Certified Solution Architect – Professional

Contents

[**Section 3: Designing Solutions for High Availability and Business Continuity** 1](#_Toc29998982)

[**Section 4: Design for operational complexity and cost control** 1](#_Toc29998983)

**12/31/2019**

# **Section 3: Designing Solutions for High Availability and Business Continuity**

* Business continuity and Disaster Recovery
* Resiliency and Fault Tolerant
* Redundancy and high availability
* Cost optimization
* Performance
* Security
* Monitoring
* Saleable and Elasticity
* Ease of deployment
* Migration and Hybrid architectures

Fault Tolerant: Ability of a system/application/infrastructure to be operational even if one or more system fails is call fault tolerant. A fault tolerant system should be the one which can recover/failover with minimal or no human interventions on a even of any failure.

AWS recommended to create a library of own AMI, with own standard & best practice. Ensure the AMI is up to date and stable for production usages.

Floating IP (elastic IP) = the IP that can be shift from the primary instance to secondary on an event of any failure.

Create regular EBS snapshots – EBS are region specific, to be used in another region we need to copy the snapshot to the desired region before creating the EBS volume from that snapshot. EBS are region specific and EBS volume are AZ specific.

Multiple site architecture: In this approach the goal is to have two or more independent copy of each application stack into two or more availability zone (site). Every application tier need to have redundant copy , in case of failure the traffic can be route to the alternative availability zone (site) – this can further be made more effective by introducing elastic load balance which can automatically route traffic between AZ, OR having a route53 health check created to route traffic between different region, alternatively one can also use elastic ip for routing traffic, so that when the instance fails the elastic ip can be remapped to another running instance in same/different AZ. Implement Autoscaling to address flatulating production load.

**01/01/2020**

Disaster Recover – Recovery plan from a system failure to resume business continuity without impacting business operation.

|  |  |
| --- | --- |
| **Benefit of having AWS as alternative site** | **Counter Argument** |
| Don’t need to negotiate contract with other provide in another region. | Avoiding vendor locking – have a multi-vendor/multi-cloud landscape. |
| One can use the same underline AWS technology across region | Need to have trained staff in more than one cloud technology – will be cost intensive to maintain multi talent pool. |
| One can use same tool/AWS build artifacts/API across region | One need to maintain different tool/build artifacts/APIs for different cloud. |

**Recovery Time Objective (RTO) –** Time taken after a disruption to restore a business process to its service level agreement as define by the operational level argument (OLA).

**Recovery Point Objective (RPO)** – Acceptable amount of data loss measure in time. For example, If the disruption occurs at 12 Noon and RPO is 15 minutes, the system should be successfully able to restore all its transactions which have occur till 11:45AM.

Note: RPO and RTO is defined by the financial impact to the business when systems are unavailable.

