EC2 Image Build: Automatically build, test and distribute AMIs

EC2 Instance store: High performance hw disk attached to our EC2 instance, lost if our instance is stopped / terminated.

EFS: Network File system, can be attached to 100s instances in a region.

EFS-IA: Cost optimized storage for infrequent class access

FSx for Windows: NFS for Windows servers

FSx for Lustre: HPC Linux file system.

ELB & ASG.

(Elastic Load Balancing & Auto Scaling Groups)

Scalability: Application / System can handle greater loads by adapting.

Vertical Horizontal (elasticity).

Vertical scalability: (Scale up/down)

Increasing the size of an instance

Eg: t2.nano → t2.large

Common for non-distributed system viz DB

HW limit

Horizontal Scalability: (Scale out/in)

Increasing the no. of instances

HS implies distributed systems

Very common for Web Applns / Modern Applns

Easy to HS.

High Availability:

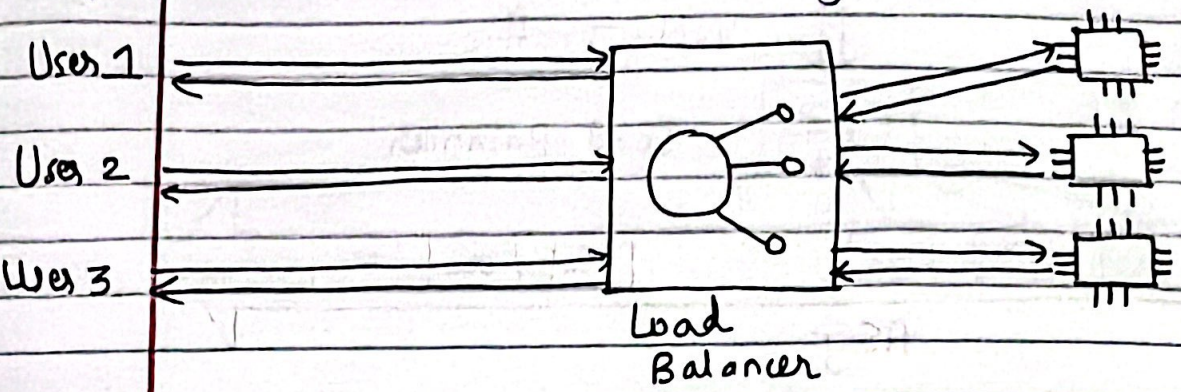
Horizontal scaling

HV means running Appln / system in at least 2 AVs.

Goal: Survive data center loss (disaster)

- Scalability: Ability to accomodate a larger load by making the hw stronger (scale up) or by adding nodes (scale out)
- Elasticity: Once scalable, auto-scaling cloud-friendly, pay-per-use, match demand, optimize costs.
- Agility: Reduce latency.

→ What is Load Balancing?



Servers that forward internet traffic to multiple servers downstream.

→ Why we Load Balancers?

1) Spread load across multiple downstream instances
2) Expose a single point of access (DNS) to your appl^n.
3) Seamlessly handle failures of downstream instances.
4) Regular health checks to your instances
5) Provide SSL termination (HTTPS) for your websites
6) High availability across zones.

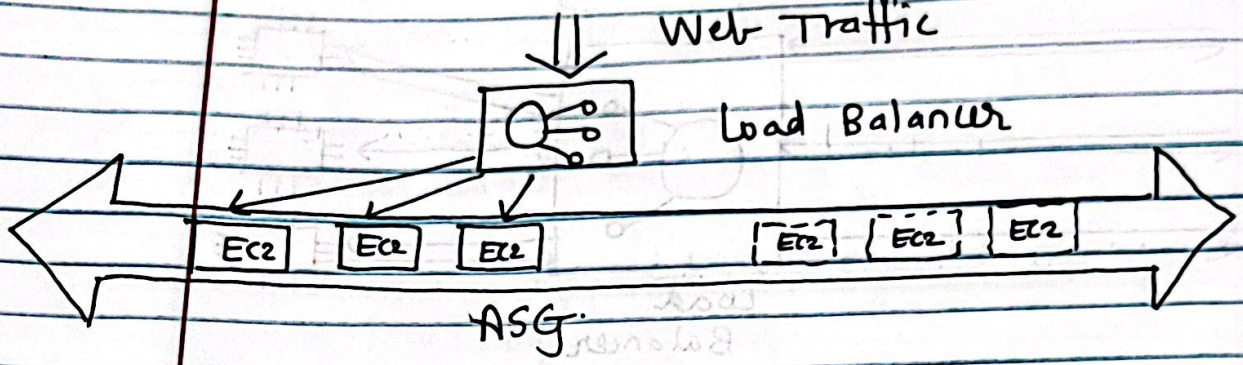
→ 3 LBs offered by AWS.

1) Application Load Balancer (HTTP / HTTPS only) Layer 7
2) Network Load Balancer (Ultra-high performance, allows for TCP) - Layer 4
3) Classic Load Balancer (slowly retiring) - Layer 4&7

What is an ASG?

1) Scale out
2) Scale In
3) Ensure: Min and Max machines running
4) Automatically register new instances to LB
5) Replace unhealthy instances.

ASG in AWS with LB



ASG: Scaling Strategies:

1) Manual scaling: Update the size of an ASG manually
2) Dynamic Scaling: Respond to changing demand
   - Simple / step scaling:
     When cloudwatch is triggered
   - Target Tracking scaling
     eg: I want the avg ASG CPU to stay around 40%.
   - Scheduled scaling:
     Anticipate a scaling based on known usage patterns.
     eg: Increase the min. capacity to 10 at 5pm on Fridays.
3) Predictive Scaling
   - Uses ML to predict future traffic ahead of time.
   - Automatically provisions the right no. of EC2 instances in advance.
   - Useful: load has predictable time based patterns.

