Certified Solution Architect – Professional

Contents

[**Section 3: Designing Solutions for High Availability and Business Continuity** 1](#_Toc37050408)

[**Section 4: Design for operational complexity and cost control** 5](#_Toc37050409)

[**Section 5: AWS Services and Strategies for Deployment and Operation/Monitoring Management** 9](#_Toc37050410)

[**Section 6: Designing Network on AWS for Complex Organization and Hybrid Architecture** 29](#_Toc37050411)

[**Section 7: Introduction to Digital Certificate and Public and Private Share Encryption** 37](#_Toc37050412)

[**Section 8: Elastic Load balancer** 38](#_Toc37050413)

**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 span 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.
  + The changes made in organization unit have eventual consistency.
  + **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. Existing consolidate billing member, can be send fresh invitation to join for full feature mode.*

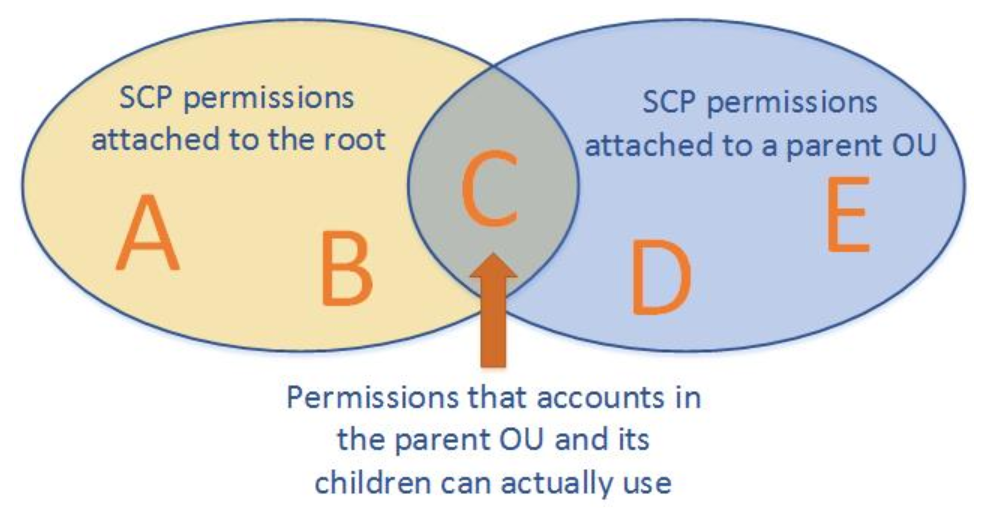
**Service Control Policies (SCP)– can prevent member account from accessing specific AWS services or even can block uses from leaving from the OU. Service Control Policy defines the services that can be used within the member account, its left to the IAM to provide required access to access AWS services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

**The master account of the organization is not affected by any SCPs that are attached either to it OR any root OR any OU master account is part of. [Master account is immunized by the SCP] AWS organization attaches an AWS managed SCP FullAWSAccess to all the root, OUs, and account which ensure building of the organization. In order to restrict access to an OU or an account replace FullAWSAccess policy with a custom polices with limited permissions.**

**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.**

****

* + **AWS Organization Integration with other services** 
    - **AWS CloudTrail:** CloudTrail from different AWS member account can be maintained in the master account, this will ensure that all trails are centrally logged.
    - **AWS CloudWatch:** CloudWatch can be configured to store all metrics from all member account into one single account. This will enable sharing of the event across organization.
    - **AWS Config:** Organization wide compliance status can be maintained, by sharing AWS Config across organization status.
    - **Artifacts:** All compliance report can be made available across organizations, anyone from the AWS organization can view the report when needed.
    - **AWS Firewall Manager:** Centrally all WAF configuration can be maintained from one single AWS account across AWS organization.
    - **AW Directory Services:** Integration of the AWS directory services, allows seamless sharing of the directory service across multiple accounts and VPCs within those account across AWS Regions.
    - **AWS License Manager:** The integration allows for cross-account discovery of computing resources throughout the AWS organizations to ensure licenses compliance of the resources in the organization.
    - **AWS RAM (resource access manager):** This helps in sharing AWS resources cross account. **(example sharing subnet, transit gateway etc).**
    - **AWS Catalogue:** Integrating AWS catalogue can help the organization to maintain a single portfolio of services hosted access the organization.
    - **AWS Single Sign** on across different account within an AWS organization.
  + **AWS Service Linked-Role (**AWSServiceRoleForOrganizations.**):** *This is an AWS managed role (i.e. one can’t edit or add policies to this role),* which helps the master account to enable trusted services in member accounts. When a new account is created in the organization OR a member account accepts an invitation to join the organization this role gets provision in the member’s account. ONLY AWS organization can assume this role. This role is NOT NEEDED if the organization is operated ONLY for consolidate billing the role can be deleted, later if all feature (organization all feature mode) needs to be enable the role needs to be restored back. Once organization all feature mode is enabled, then the AWS Service linked role can’t be deleted. AWS Service Linked role are excepted from AWS Organization Service Control Policies (AWS OU SCP). AWS OU SCP never effected AWS Service Linked Role.
* **AWS Organization best practices** 
  + Use AWS cloud trail to monitor and log API Calls in the mater data S3 bucket.
  + Don’t add AWS resources to AWS Master Account, except for those which are absolute necessary.
  + Follow Least Privilege Principle.
  + Provide AWS OU SCP at OU level then doing it at the account level. Policies applied at the OU level, automatically inherited by the accounts so one didn’t have to add policies to each individual account within the OU.
  + Avoid have SCP at the root level, this will bar all the accounts to avail the service.
  + Use whitelisting OR blacklisting, DON’T mix whitelisting and blacklisting – as it creates confusion during debugging.

