Compute Service

EC-2 → It's units are called Instance.

Step-1

Select an AMI

Step-2

Configure networking and security

Step-3

Choose the instace type

Step-4

Choose the AZ.

Step-5

Attach storage (EBS)

Step-6

start EC2.

You could have maximum of 20 instance attached to it with 2 default high I/O instances.

A series has Medium / Large ~~Instance~~ size.
M series just has large size
T series has all 4 instances available.

Types of Instance.

- General purpose.
  - T2, M5, M4, and M3.
  (M series, T series and A series)

  Provides a balance of memory and network resource.

Something that used fairly consitently but not vigourslyC like 1000 user accesing it).

We can use hardware virtualization (as it needs less computing and less memory generally).

T2 class provides burst performance.
- credits accrue during idle times.
- Credits are used to give burst performance

M does not have burst performance
- are good for development and staging.

- Compute Optimized.
  - C5, C4 and C3.
  - Useful for CPU intensive application.
    e.g. - media coding
    - Intensive batch jobs
    - Many concurrent users
    - Intensive batch Jobs
    - Gaming server.

- Memory optimized.
  - X1e, X1, R4 and R3.
  - Usefull for high memory requirements
    e.g. processing large datasets.
    - In memory databases.
    - big data processing.

- Storage Optimized
  - H1, I3 and D2.
  - Usefull for high sequential read/write to local storage.
    - Relational database.
    - Data warehousing.
    - Image storage and processing

Note → You can use memory optimized instance with right EBS volume to get storage optimization.

- Advance computing
  P3, P2, G3 and F1

  → mostly need for GPU.
  -

~~Storage~~
Cost depends on 3 things

- Have the instance running.
- Have the appropriate storage.
- Have the appropriate network throughput.

Pricing category for EC2

- On-demand → Don't know your workload.
  - charge for usage time at a flat rate.
  - Billed in 60 second increment rounded up.
    E.g. Use 20 sec → get billed for 1 min
    " 90 sec → " " " 2 min
  - Use for short term workload that cannot be interrupted

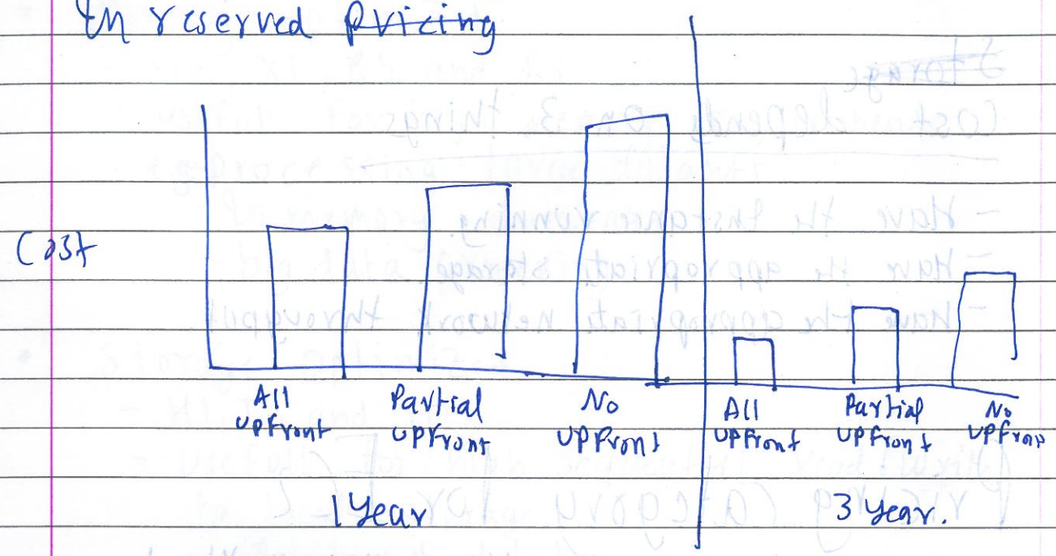
- Reserved.
  - Hours
  - Reserve usage minimum 1-year
  - Can be far less expensive than on demand.