* Disaster Recovery Plan for Commute from AMI backup aspect:
* AMIs can help in restoring failed instance quickly.
* AMIs can be used within the context of the autoscaling to scale out DR site on an event of primary site failure.
* AWS recommended to maintain a library of preconfigure AMIs, including the application stack, so that on an event of a disaster these can be used to create new (EC2) instances to replace the old/failed once.
* Keep the AMIs available on DR sites.
* Disaster Recovery Plan for Storage from S3 Fault tolerance aspect [FOR OBJECT STORAGE]:
* Can be used as primary object storage.
* Highly redundant, with 99.99999999999 (11 9’s) redundancy rate. Objects stored within S3 buckets are automatically backed up in multiple facilities within the region.
* Additional security can be provided within S3 bucket for ensure high degree of data retention through MFA for delete operation, versioning, cross region replication – for auto backup in DR site, bucket level policies & object level policies.
* One can also use S3 Glacier storage class for archiving/storing data at very LOW COST, however RTO needs to be set to meet the recovery time as object archive in S3 Glacier takes longer time (usually 3-5 hours).
* Disaster Recovery Plan for Storage from EBS Fault tolerance aspect [FOR BLOCK STORAGE]:
* Point-in-Time snapshot can be created to backed up EBS content into S3 bucket.
* Snapshot created can be used to create a new EBS volume & then connect it the EC2 instance to replace failed EBS volume.
* EBS volume data are stored in different system within the Availability Zone (AZ), this provide protection from single system failure within AZ.
* Disaster Recovery Plan for Storage from Storage Gateway prospective
* Storage gateway provides an easy means to backup/store on-premises data on to AWS cloud. Customer can download VMware image from AWS console, and install a VM instance from the same on premises which will provide NFS/iSCSI interface for backup/storage.
* Storage Gateway – File Gateway (NFS interfaces) [OBJECT STORE]: files are asynchronously backed up into S3 bucket, there is a one-to-one mapping between on premises files and S3 objects. User can ALSO access the S3 objects(files) from the S3 bucket. S3 sub resources like lifecycle policy, versioning etc. can be implemented on the uploaded file(objects).
* Storage Gateway – Volume Gateway - Cached | Storage Mode (iSCSI interface) [BLOCK STORE]: Unlike File gateway, in case of the volume gateway, on premises files are stored into S3 bucket as BLOCK store. There are two possible operating modes – *Cached Mode* where frequently use files are cached on-premises while its asynchronously backed up on S3 bucket as volume store, in *Storage Mode* there is NO local file store everything is backed up on S3 as EBS snapshot. One can’t access the files stored in S3 directly – to access the files one need to create EBS volume from form the snapshot (*in case of cache mode, first snapshot needs to create from the volume store prior to creating EBS volume*).
* Storage Gateway – Tape Gateway: it provides an iSCSI VTL (Virtual Tape Library) interface to backed up on premises files into virtual tape data store within S3 bucket or can be achieve into S3 Glacier. It consists of virtual media changer, virtual tape drive, and virtual tapes.
* Disaster Recovery Plan for Storage from AWS Import/Export prospective:
* If large amount of data needs to migrate (backed up) quickly into AWS one can use AWS import/export feature – solution like snowball, snowball edge and snowmobile can be leverage for the same.
* Rule of thumb – “*if the data takes more than week to transfer over the available connection (VPN/Direct Connection/Open internet), then better use AWS import export instead*”.
* Disaster Recovery Plan for Storage from VM Import/Export prospective:
* Ease means to transfer VM image (hypervisor base virtual image) from/to AWS.
* This can be helpful in creating DR site at AWS or on premises.
* NOTE: VM Export is ONLY AVAILABLE to those instances which are initially brought in using VM import feature. AMI based EC2 instance cannot be exported using AWS Export feature.
* Disaster Recovery Plan for Storage from RDS Fault Tolerance aspect:
* RDS Multi-AZ: Primary-to-Secondary synchronous data replication within a region. On an event of any failure the primary DB instance DNS entry will be automatically swapped with secondary DB instance. Help in archiving zero downtime during system patching/upgrade.
* Read Replicas: For read intensive applications, a read-only DB instance can be provide with asynchronous data replication from the primary instance. On an event of any failure read replicas can be promoted
* Automated Backup: Automated backup in conjunctions with transaction logs to help recover failed DB instance with RPO up to 5 min.
* Manual Snapshot: this are necessary for backing up a DB instance OR restoring a DB instance in another region.
* Can be synced from on premises database to RDS and vise-versa – failover from the on-premises to AWS OR AWS to on-premises is possible.
* RDS instance size can be upgraded to a bigger size BUT can’t be downgraded to a smaller size.
* For DynamoDB one can copy the data into a S3 bucket and using data-pipeline it can be copied to another region DynamoDB. Also, one can use cross-region replication of the DynamoDB to replicate the data into another region DynamoDB table. For Sync operation (preparation phase), the dynamo DB can be created with lower read-write capacity, during recovery phase the capacity can be altered to meet the production demand.
* For redshift database: During the preparation phase one can create snapshot of the database in to S3 within same/different region, during the recovery phase redshift cluster can be (re)created from the copied snapshot.