# **Section 5: AWS Services and Strategies for Deployment and Operation/Monitoring Management**

**01/17/2020**

* **CloudFormation** 
  + Infrastructure as cloud – create /update/delete/version control “***stack***”, collection of AWS resources combines together in a single template.
  + Simplify Infrastructure management
  + Quick replicated of an Infrastructure in different region
  + CloudFormation will rollback stack creation, if any of the AWS failed to provision.
  + “***Change Set***” will hold ONLY the modified delta of the CloudFormation template, one can review the change and decide upon applying it OR rejecting it.
  + ***Deletion policy*** will help in retaining specific resource/information when a specific template is deleted. E.g. when deleted an EBS volume one might take a snapshot to retain its data, similarly for RDS instance before deleting an RDS instance one may like to take a snapshot.
* **Cloud Formation Template Anatomy**

|  |  |
| --- | --- |
| **CloudFormation Template Header** | **Description** |
| AWSTemplateFormatVersion (Optional) | The version of the CloudFormation Template usually defined by the version date. |
| Description (Optional) | Short description of the template, that define the stack. |
| Metadata (Optional) | Metadata of the template which can be added to describe the template. |
| Parameters (Optional) | Value that are passed to the template are called Parameters. One can refer the Parameter section from the resource and the output section.  Sample Parameter  "Parameters" : {  "InstanceType" : {  "Description" : "Amazon EC2 instance type",  "Type" : "String",  "Default" : "m4.large",  "AllowedValues" : [ "t1.micro", "t2.micro", "t2.small", "hs1.8xlarge", "cr1.8xlarge", "cc2.8xlarge", "cg1.4xlarge"]  } |
| Mappings (Optional) | Custom define Key/Value mapping which can be use to specific conditional parameters. It works like a lookup table  One can easily map the value matching to the key using intrinsic function fn:FindInMap in both resource and output sections. |
| Condition (optional) | Condition that control whether certain resources are created OR whether certain resource properties are assigned a value during stack creation or update.  For example: for test environment multi-AZ setup is not desired, this can be achieved through CloudFormation condition element. |
| Transform (optional) | This is to use for serverless applications (aka Lambda based application), to specify the version of the serverless application module (SAM). On specifying Transforms element, one can use AWS SAM syntaxs to declare resources in the template. |
| **Resource** (**required**) | This is required to specifies the stack resources and their properties for example – for specifying Amazon EC2 instance, s3 bucket. |
| Outputs (optional) | Describes the values that are returned whenever one views the stack properties. When one declares an output for an S3 bucket name, on calling aws cloudformation describe-stack CLI command it displays the name of the S3 bucket. |

* **Intrinsic functions:** This are the built-in functions which helps in managing the stack. Intrinsic functions can be used following places within AWS cloud formation template – Metadata, Resource and Output sections. It can also be used in update policy attributes. One can also use intrinsic function in conditionally create stack resources.