- Spot category.
  - Bid on unused compute time.
  - upto to a 90% discount over on demand

- Scheduled Instance
  - Enables you to purchase capacity ~~requirement~~ reservations that recur on a daily, weekly or montly basis
  - You Reserve the capacity in advance so that you know it's available when you need it.

Types of reserved Instance

In reserved Pricing Instance



Default user name while connecting to an instance is ec2-user

Shared tenancy is default. tenancy. modes

(Please review tenancy).

Dedicated hosts

- physical machine.
- Used by one customer
- must be explicity configured.

- Pros
  - More accurate licensing management.
  - More detail reporting.
  - Compliance management.
  - Determine host placement during instance restart.

Cons
- costly

Dedicated Instance

- It also runs on physical machine.
  - only one instance running on that machine.
  - on restart it may be moved.
  - your DI can share with other DI of your own.
- Used by one customer.
- Must be explicity configured.

Pros
- Runs on hardware dedicated to the customer.
- Provides performance advantage of a dedicated hosts.

Standard RI - most discount (upto 75%)
Convertible RI - These provide a discount (upto 54%) and capability to change the attributes of the RI.

AMI

- Blueprint with server configuration details.
- Similar to localized imaging solution.
- All "Instance" created from an AMI

Source
- Free Amazon
- marketplace.
- community
- ~~Public : Anyone.~~

Who can launch AMI
- Public - Anyone
- Explicit - Specified.
- implicit - Owner.

Default.

- Hardware Virtual Machine (HVM)
  - AMI's fully virtualizes the hardware
  - Required hardware assited virtualization.

PV AMI
- Runs on hosts without specific support for virtualization.
- Dosent perform as well as HVM AMI's

Instance Root volume.
- Contains the boot sector.
- Boot sector's initiates the boot loader
- Boot loader launches the OS.

Instance Root volume.

1. Instance store - backed AMI
   - No support for the stop action
   - On failure, data in the instance is lost.
   - Root volume is stored is S3 bucket.

2. EBS backed AMI
   - Root volume stored in an EBS volume.
   - support for the stop action.
   - On failure, data in the EBS volume is not lost.

EC2 status checked

→ By default AWS EC2 service perform automated status check every one minute.

→ This is done on every EC2 instance to identify and H/W or S/W issue.

→ Status check is built into the AWS EC2 instance.

→ They cannot be configured, deleted or disabled.

→ EC2 services can send its metric Data to AWS cloudwatch every 5 minutes

→ Enabled detailed monitoring is chargeable and sends metric every 1 minute.

→ You are not charged for ec2 instance, if they are stopped.

When you stop an EBS Backed EC2 Instance

→ EBS volume attach to Instance
→ Any data cached in RAM or instance store is gone.
→ Instance retains its private IPV4 or IPV6.

but it will return public IPV4.

→ Instances Retain its Elastic IP address.

EC2 Termination

→ EBC root volume gets deleted (you can change the config for this as well)

→ But it will not delete external EBS volume

→ Metadata will not be encrypted

Imp

Data in instance storage is not lost when its rebooted.

Data is lost when
→ underlying drive fails.
→ Instance is stopped or terminated.

launching Instances

- Bootstraping
  - Providing code to be run on an instance at launch.
- VM Import
  - Importing existing virtual machine.

Instance Metadata

- security group
- Instance ID
- Instance type
- AMI base of the Instance.

Security group

You could have 5 sg per instance.

- An instance receives a default security group from the VPC.
  - default setting
  - other security group may be attached.
  - default security group may be attached.

ECS

It helps to run docker containers.

Docker container helps application run on any ~~pipe~~ platform or os.

↓

Just like Java runtime can help us run any java code.

ECS Features

- no virtual machine build required.
- Uses Amazon fargate to automatically build env.
- Can use EC2 instance for more control.

Container server

Elastic beanstalk.