**01/04/2020**

* Disaster Recovery Plan for Networking, on an event of a disaster the ability to quickly shift the network settings from the production site to the DR site is very much required.
* Route53 AWS managed service helps in quickly change and restore network connectivity (routing) on an event of a failure.
  + - Route53 is an AWS managed, highly available and highly scalable global service.
    - Route53 automatic Health-check configuration can route incoming traffic to healthy site.
* Using Elastic IP (floating IP): these are the static IPs that can be configured to the system, on an event of any disaster these IPs can be swap with the DR instances to quickly resume the production workload.
* Elastic Load balancing:
  + - It can load balance the traffic within a given region
    - Elastic IP addresses can be allocated to the ELBs, which make the transition simple on an event of any disaster. (DNS can configure to pass the load to the elastic IPs, which can load balance the traffic among healthy instances).
* Amazon VPC: On an event of a disaster Amazon VPC can help in extending the network topology to the cloud. This is mostly applicable in extending the on-premises application to the cloud during disaster.
* Amazon Direct Connect: This helps in connecting the AWS cloud network to the on-premises network by providing a low latency, high bandwidth dedicated physical connection.
* Disaster preparation phase: Before disaster strikes, the stage where the disaster/backup plans are made are called disaster preparation phase.
* Disaster recovery phase: After disaster strikes, the steps that needs to be taken to recover from the disaster is call disaster recovery phase.
* Different Strategies for Disaster Recovery are: These strategies can be mix and match to find a right solution that meets the RPO and RTO expectations.
* **Backup & Restore** 
  + - **Backup solution:** Snapshot | Object (file copy) | Volume copy.
    - **Storage solution:**  S3|Glacier | Tiring solution (according to the usages storage class will be selected).
    - **Data transfer solution:** Over internet using custom solution| AWS File Gateway | AWS Volume Gateway – Stored/Cached mode | AWS Tape Gateway | AWS import/Export | AWS VPN Connect |AWS Direct Connect
    - **Key Consideration:** 
      * Select appropriate means: tools/services for backup/storage/data transfer solutions.
      * Define retention policy – till when a backed-up objects/files/snapshots needs to be stored? How soon they need to be made available? Can they be archived?
      * Ensure adequate security measures are applied to secure backed up contents.
      * Plan regular game-day/dry run to ensure the data are correctly backed up and can be restore correctly when needed.
* **Pilot Light:** DR strategy where a minimal version of the cloud is always running on the DR site. During preparation phase: The core component will be identified and a minimal version of the solution will be running on the DR site and will be in Sync with the production. During restore phase the other components on the solution can be quickly provisioned and the production traffic will be routed to the DR site. RDS site may be upgraded to meet the production traffic.
  + - **Key Consideration** 
      * Backed up AMI can be use to provision the remaining component.
      * Running Instance size should be upgraded to meet the production demand.
      * Add resilience to the DR Site: Plan for fallout instance failure.

**01/06/2020**

* Warm Standby: Scale down version of the fully functional version of the application will be running on the DR site. On an event of any disaster, production traffic will be automatically route to DR site and the application will be automatically scaled up to meet production requirement.
  + - **Key Consideration**
      * Increase the size of the EC2 instance fleet (horizontal scaling).
      * Increate the sized of the EC2 instance (vertical scaling).
      * Implement Route53 health checks to automatically route production traffic to DR site, on an event of a disaster. OR manually change the DNS entries of the Route53 to route production traffic to DR site on an event of a disaster.
* Multisite: In case of multi-site (active-active) DR strategy, a fully functional redundant site will be running in parallel to the production site. ON an event of any failure production traffic will be severed by the DR site.
  + - **Key Consideration**
      * Configure weighted or equivalent routing mechanism to distribute production load between production site and DR site. (manually overwrite the Route53 DNS configuration to route all traffic to DR site)
      * Based on the RPO, synchronous or asynchronous replication need to be incorporate.
* Data Replication: For data replication following key things needs to be keep in mind.
* Distance between the sites – larger the distance more latency & jitter.
* Availability of the bandwidth – How two sites are connected? Type of connection VPN/Direct Connection etc.
* Data rate required by the application – it should be always lesser than that of the available bandwidth.
* Replication methodology: serial or parallel connection for replication.
* Synchronous (*150 ms latency*) or Asynchronous replication: Synchronous replication data are automatically update; Asynchronous replications are eventual consistent.
* Self-Healing: The following services are self-healing services.
  + SQS (Simple Query Service) can help in decoupling the applications which can help in automatic reprocessing of the failed instances without any human interventions
  + CloudWatch with EC2 instances autoscaling, which can automatically detect un healthy instance using CloudWatch metrics and terminate unhealthy instances.
  + EC2 autoscaling instances – which can be use to replace the unhealthy instance with healthy once.
  + AWS Glacier/S3 perform integrity check and ensure that all objects are constantly availability across different S3 systems.