|  |  |
| --- | --- |
| **Intrinsic Functions** | **Description** |
| Fn::GetAtt | Fn::GetAtt returns the attribute of a resource in the template.Fn::GetAtt:[*logicalNameOfResource*, *attributeName*] |
| Fn::FindInMap | Fn::FindInMap returns the corresponding value mapped to the keys in a two level map that is defined under the AWS cloud formation mapping sections.  {"Fn::FindInMap" : [ "*MapName*", "*TopLevelKey*", "*SecondLevelKey*"]} |
| Fn::GetAZs | Fn::GetAZs returns an array of listOFAvailabilityZone for a given region  { "Fn::GetAZs" : "*region*" } |
| Fn::ImportValue | Fn::ImportValue returns the value of an output exported by another stack. This is typically used to create a cross-stack reference.  Things to keep in mind regarding the cross-stack referencing   * Cross-stack reference cannot be created across region (within region ONLY). * For each AWS account Export Name must be unique within a region. * One can’t delete a stack if it’s been referred by another stack. * One can’t modified/remove an output value that is reference by another stack. |
| Ref | The intrinsic function ref returns the value of the specific parameters or resources.  {"Ref": "*logicalName*"} |

* To further control resources within a stack one can add additional relationship and behavior of the resources specified in AWS CloudFormation template using following CloudFormation template attributes.
* **CreationPollicy**: TO delay the resource creation until other resource is created successfully (unit CloudFormation Template received specific number of success signal) or based on set timeout. Currently ONLY ***AWS::AutoScaling::AutoScalingGroup***, ***AWS::EC2:Instance*** and ***AWS::CloudFormation:WaitCondition*** supports CreationPolicy. One can use cnf-signal helper script to send signal back, notifying about the progress of the stack creation. (cnf-signal are helped script available***/opt/aws/bin/cfn-signal***)
* **DeletionPolicy**: If you want to retain (preserved) or create a snapshot (backed up) for resources before deleting the stack. This is also applicable to the stack update operation that leads to resource been deleted from the stack. (*Note: this is NOT APPLICABLE to the resources whose physical instance is replaced during stack update operation. For example, where AWS resource property is updated and cloudFormation template replace the resource with a new one.*)

**Deletion Policy Attributes**

**- Delete:** if delete attribute of the deletion policy is attached to any resource then resource will be deleted once the AWS CloudFormation stack is deleted. By default, if there is no deletion policy attached, then resource will be deleted from the stack. *However: Default deletion policy for RDS (AWS::RDS::DBCluster) is snapshot. For successfully deleting a S3 bucket, all the content within the S3 bucket needs to be deleted.*

**- Retain:** if retain attribute of deletion policy is attached to a resources AWS will NOT delete the resource(s), when AWS cloudFormation template stack is deleted.

**- Snapshot** snapshot attribute of the deletion policy can be attached to the AWS resources which supports snapshots like EBS volume, ElasticCache , RDS , Redshift etc. AWS cloudFormation creates a snapshot of the resource before deletion the resources.

* **DependOn Attribute**: Specifies the resources creation depends on creation of the other resources.
* **MetaData Attribute**: TO add more metadata to the resources.
* **UpdatePolicy Attribute**: How the resource updates need to be carried out.
* **UpdateReplacePolicy Attribute**:

**Nested Stack:** Moving the repeatedly using AWS resources within a main template into a separate stack within the main template and refer is where the common AWS resources are needed.

**Cross-stack-reference:** In order to improve the maintainability and ownership of the stack, standard infrastructure designs are grouped under single stack and been reference by other stack using cross-stack-references. For example, VPC networking related components can be placed under network-stack and the web-application stack can simply reference the resources from the output of the network stack which is maintained by the networking team.

**Stack-Policy:** The CloudFormation stack policy is a JSON document that defines what can be updated as part of a stack update operation. To set or update the policy, the IAM users or roles must first have the ability to call the cloudformation:SetStackPolicy action.

**AWS CloudFormation Best Practice**

**Planning and organizing**

1. Organize Your Stacks by Lifecycle and Ownership
2. Use Cross-Stack References to Export Shared Resources
3. Use IAM to Control Access
4. Reuse Templates to Replicate Stacks in Multiple Environments
5. Verify Quotas for All Resource Types
6. Use Nested Stacks to Reuse Common Template Patterns

**Creating templates**

1. Do Not Embed Credentials in Your Templates
2. Use AWS-Specific Parameter Types
3. Use Parameter Constraints
4. Use AWS::CloudFormation::Init to Deploy Software Applications on Amazon EC2 Instances – use CloudFront helper script cfn-init to install software and deploy application.

|  |
| --- |
| **Cloud Developer Kit (CDK):** is an open source framework to model and provision your cloud application resources through CloudFormation template. In CDK everything is a construct, developers can design, compose and share their custom resources that incorporate unique requirement. CDK can be model either in java or .net. |