IAM

- Manage acces to AWS
- suppots users, group and roles

Iam Concept

Resources → Things on which action can be taken
principals → Things that can take action.
Users
Groups
Roles.
Policies.

Principals $\xrightarrow[\text{using}]{\text{authorised}}$ to take "Action" or resource.

identity based policy or group based.

Target tracking policies. - Increase or decrease the current capacity of the group based on a target value for the specific metric. This is similar to the way that your thermostat maitains the temp of your home.

Step scaling → Increase or decrease the current capacity of the group based on a set of scaling adjustment. Known as step adjustment, that vary based on the size of the alarm breach.

Does not support/wait for a cool down timer

Support a Warm-up timer taken by newly launch instance to be ready and contribute to the watched metric.

Scaling policies

→ Define how much you want to scale based on defined conditions.

→ Autoscaling group uses alarms and policies to determine scaling.

→ For simple or step scaling, a scaling adjustment can't change the capacity of the group above the max group size or below the min group.

Predictive Scaling → It looks at historic patterns and forecast them into the future ~~for scaling~~ to schedule change in the no of EC2 instances. It uses ML model to forecast daily and weekly data.

Auto scaling policy

- ~~Static~~ manual
- Dynamic
  - Target tracking
  - Simple scaling policy
  - Step Scaling Policy.

cool down → It will wait for certain time ~~to~~ before creating ~~or removing~~ a new instance irrespective of anything

warm up period → would wait for the instance to warm up. and check the overall CPU utilization.

Working with elastic load balancer with auto scaling group.

→ We can attach one or more elastic load balancer to our autoscaling group.

→ both of them must be in same region.

→ Once you do this, any EC2 instance existing or added by the Autoscaling [illegible] be automatically registered.

→ Instance and load balancer must be in same VPC.

Imp

→ You do not need to register those instance manually on Autoscaling group. defined ELB.

Health checkup.

→ Autoscaling, classifies its EC2 instances health status as healthy or unhealthy.

→ By default, Asg uses EC2 status check only to determine the health status of an instance.

→ When you have one or more ELB defined with the autoscaling group. You can configure Autoscaling to use 'both' the EC2 health check and the ELB health check to determine the instance check.

→ Health check grace period is 300sec by default.

Don't do it.

→ If we set 'zero' in grace period the instace health is checked once it is in service.

→ untill the grace period timer expires any unhealthy status reported by EC2 status checks, or the ELB attached to the autoscaling group will not be acted upon.

After grace period expires, Auto scaling group would consider an instance unhealthy in any of the following cases.

- EC2 status check report to Auto scaling group an instance status other than running.

→ If ELB health check are configured to be used by Auto Scaling than if the ELB reports the instance 'out of service'.

→ unlike AZ rebalancing, termination of unhealthy instace happen first. than autoscaling attempt to launch new instance to replace the one terminated

→ elastic IP and EBS volume get detached from the terminated instance you need to manually attach them to new instance.

Stand by state

→ you can manually move an instance from an ASG and put it in standby state
→ Instances in standby state are still managed by autoscaling.
→ Instances in standby state are charged as normal, in service instances.
→ ~~Instances~~ They do not count towards available EC2 instance for workload.
→ Autoscaling does not perform health check on instance in standby state.

Components

1. launch configuration - like instance AMI, type, key pair, security group.
   which EC2 would we work on.

2. Auto scaling group

   group name, group size, VPC, subnet
   health checks Record.

3. Scaling policy

   metric type, Target value.

Imp.

AS always tries to rebalance the instance distribution across AZ's.

While rebalancing, ASG launches new EC2 instance, where there are less EC2 present, and then terminates the instances from the AZ, that made more.

We can attach a Running EC2 instance to an AutoScaling group by using AWS Console or CLI, if the below mentioned condition are met.

→ Instance must be in Running state. (not terminated or stop).

→ AMI used to launch the EC2 still exist.

→ Instance is not part of another EC2 group.

→ If the existing EC2 instance under the ~~autobalancing~~ Autoscaling group, plus the one to be needed.