ELB & ASG Summary *

High availability vs Scalability (Vertical and Horizontal) vs Elasticity vs Agility in the cloud.

Elastic Load Balancers (ELB)
Distribute traffic across backend EC2 instances, can be multi-AZ
Supports health checks
3 types: Application LB (HTTP-L7), Network LB (TCP-L4), classic LB (old).

Auto Scaling Groups (ASG)
Implement elasticity, across multiple AZ
Scale EC2 based on demand
Replace unhealthy
Integrated with ELB.

Simple Storage Service (S3).

Backup and storage
Disaster Recovery
Archive
Hybrid Cloud Storage
Application hosting
Media hosting
Data lakes & big data analytics
Software delivery
Static website.

Nasdaq stores 7 years of data into S3 Glacier.
Sysco runs analytics on its data and gain business insights.

Store objects (files) in buckets (directories).
Buckets must have a globally unique name (across all regions all accounts)
Buckets: defined at region level
S3 looks global but buckets are created in a regio

- Objects (files) have a key
- No concept of "directories" within buckets.
- Max object size is 5TB
- If uploading > 5GB, must use "multi-part upload".
- Metadata (list of text key / value pairs - system or user metadata)
- Tags (Unicode key / value pair - upto 10) - useful for security / lifecycle
- Version ID.

S3 Security:

1) User based:
IAM policies : which API calls should be allowed for a specific user from IAM console.
2) Resource based:
   1) Bucket tis Policies - Allows cross account
   2) Object Access Control List - Finer grain
   3) Bucket Access Control list - less common
3) Encryption:
Encrypt objects in Amazon S3 using encryption keys.

S3 bucket policies:

1) JSON based policies:
   - Resources: buckets and objects
   - Actions: set of API to Allow or Deny
   - Effect: Allow / deny
   - Principal: The account or user to apply the ~~appto~~ policy to
2) Use S3 bucket for policy to:
   - Grant public access to the bucket
   - Force objects to be encrypted at upload
   - Grant access to another account (Cross account)

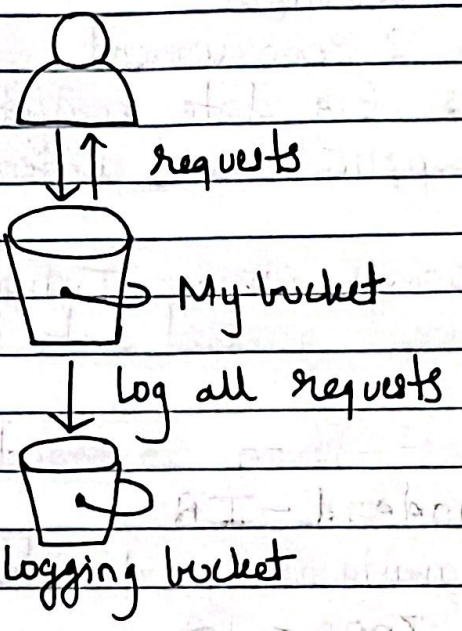
403 (Forbidden) error: Bucket policy is currently not allowing public reads.

Amazon S3 - Versioning:

Version your files in Amazon S3
Enabled at the bucket level
Helps protect against unintended deletes (ability to restore a version).
Easy to roll back to previous version.
Suspending versioning does not delete the previous versions.

S3 Access Logs:

For audit purpose, log all access to S3 buckets
Very helpful to come down to the root cause of an issue, or audit usage, view suspicious patterns etc.



S3 Replication (CRR & SRR)

Must enable versioning in source & destination
Cross Region Replication (CRR)
Same Region Replication (SRR)
Buckets can be in different accounts
Copying is asynchronous
CRR: Use cases: compliance, lower latency access, replication across accounts.

- SRR - Use cases: log aggregation, live replication, between production and test accounts.

→ S3 Storage Class:

1) Standard - General Purpose
2) Standard - Infrequent Access (IA)
3) One Zone - Infrequent Access
4) Glacier instant retrieval
5) Glacier Flexible retrieval
6) Glacier Deep Archive
7) Intelligent Tiering

→ S3 Standard - General Purpose
- 99.99% Availability
- Used for frequently accessed data
- Low latency
- High throughput
- Sustain 2 concurrent facility failures
- Use cases: Big data analytics, mobile & gaming applications, content distribution...

→ S3 Storage Class - Infrequent Access
- less frequent accessed data but rapid access when needed.
- Lower cost than S3 standard.
  - S3 Standard-IA
    99.9% availability, disaster recovery, backups
  - S3 One Zone - IA
    99.5% availability, Storing secondary backup copies on-premise data, or data you can recreate

→ S3 Glacier storage class:
- Low cost, archiving / backup
- Price for storage + object retrieval cost
- Glacier instant Retrieval: Millisecond retrieval, great for data accessed once a quarter, min. storage duration of 90 days

Glacier Flexible Retrieval: Min storage - 90 days, Expedited (1-5 mins), Standard (3-5 hrs), Bulk (5-12hrs) - free

Glacier Deep Archive: Long term storage, standard (12 hrs), Bulk (48 hrs) · Min storage: 180 days.

S3 Intelligent-Tiering

Small monthly monitoring
Auto-tiering fee
No retrieval charges
Frequent access tier: default tier
Infrequent access tier: objects not accessed for 30 days
Archive instant Access tier: " for 90 days
Archive Access tier: configurable from 90 to
Deep Archive Access tier: config from 180 - 700 days.