* **Cloud Formation Best Practice – Planning & Organization** 
  + Organize Your Stacks By Lifecycle and Ownership
  + Use Cross-Stack References to Export Shared Resources
  + Use IAM to Control Access
  + Reuse Templates to Replicate Stacks in Multiple Environments
  + Verify Quotas for All Resource Types
  + Use Nested Stacks to Reuse Common Template Patterns
* **Cloud Formation Best Practice – Creating Template** 
  + Do Not Embed Credentials in Your Templates
  + Use AWS-Specific Parameter Types
  + Use Parameter Constraints
  + Use AWS::CloudFormation::Init to Deploy Software Applications on Amazon EC2 Instances
  + Use the Latest Helper Scripts
  + Validate Templates Before Using Them
* **Cloud Formation Best Practice – Managing Stacks** 
  + Manage All Stack Resources Through AWS CloudFormation
  + Create Change Sets Before Updating Your Stacks
  + Use Stack Policies (*Stack policy helps in preventing resources from unintentional update of the resources*)
  + Use AWS CloudTrail to Log AWS CloudFormation Calls
  + Use Code Reviews and Revision Controls to Manage Your Templates
  + Update Your Amazon EC2 Linux Instances Regularly
* **Cloud Formation Best Practice – Securing Stack** 
  + **Limiting Access to CloudFormation Stacks with IAM – Limit the IAM user access to the CloudFormation API. *Who can do what?***
  + **Use IAM conditions for CloudFormation to limit the creation and update of the AWS resources using CloudFormation using** 
    - **cloudformation:TemplateURL** – using this policy one can ensure that all create/update/delete stack call are made by the user using a specific CloudFormation Template.
    - **cloudformation:ResourceTypes** – Using resourceType parameter one can enforce that what type of AWS resources are getting created/updated/deleted
    - **cloudformation:StackPolicyURL**– The CloudFormation stack policy is a JSON document that defines what can be updated as part of a stack update operation. To set or update the policy, your IAM users or roles must first have the ability to call the cloudformation:SetStackPolicy action. Using StackPolicyURL parameter one can enforce the stack policy that needs to be used

**PENDING – OPSWORKs Note be added later.**

**03/21/2020**

* **ElasticBeanStack:** AWS Elastic Bean Stack is an eassy to use service for deploying scalable web application and services development with supported programming languages on familiar servers like Apache, Nginx, Passenger, Docker, tomcat and IIS.

|  |  |  |
| --- | --- | --- |
| **CloudFormation** | **AWS Ops Work** | **Elastic BeanStalk** |
| Where the target is the create a low-level infrastructure across different AWS region. | Where the target to deploy application in AWS infrastructure with more control on the deployment using chef /puppet base scripts. | Where the target is to deploy application into selected servers with minimum maintenance possible and visibility to application health monitoring. |
|  | While onboarding docker image into Opswork, one need to define granular details to onboard the docker image to create Elastic Container Service. | While onboarding docker image into Elastic BeanStalk with minimal configuration required. |
|  | There is NO direct integration of Ops Work with other AWS services like S3 RDS etc. However, one can use service role to access other services from Ops Work. | There is a direct integration of Elastic BeanStalk with other AWS services like S2, RDS etc . |

* ElasticBeanStalk provides
  + Supporting of major programming language like java/go/ruby/.net/php/node.js/python.
  + Supporting different platform for running major programming language.
  + Supporting web containers like Tomcat / Passenger/ Docker etc.
  + Can also support custom build platform for hosting not supporting languages/platforms.
  + It also provisions the resources needed to run the application like EC2, load balancer etc.
* How to deploy an application on elastic bean stack.
  + Develop the application
  + Upload the application bundle, like .jar .war file etc.
  + Provide information about the application
  + Elastic Bean Stack automatically launch the application and provision the resources needed for the hosting of the application.
  + Once the environment is provision and the application is successfully launched on the provision resource, elastic bean stack help in monitoring the application and also deploy a new version of the application if needed. Application monitoring metric and environment status can be monitored through AWS management console or through AWS command line interfaces.

Note: Elastic Bean Stalk does not provide data persistence, within EBS. Once the EBS is stopped the data persistence within its resources will be lost. It’s ADVISABLE NOT TO host RDS instance within EBS instance as the RDS data as well as the snapshot will be lost once the environment is stop. One can host the RDS instance outside of the EBS and refer the same using the CName within the application or store the files within S3 buckets.