- You can manually remove EC2 instance from an Autoscaling group using AWS Console or CLI.

→ You can then manage the detached instances idependantly or attach it to another autoscaling group.

→ When you detach an Instance, you have the option to decrement the autoscaling group desired capacity.

→ If you do not, the autoscaling group will launch another instance to replace the one detach.

→ When you delete an Autoscaling group its, Parameters like maximum, minimum

and desired capacity are all set to zero. Hence, it terminates all its EC2 instance.

→ If you want to keep the EC2 instances and manage them independently, you can manually detach them first, then delete the ASG.

DynamoDB

- NoSQL
- Provides special Feature
  - milisecond latency at any scale
    - very fast read an write.
  - stored is SSD.
  - spread across 3 distinct data center

- Read consisteny types.
  - Eventual consistent. reads.
  - strongly consistent reads.

- storage

- Columnar data store.
  - Sequential reads
  - Very fast reads.
- Data compression.
- MPP (Massively parallel processing
- SSL transit encryption
- AES-256 storage management.
- It operates in 1 AZ.
- Snapshot can be restored to new AZ.

Aurora replicas

- up to 15 Aurora replica.
  - Automatic failover.
- Up to 5 MysQL read replica.
  - no automatic failover

OLAP

- Data warehouse database
- optimized for online Analytic processing
- AWS managed.
- Pricing

Redshift.

- single node.
  - 160 GB.
- Multiple node.
  - leader node.
    - connection and queires.
  - compute node.
    - store data and execute queries and calculation.

Database security

- RDS support "at Rest" encryption.
- must be enabled at creation time.
- can be enabled on recovery

- admin access based on IAM.
- Data access based on database capab
  - CRUD.
  - DB admin

Aurora.

- Relational DB
- optimized for online ~~processing~~ Transaction processing. (OLTP)
  - very fast writes.
- My SQL-compatible database system.
- increased performance from My SQL.

Scalability

- Scaling the instance.
- Auto scaling in supported in RDS
- ~~Changes the type/class~~.
  - can be scripted with CLI commands.

- Read Replica.
- Read only Copy of the database.
- offload read-only traffic from the main database.

- Multiple instances can be in different region.

Clustering.

- Multiple servers.
- One database with replication
- Increase availability
- Automatic Failover
- Increased cost.

Standby Instances

- Multiple servers.
- One database with replication.
- Increase recoverability.
- no automatic failover

long polling

→ Is preferred to regular / short polling. It uses fewer and request cost by

- eliminating false empty responses by querying all the servers.

- Reduce the number of empty responses by allowing AMazon SQS to wait untill a message is available in the queue before sending a response, unless connection timeout.

- Receive message wait time is set to a non zero value

→ Billing is ~~done~~ same for both patting

Short - Polling

A request is returned Immediately even if the queue is empty.

- It does not wait for message to appear in the queue.
- It queries only a subset of the Available servers for message (based on weighted random distribution)

→ default by SQS
→ Receive message wait time is set.

→ More request are used which implies higher cost.

How Amazon SQS charge

1) API action → Every Amazon SQS action count as a request.

2) FIFO request → API action for sending, receiving, deleting and changing visibility of message from FIFO queue are charged at FIFO Rates.

(3) Contents of request → A single requests can have from 1 to 10 message, upto a maximum total payload of 256 KB

(4) size of payload → Each 64 KB chunk of a payload is billed at 1 requests.

(5) Interaction with Amazon S3.
(6) Interaction with AWS KMS.

Standard Queue	FIFO Queue.
→ High throughput.	→ limited throughput (300 TPs)
→ At least one delivery	→ exactly one processing
→ Duplicacy is possible	→ Strict ordering First in first out
→ Best effort ordering	→ FIFO Queue are limited to 300 trasactions per second but have all capabilities of standard queue.

