offers **two types of load balancers** that both feature high availability, automatic scaling, and robust security.
- These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.
- The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures.
- Two components:
  - the load balancers
  - controller service – verify the load balancers
- To ensure traffic is evenly distributed: **"Enable Cross-Zone Load Balancing"**
- Connection draining is the concept of ensuring traffic are not sent anymore to instances that are deregistering or unhealthy.

EC2

- Cluster group can only be in one AZ
- Amazon's SLA guarantees a Monthly Uptime Percentage of at least 99.95% for Amazon EC2 and Amazon EBS within a Region.
- EBS volumes can be attached to an ec2 instance in the same AZ
- The AMIs will need to be copied to the new Region prior to deployment.

RDS

- By default, the scan operation processes data sequentially. DynamoDB returns data to the application in 1 MB increments, and an application performs additional scan operations to retrieve the next 1 MB of data.
- The easiest way would be to take a snapshot of your DB Instance outside VPC and restore it to VPC by specifying the DB Subnet Group you want to use.

- To automatically failover from one geographic location to another you should use Multi-AZ for RDS.
- You should implement database partitioning and spread your data across multiple DB Instances.
- Databases generally do not require public access from the Internet, so a private subnet is the better choice from a security perspective. /28 is the smallest possible subnet in an AWS VPC.
- ***RDS replication : MULTI-AZ – Synchronous , Read-Replica – Asynch***
- At this time, you cannot have a multi-AZ copy of your read replica.
- Read Replicas are supported by Amazon RDS for MySQL and PostgreSQL.
- Infrequent IO:Amazon RDS Magnetic Storage would be the most suitable.
- At the present time, encrypting an existing DB Instance is not supported. To use Amazon RDS encryption for an existing database, create a new DB Instance with encryption enabled and migrate your data into it.

SMS (Server Migration Service)

- Improvement of VM Import/Export
- Simplify migration process, orchetrate multi-server migrations, test, support, minimize downtime
- 50 concurrent VM migrations per account
- 90 days service usage

Others

- Amazon DevPay and FPS – for paying
- It's always best practice to grant users access via IAM roles and groups even if they only need access once
- SWF has a gurantee that processes are only executed once against SQS
- Availability Zones offer you the ability to operate production applications and databases which are more highly available, fault tolerant and scalable than would be possible from a single data center.
- You can use **AWS Config** to continuously record configurations changes to Amazon RDS DB Instances, DB Subnet Groups, DB Snapshots, DB Security Groups, and Event Subscriptions and receive notification of changes through Amazon Simple Notification Service (SNS).
- SSD volumes must be between 1 GiB – 16 TiB.
- Economies of scale: The AWS Well-Architected framework has been developed to help cloud architects build the most secure, high-performing, resilient, and efficient infrastructure possible for their applications. This framework provides a consistent approach to application and solution architecture that will scale with your needs over time.
- AWS Config – enables you to keep track of all the config you have for your rersources
- In **cloud** computing, **elasticity** is defined as "the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible**"**
- **Paying account and linked account for Consolidated billing**

Trusted Advisor

- **Covers Performance, cost optimization, security and fault tolerance**

Exam Notes part 3

- **Glacier** – 10 gb free retrieval under free tier
- **Instance stores** – cannot be in stopped state; they are either terminated or running
- **EC2 Instance states lifecycle:**

    - pending
    - running
    - rebooting
    - stopping
    - stopped
    - shutting-down
    - terminated
- **Cloudfront request if files is not in cache:** holds the request until origin server serves it in the cache of the edge location
- **Cloudformation parameters:**
    - Outputs: for specifying the outputs
    - Parameters :(name etc)
    - Mappings
    - Resources: resources of cloudformation, type of resources
- **Multi AZ RDS** helps in maintenance tasks as it will initiate auto failover in standby
- **AWS VPC Platforms**
    - EC2 Classic
    - EC2-VPC
- **RDS Soap webservices calls uses HTTPS only**
- **Lambda** supports
    - Java
    - Node.js
    - Python
    - C#
- RDS Read Replicas limit: 5
- **AWS CodePipeline** is a continuous delivery service that enables you to model, visualize, and automate the steps required to release your software.
- **CloudTrail logs to S3 bucket** = 5minutes
- **Oracle BYOL (**bring your own license): Standard Edition 2 (SE2), SE1, SE, Enterprise Edition (EE)
    - Included license: SE1, SE2
- **http://status.aws.amazon.com = AWS Service Health Dashboard**
- **Resources that can't be tagged:**
    - Dedicated hosts
    - Elastic IP
    - Instance Store volumes
    - Key-Pair
    - NAT GW
    - Placement groups
    - VPC endpoint/flowlog

AWS Limits

Amazon Web Services Soft Limits – they are the limits by default but this can be increased by sending a request to Amazon

IAM

- 100 Groups
- 250 Roles
- 5000 users

RDS

- 40 DB instance
- 100 TB Total Storage
- 35 days maximum backup
- DynamoDB
  - 400kb item limit size
  - 1 read unit – 4kb
  - 1 write unit – 1kb
  - Batch Operations – 16mb
  - 3 geographic locations
  - Scan maximum – 1mb
    - local secondary index – size of all indexed items = 10gb or less
- Aurora DB
  - retains 6 copies (3 Az with 2 copies)
  -

S3

- 100 Buckets per account
- 5 TB size limit per bucket

EC2:

- 50 instances per region
- 500 Security group
- 100 Rules per security group
- 5000 key pairs

VPC:

- 5 Elastic IP
- 5 VPC per region
- 5 internet gateway
- 5 NAT Gateways
- 50 customer gateway per region
- 50 VPN connection
- 50 outbound/inbound per group
- 200 subnet
- 200 routes per table

CloudWatch

- Store metrics after deletion: 2 weeks.

Regions

- There are west-east and central regions

AWS Support Service Level

- Basic
- Developer – non critical – 12 hrs, general – 24 hrs
- Business – Business impaired issues : 4 hour response time, business down – 1 hour response
- Enterprise – Critical problem : 15 minutes response time

Underlying AWS you can Access:

- Elastic Map Reduce
- Elastic Beanstalk
- OpsWork
- EC2