# **Section 4: Design for operational complexity and cost control**

**01/09/2020**

* Possible AWS Infrastructure configuration from account prospective are
  + One Account for the company, and individual account per employee. [Typically situated for small company.]
  + One Account for the company, and IAM user per employee. [Seen mostly on mid-size companies.]
  + One Account for the company, and account per company department. [Most common.]
* For above possible infrastructure configuration, one can use one of the following bill consolidation process:
  + Manage separate bill for each account [Maintain separate payer account}.
  + Have consolidated bill for each of the linked account, into a master account (payer account). [single consolidated billing.]
* Advantages of having a consolidated billing are:
  + Single bill – need not to deal with each account/department individually.
  + Ease tracking – per account level usage
  + Ease of user maintenance & standardization across all account(department) of the organization.
  + In case of consolidated billing, AWS consider all linked account as one single account (master account), where one can take cost benefit as AWS services are priced in tire/usages with increasing usage unit price reduces.
* **AWS Cost Explorer**
  + Helps in viewing AWS cost as a graph
  + Filter graphs as per – Region, Availability Zone, AWS service, API Operation, custom cost allocation tag, Amazon EC2 instance type, purchase options, usages type, usage group and more.
  + Cost can be filtered per linked account, in case of consolidated billing.
  + Cost forecasting base on historical data.
* **AWS Budget** 
  + Two budgets free per account, up to 20,000 budgets can be configured. Fee will be applicable for budgets exceeding free budgets.
  + AWS Budget helps in planning & managing service usages, service cost, and instance reservations.
  + Following things can be viewed with ease in the AWS Budget
    - Account usage to date, including how much capacity of reserved instance has been utilized.
    - Current usages against the budgeted usage or free tier limit.
    - Predicted usage/bill at the end-of-the month.
    - SNS notification can be configure which can notify when the forecast exceeds the budget cost per account/service OR when the usages exceeded the budgeted cost.

Different Types of budget

* + - **Cost Budget:** Plan how much is plan to spend on a service
    - **Usage Budget:** how much is plan to spend on a single/multiple service together.
    - **RI (Reserved Instance) Utilization Budget:** Define utilization threshold and received alerts when the utilization is below the threshold limit, this protect from under-utilization of reserved instance (RI).
    - **RI (Reserved Instance) coverage Budget:** Define coverage threshold and received alert when the number of instance hours fall below the budgeted hours. This ensures that RI instance hours are in aligned with the budgeted hours.
  + Linked account can create budgets of their, using IAM policy one of the member account (lined account) can be granted access to read/create/edit/delete budget in other linked account – this is typically done to provide full control to financial department (linked account) to manage other account/master account budget – using AWS organization one can provide/revoke access to linked account from access AWS budget.
  + AWS cost explorer helps in visualizing the costing using different graphs - if cost explorer is NOT ENABLE, it gets automatically enabled when the budget is created for the very first time.

**Best Practice Recommendation from AWS**: Its recommended NOT to run any AWS services in master account, with ONLY exception to host a single S3 bucket where all invoices/bills are store for future references.

There are two ways to organize infrastructure from Billing prospective: a) using AWS organization & consolidate billing b) having a single account with multiple VPCs for each individual group.

|  |  |
| --- | --- |
| **Consolidated Billing** | **Single Account with multiple VPCs.** |
| Easier form AWS architecture prospective | Simple billing, one bill for everything |
| Volume discount – all discount related to all linked account can be better mange when consolidated | Per account discount – easy governance. |
| Need IAM policy account accounts to view manage account bill/budget | NO IAM policy needed. |
|  | Complex policies, to restrict/allow resource level access. |
| Tag is complex, as it needs to be done across different accounts. | Tagging is simple, as it needs to be managed within a single account. |
|  | Setting VPCs get completed from costing prospective. |

Note: In case of consolidate billing by default Reversed Instances are shared within all linked accounts. This can be disable if desired (Reserved instance sharing = turn on/turn off from the billing and preference page) by the master account (payee account).

**01/15/2020**

* + AWS CloudWatch can help in creating alert based on the usages and can send alert notifications for
    - When a set threshold of a service usages is reached as per the threshold defined in the billing alerts.
    - When the usage exceeds the amount that is defined in the billing alerts.
    - When signed up, it can also send alerts/notification when the pricing is changed.
  + AWS Tags are the key value pairs which can be used for organizing and categorizing AWS resources. There are two try of tags a) system (AWS) generated and b) custom tags a tag cannot be deleted or merged. Logical grouping of the resources is useful in – projects, cost centers, development environment, application, department.
  + Resource group: Group of AWS Resources that share one or more common tags ***within a same AWS region***. Resource group can be of two types – tag base and CloudFormation stack base. Resource group can have nested group from the same AWS region. They can be used for cost exploration or for other bulk AWS operations.
* **AWS Cost Optimization** 
  + Right Sizing, selecting the right EC2 instance family and size. This needs to be done based the performance benchmarking (based on workload utilization, CPU, RAM, networking , storage size, and I/O – which storage classes to be selected , should be selected busting throughput or provision throughput).
  + Continuous monitoring & tagging
  + Leverage AWS service for cost optimizing by finding right (new) services.
  + Cost explaining for cost optimizing
  + EC2 instance purchasing option – OnDemand EC2 instance are most expensive. Don’t use it forever, use it for a short spam of time (during initial phases) before shifting to the Reserved instance. (One of the reasons for choosing OnDemand instances is when the instance type is new/high demand and availability cannot be guaranteed within a specific AZ). **[RI applicable for the Availability Zone OR a particular Region – base on the choice instance will be reserved]. [One can also have scheduled reserved instances – where instance will be reserved for a specific scheduled.]**
  + **Dedicate Host:** Where a specific host is reserved for the client, NO other client instance will be running on that host. This helps in saving by bringing your existing per-socket licenses, per-core licenses or per-VM licenses.
  + **Dedicated Instance**: Pay per hour for the instance that are running on single tannate hardware.
  + **Capacity Reservation:** reserve capacity for your EC2 instance in a specific availability zone for any duration.
  + **Reserved Instance Attribute:** 
    - **Instance Type:** EC2 instance type for example m4. large – instance family/size.
    - **Scope:** The availability of the instance scope – within region or within zones (RI limits - **20 max zonal reserved instances and 20 MAX for each of the availability zone.**)
    - **Tenancy:** it’s a single tenancy (dedicated) or multiple tenancy.
    - **Platform:** Windows or Lunix