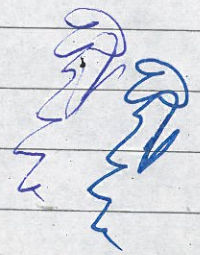
SQS

~~Simple Queue~~

Visibility timeout → When some server or AWS entity picks a resource in in the queue it becomes invisible for ~~30 sec~~. particular time

But suppose the AWS resource is not able to process it, than ~~the~~ ~~queue~~ resource would an API call to SQS for ~~more~~ extendinding the visibility timeout.

If the ~~the~~ timeout happen's than the queue item again reappears.



If a queue item get's failed to get processed. than it gets tranferred to dead letter queue.

↓

only recommeded for simple queue.

SQS is pull service.

SNS is push ~~s~~service.

You can use Autoscaling to scale the EC2 Fleet processing the SQS message as the queue size increases

- Cognito controls access to AWS resource
  - Define roles
  - map users to role.

Elastic Map reduce.

- Distributes processing across clustering.
  - implements a managed hadoop framework.
- Pulls data from S3 bucket.
- Uses EC2 instance.
- User defined the number you needed clusters
- Masters
  - coordinates job distribution across core and task node.
- Core.
  - Runs tasks assigned by the master node.
  - Stores data in the cluster.
- Tasks.
  - runs tasks that do not store data.

Cloud watch

- monitors the cloud and on-premises system.
- Dashboard.
- logs.
- Events → something happened.
  - ↓ executes Alarms
  - ↓ Can SNS notification
- Alarms

Why use it.
- monitor critical system
- receive notification. related to performance and security.
- push on-premises logs into the cloud.
- Take automatic action based on alarms

Organizations

- Collections of AWS accounts
- Centralized
  - one management interface
  - Billing
  - Account management.
- No additional charge.

Trusted Advisor

Elastic transcoding → take media file and convert it to required

- media transcoding
  - jobs - do transcoding
  - Pipeline - Queue to manage jobs.
  - Presets - setting to convert media
  - notification - SNS used to notify of job status.

- Transcoded files placed in original bucket.

Translate

- Translate content into different languages for geo localization.

- On-demand language

Elemental mediastore

- Video origination and storage service.
  - Containers
  - folders
  - endpoints
  - object.
  - policies.
- Live video streams.
  - elemental mediastorage

Don't do it for saved video → S3 is good for it.

Transcribe

- Speech to text.
  - Audio
  - Video
    - generates closed caption files.

Integrates with translate.

Rekognition

Analyis image and video analysis
- Find people
- Speech
- Object.

like google lens.

- Can run agaist S3 buckets.
  - Enhanced search based on analysis results

Improves search capabilities. in video file.

AWS workspaces

- virtual desktop in AWS
  - linux
  - windows 7
    - optional office bundal.
  - windows 10. (BYOL).
  - ~~virtual D~~

- Persistent D: drive.
  - virtual D: drive.

Appstream

- virtual application in AWS.
- Commonly used for custo-developed apps.
- Also used for commercial Apps.

Elasticache

- In memory caching for database.
  - Memcached
    - simplest model for implementation.
  - Redis
    - Hippa and PCI-DSS complaint.

HSM

Helps to encrypt and decrypt information.

When you create a ~~cluster~~ cluster the ~~encryp~~ encryption process is left to HSM for doing the job of encryption and decryption.

Directory service

Helps create AD.

Cloud search

Creates search engines.

Elastic search

AWS Glue

ETL service.

Quicksight

Business analytics.

Athena

Query S3 buckets.

Devops tools

AWS Solution

Transit gateway

Acts like a hub in a star network.

for VPC. (for peering, VPN, on premise sol^n etc)

Backup

Cost explorer

Historic data,

Operation Excellence.

~~Preparing properly~~

- Pre Pare
- operate
- evolve.

Pre Pare

- Understand workloads and expected behaviour.
- consideration
  - Operational priorities →
  - Design for operation. → designing operation based on priorities
  - operational readiness → Being able respond

Operate

- Monitor
  - Environment health
  - Discover business with technical insight.
- Respond
  - Security
  - Reliability
  - Performance
  - cost.

