

Domain 2: High-performing architectures

Elastic vs scalable

Although they both mean adapting to dynamic environments they don't really mean the same thing.

Scalable means that you are allocating resource expansion on a more persistent level to meet workload growth. Take the example of pizza place, you notice that there is a steady growth in the popularity of pizza demand, so as the owner of a pizza chain you decided to open up another pizza restaurant anticipating the future growth. It is more permanent.

Elastic means that you are dynamically allocating resources to existing infrastructure in reaction to immediate demand fluctuations, there is a sudden pizza demand because of say Super Bowl or World Cup you would allocate more baker and cashier for your restaurant chain for that day, then the next world it would resume a normal operation.

As you can see scalable is more permanent, and elastic is on the fly for that specific peak demand. They work together to reduce cost while also ensuring that customers are meeting their demands.

Elastic and scalable compute solution

Just because your application is hosted on AWS it doesn't necessarily mean it is inherently scalable by default.

How scalable they are depend on what services you have chosen and how have you configured them.

Some services are scalable by default without you having to do anything, AWS Lambda is scalable and elastic by default, you don't need to make more AWS Lambda instances to scale up, or elastic.

EC2 on the other hand is not inherently scalable, but you can configure it to be scalable. You need to know what your application needs are, before you can choose the instance type. You should be able to pick the appropriate EC2 instance type for the corresponding workload.

Cloudwatch alarm

This is used to trigger scaling event, and you would have to pick a metric to monitor and use as a deciding factor on when to scale up your application.

The common metric that is used for monitoring is CPU utilization.

High-performing and scalable storage solution

Each storage solution have its use cases, and scaling ability. EBS vs S3 or EFS.

A EC2 attached to a EBS volume, when that EBS volume fills up it will not automatically increase the volume on its own.

A EC2 attached to a EFS filesystem, the filesystem will grow and shrink automatically without you have to do anything as you remove or add files.

You should also determine which storage solution is the best fit based on future storage needs. Each storage service also have a upper bound limit, the maximum size capacity.

Performance

You should also be able to determine which storage you should use according to its performance, the IO access speed is it fast or is it slow?

EBS volume are extreme low latency, but they can be configured to have a certain latency depending on the use cases.

Performance improvement

You should be familiar with some of the API that you can use to increase performance of data uploads or data retrieval.

For S3, data upload know the API or CLI command. Multipart uploads. Amazon S3 accelerator. Caching with Amazon CloudFront to improve retrieval speed.

High-performing networking solutions

You should be able to pick the best networking solutions for a workload given a set of requirements or circumstances.

Under a hybrid system, a company might still be hosting some data center for part of their application, and the other part utilizing AWS. The data center and the AWS architecture need a secure way of transferring data and messages across these two different system, securely and reliable.

You can connect data center and AWS via AWS managed VPN or AWS Direct Connect. Which one to pick will depend on the use cases, and their performance differs.

AWS Transit gateway, can be used with VPN or direct connect, to connect multiple VPC to a remote network.

Direct connect

High-speed, low-latency connection that let you connect on-premises (local) infrastructures to AWS services. The connection is made possible by dedicated lines and bypasses public internet to reduce network unpredictability and congestion.

However, the ease of use is not there.

AWS VPN/Site-to-Site VPN

Let you create an encrypted connection between Amazon VPC and your on-premise infrastructure over public internet.

It let you connect existing on-premise network to your AWS VPC as if they were running locally.

It is easy to set up and install.

AWS transit gateway

You would use transit gateway to connect multiple VPC together with on-premise network.



AWS VPN CloudHub

If you have multiple on-premise sites then you can use VPN Cloudhub to connect the VPC to all of those on-premise sites in a secure manner

AWS VPN CloudHub

Connectivity between VPCs

1. VPC peering: Point to point network connectivity between two VPCs only. But you can create a full mesh network that uses individual connections between all networks. This will require lots of work if you are planning to connect multiple VPCs with each other.
2. AWS Transit gateway: You can use this to solve the problem with VPC peering, creating a full mesh network all interconnected connections are much easier with transit gateway. You can connect existing VPCs together.
3. AWS PrivateLink: Doesn't use public internet for access services hosted on AWS. It connects services and not VPCs together.
4. AWS Managed VPN: Used for establishing connection between on-premise and VPCs

Amazon Route 53

Route 53 has a feature called geoproximity routing that routes the user request to the AWS service that is physically closest to the user.

Route 53 is a DNS server.

AWS Global Accelerator

Improves availability and performance for your public internet applications

Also consider caching using AWS CloudFront to improve performance.

Data transfer services

Multiple ways of transferring data to AWS that you should be familiar with.

1. AWS DataSync
2. AWS Snow Family
3. AWS Transfer Family
4. AWS Database Migration Service

Depends on the amount of data, type of data, destination, and source one will be better than the other.

High-performing database solutions

Need to know which services need to use non-relational, relational, and graph databases.

Amazon RDS can be high performance, but DynamoDB

Know the performance differences database services, RDS vs Aurora.

Aurora is MySQL and PostgreSQL compatible relational database. Have performance, availability, simplicity, and is cost effective. Compared to RDS, the performance is higher and more consistent.

RDS is managed SQL database, easy for provisioning, setup, patching, and backup.

RDS Storage types

General purpose SSD: This is for general cost-effective general use. Best for development and testing environments.

Provisioned IOPS SSD: Designed to meet intensive I/O intensive workloads. Best for production environments.

Magnetic: Used for backward compatibility. For workloads where data is accessed less frequently and small database.

Scaling strategies

Amazon DynamoDB scales storage under the hood automatically, but you also still have control over scaling throughput or you can auto scale as well to increase or decrease scaling on the throughput.

RDS also let you do auto scaling for storage, but to scale CPU usage you need to update the DB instances.

Database caching

Different database have caching features that you should be familiar

Revision #2

Created 5 January 2023 14:31:58 by Tamarine

Updated 6 February 2023 18:24:02 by Tamarine