* For Dev environment, one can consider using the elastic bean stock as an alternative for test database, but recommended for PROD environment.
* AWS EBS resources can be accessible directly, however it’s advisable not to alter their configuration directly as it may become unusable.
* One can rebuild an environment which has been terminated with 42 days, once terminated environment is rebuild it creates a new resource with the same name, ID and configuration – however data stored earlier will be lost.
* **Elastic BeanStalk Application**: Application server as a container for the environment that run the web-application and the version of the source code, saved configuration, logs and other artifacts that one creates for using Elastic Bean Stack. (In case of Elastic Bean Stack, Application is conceptually compare to a folder.)
* Deleting an application is Elastic BeanStack Application, results in deleting the application source code, logs , configuration and all the version of that application.
* **Application Version**: Application Version is a part of an application which refers to a labeled iteration of deployable code for the web application. Application version points to a unique labeled deployable package of the source code store under S3 location. One can deploy an application version to an running environment Or create a fresh application version of the code and deploy the same to the environment.
* **Environment**: With respect to Elastic BeanStack an environment is a version that deployed onto AWS resources. An environment can deploy ONLY ONE version at a time, if multiple different version of the code need to deployed side-by-side it needs multiple environment.
* **Environment Tier**: there are two types of environment tier – *webserver environment* & *worker environment*. Webserver environment are the once that servers HTTP request, while worker environment are the once that servers pulling task from the Amazon SNS (Simple Notification Queue) for processing.
* **Environment Configuration:** Collection of setting for the underline AWS resource for hosting EBS Application, if the configuration changes – new settings are applied on the underline AWS resources or deploys (change) the underline AWS resources with the new settings.
* **Configuration Template:** Template for creating environment configuration for (EBS) elastic beanstack.
* **Elastic BeanStack –** web container is a component of the webserver that runs Java Servlets / JSP. Apache webserver is HTTP server that manages static pages, caching, redirection while tomcats supports webapplication features.
* **Elastic BeanStalk Host Manager (HM) –** agent deployed to EC2 instance by the Elastic BeanStalk for deployment and monitoring of the application. Following are the task that are performed by an EBS HM (host manager).
  + Deploy the application into EBS – EC2 resources.
  + Aggregate Event and Metrics for retrieval via console /API/ Command Line Interface.
  + Generate Instance level events.
  + Monitor Application Level Log file for critical errors. (this feature is not available on OpsWorks).
  + Monitor Application Server.
  + Patching Instance component.
  + Rotating Application log and publishing it directly to configure S3 bucket.
* **Elastic BeanStalk Custom Platform or Custom AMI** – If supported default pre-build platform does not supports the customer requirement customer can opt for using Elastic BeanStack Custom Platform feature where one can build their desired platform from scratch and use it within EBS for hosting custom application.

In order to start with the building custom platform, one needs to starts with building AMI from a supported operation system, then create an Elastic Bean Stalk platform using Packer an open source tool for creating machine images for many platform including AMI which can be used within EC2 instances.

When to use custom platform?

- for hosting legacy application which are NOT supported by the Elastic BeanStalk platforms.

- to reduce the startup time of the application, as Elastic BeanStalk needs to run multiple scripts to build the desired environment for deploying the application, ALL these configuration (scripts) can be move to the AMI instead of running it during the installation process they by reducing the startup time of the application.