Evolve

- learn from experience
- share learning
- Scale
- Improve.

Resilient Design.

- Provides reliability
- Automation
  - Recovery
  - Scaling
  - Backups

Performance

- The key to performance is Autoscaling

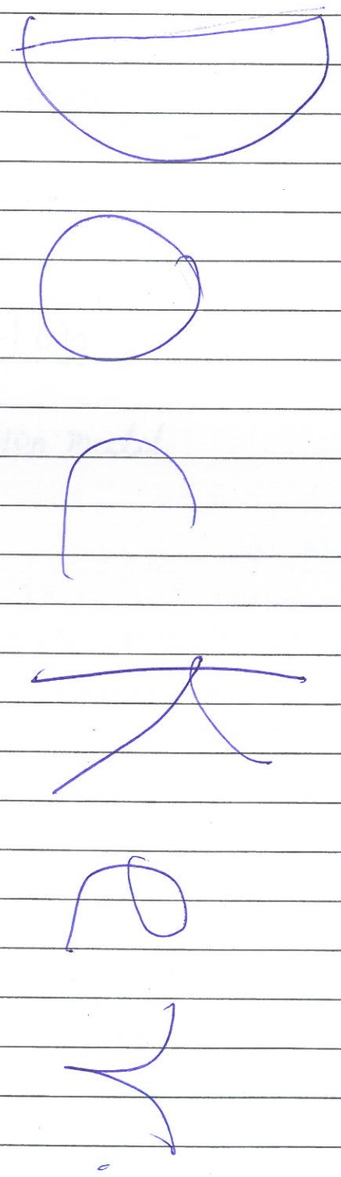
  EC2 instances can be scaled Automatically
  - logging of scale action should be in place.

- Database service can be scaled quickly.

Security pillar

Cost optimization

- Adopt consumption model

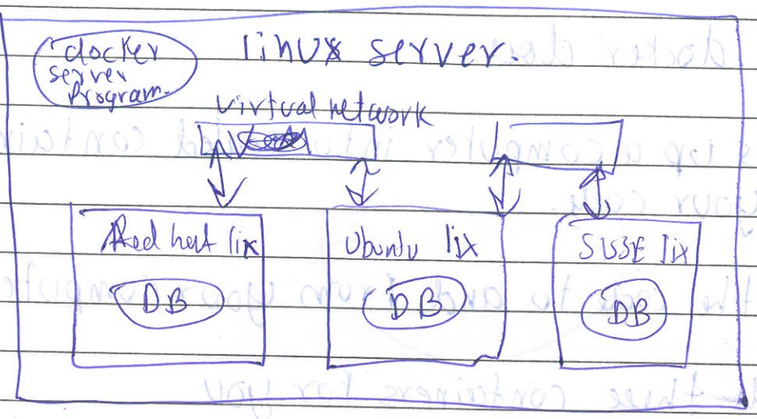


What docker does.

- Carves up a computer into sealed containers that run your code.
- Cuts the code to and from your computers.
- Builds these containers for you.
- Is a social platform for you to find and share contianer which are different from virtual machines.

What is container?

- A self contained sealed unit of software.
- Contains every thing required to run the code.
- Includes battery and OS.
- A container includes
  - code
  - configs
  - processes
  - networking
  - dependencies
  - OS.



About docker.

- A client program named docker.
- A server program that manages linux systems
- A program that builds containers from code.
- A services that distributes container (so you can share your container and other can find yours).

The Docker Flow

Image → It's just enough part enables you to do what you

Docker commands.

docker images → gets you all the images in docker.

docker ps -a → gives you the docker container created
'-a' flag → all the container which are stoped.

docker run -ti <Image name:tag> → runs the docker image.

-ti Open terminal for you. else docker runs and stops the container.

Memory limit

docker run --memory <max memory> image-name

CPU limit

docker run --cpu-shares relative to other container

Lessons From the field.

- Don't let your container fetch dependencies when they start.

- Don't leave important things in unnamed stopped container