**Offering Class:** There are two offering classes for Reserved Instance, a) standard and b) convertible. Standard instance can ONLY be modified but CAN’NT be exchange for another Reserved Instance (RI) with different RI attributes. In case of convertible RI, they can be exchange for other RI instance with different RI attributes, like standard offering classes convertible offering classes can also be modified. Standard offing class instance can be sold in market place BUT convertible offering class instance can’t be sold in AWS market-place.

**Once a reserved instance is purchase, it cannot be returned – it can only be modified, exchange, sell.**

* + **AWS Spot Block Instances:** Spot instance which can be blocked for a definite duration (1-6 hours). The pricing is dependent upon the request duration and the availability of the capacity. Instance get automatically terminated at the end the duration or can be manually terminated. NOTE: using RequestSpotFleet API function one can launch a large number of EC2 sport instance at same time. To run RequestSpotFleet one needs to provide Target Capacity , Maximum Bid price , Launch Specification(AMI Id, VPC, subnets or availability zones, security groups, block device mappings, user data, and so forth) , IAM Fleet Role
  + **Elasticity**
    - Implement autoscaling – horizontal scaling of the instance when needed scale out and when NOT needed scale out.
    - Non-production workload – start when needed and shutdown (*deprovision)* when not needed.
    - Go Serverless where possible – pay ONLY for the used computation.
    - Use managed services instead of EC2 workloads.
* **AWS Organization:** AWS organization is a service for managing accounts at scale. **It’s a global service, can be reached from US EAST url.** 
  + **AWS OU are used for centrally managed accounts, automated creation of the AWS accounts, OUs, and Hierarchy AND can be used to enforce Service Control Policies.** (*AWS OU does not manage or control cross account permission; it’s been done by the IAM policies*). AWS OU can override permission that are otherwise granted by the IAM policies.
  + **AWS Organization Unit is a container of other Organizational Unit (OU) or AWS accounts.**
  + **OU Policy** – are the rule/policy to control AWS services. If the **Service Control Policy** is attached to the root of the OU then the policy is applicable to all the OU. If the **Service Control Policy** is applied ONLY to the AWS account then that is applied only to that particular account.
  + **Feature of AWS OU**
    - **Create Account – OU can be used to create member accounts**
    - **Add new accounts – OU can invite new member account to it**
    - **Can apply service control policies to the member accounts**
    - **Consolidated single billing for all the member AWS accounts**
    - **Manage Hierarchical grouping of the AWS account as per – budgetary needs, security needs, compliance needs. [ONE AWS OU can have up to 5 nested OUs]**
    - **Manage AWS Service / individual API action for the member accounts, once applied this will override the IAM policy access provided by the member account administrator.**
  + **Mode of operation for AWS Organization** 
    - **Consolidate billing** *(default feature).*
    - **All features** *(additional feature like service control policies are applied in addition to the consolidate billing) – Once a member account accepts to join a OU request sent by the Master Account by default all OU features will be enabled for the member account. This is primarily done to implement restriction applied from the master account using service control policies.*

**Service Control Policies (SCP)– can prevent member account from accessing specific AWS services or even can block uses from leaving from the OU. Service Control Policy defines the services that can be used within the member account, its left to the IAM to provide required access to access AWS services that are granted by the Service Control Policy for that particular account.**

**There are two possible ways to manage member account AWS services**

**- using SCP whitelisting – in this SCP defines the services that can be access within the member OUs/Accounts and rest other services are explicitly gets denies.**

**- using SCP blacklisting (default) - in this SCP defines the services that are denied within the member OUs/Accounts and rest other services are implicitly allowed.**