**03/22/2020**

* **Elastic BeanStalk – for docker container base applications:** A docker is a standardized unit of software development, containing everything needed to run including library binary files, system tools, code and runtime. The main benefit of using docker container is the ability to quickly and easily spin lightweight repeatable environments using minimal codding.
* One of the possible ways to onboard a custom application into elastic beanstalk is to create a docker image and use the docker image to build the Elastic Beanstalk application.
* Environment Variable define within the Elastic BeanStalk are passed directly to the container image.
* Elastic BeanStalk infrastructure automatically takes care of the capacity provisioning, load balancing , auto scaling and application health monitoring.
* If the docker container with the Elastic BeanStalk crashed / killed, Elastic BeanStalk restart the docker container automatically.
* Docker Platform configuration available within Elastic BeanStalk
  + **Single container Configuration**: It can be use to deploy Docker image describe in docker file or Dockerrun.aws.json (optional) definition and source code EC2 instances running in an Elastic Beanstalk environment.
  + **Multiple container Configuration**: Use Elastic Container Service (ECS) to coordinate deployment of multiple docker containers to Amazon ECS cluster in Elastic BeanStalk environment. Instance in the environment each run the same set of containers, which is define in dockerrun.aws.json file (in this case dockerrun.aws.json is mandatory).
  + **Pre-Configure docker platform configuration** **with supported language**: There are many pre-configure docker platform that can be use to run popular combination software stacks like java with glassfish etc. This option does not provide means to configure the software that application runs on. However, to customized the preconfigure docker platform to install additional software that the application needs, one can add a docker file to the application root folder.
* Elastic BeanStalk integration with other AWS Service
  + By default, Elastic BeanStalk creates s3 bucket in the same region where EBS is running, this bucket will be used by EBS for storing object required for proper operation of the application. The buckets are NOT encrypted, but the same time secure as it does not have public access to it.
  + Elastic BeanStalk can be integrated with Elastic File System (EFS), it can be mount to instance across availability zone, which can use it like local storage drive without changing the code.
  + Elastic BeanStalk can be integrated with RDS, EBS instance can refer the RDS using CName value so that in case of multi-az RDS instance once the primary DB instance fails and it get switch over the secondary instance one don’t have change the db end-point.

NOTE: Don’t use the EBS generated security group to all access to the RDS instance, as it will create a dependency on the RDS security group, which will prevent it from deleting when the EBS instance is terminated. (All other EBS resources will be deleted on termination, but NOT the security group). Solution to this problem is to crate a new security group (outside EBS) allowing DB communication with the security group and attach it to both EBS EC2 instance and to the RDS instance, this will help in removing the dependency).

* + Elastic BeanStalk can access dynamo DB (created within or outside the Elastic BeanStalk).
* Monitoring in ElasticBeanStalk – default ElasticBeanStalk is monitored by Cloud Watch. Elastic BeanStalk can also be customized to send custom metrics to CloudWatch. CloudWatch then can be used for monitoring OR creating alerts to implement decisions or automatic notification based on the define rules.
* Elastic BeanStalk application logs are automatically send to CloudWatch, which can be configured to also forward to an S3 location or archived at S3 location. Each of the EC2 instance of the Elastic BeanStalk has an CloudWatch Log agent installed which helps in publishing metrics datapoints [custom datapoints] to CloudWatch to each configure Log Groups. [Each Log Group applies its own filter pattern to determine what log streams events to send CloudWatch as data points – each logstreams can be configured to have their own retention, monitoring, access control settings].
* Additionally, enhance monitoring can be enable for Elastic BeanStalk EC2 instance. Each of the AMI for Elastic BeanStalk comes with health agent installed on it which helps in supporting monitoring of the enhance health.
* Elastic BeanStalk can export log data from different log group to be exported to S3 location, a same S3 location be used to store logs from multiple different log group (source). In order to separate the log data , different prefix can be attached to each of the log groups. [additionally lifecycle policies can be created on the S3 bucket to archive the data into S3 Glacier Storage class.]
* Logging for docker container within Elastic BeanStalk : One needs to configure application to write logs into a specific folder and then those logs can be configured to send to S3 bucket.
* CloudTrail can be enabled to Elastic BeanStalk API calls – every call made to Elastic BeanStalk can be made to record in cloudTrails logs which can then be send to S3 bucket.
* Best-Fit to use Elastic BeanStalk
  + Quickly Deploy application for prototyping or for testing.
  + To deploy application to AWS with no or minimal knowledge of AWS infrastructure.
  + Migrating custom web application to AWS infrastructure.
    - Create a docker file including all application dependencies.
    - Create a docker image from the docker file
    - Upload the docker image to public or private docker image repository.
    - Deploy the docker image using Elastic BeanStalk as single docker image or multiple docker image.
  + To be used in software development project by the software developers.
* When NOT to use Elastic BeanStalk
  + When team has expertise on build secure/scalable AWS infrastructure, and needs more control on the AWS infrastructure.
  + Team needs to recreate the AWS infrastructure quickly across different AWS regions.

**03/25/2020**

* **CloudWatch:** Provide required visibility to the AWS Infrastructure.
  + It provides system wide visibility into resource utilization, application performance and operational health.
  + Monitor AWS resources and Applications in real time.
  + CloudWatch can be use to collect and tract metrics – it can be AWS generated and custom metrics.
  + CloudWatch metrics can be viewed from Amazon CloudWatch console, AWS CLI, CloudWatch APIs, Or using AWS SDKs.
  + CloudWatch can collect data point securely from VPC resources using CloudWatch VPC interface endpoint.
  + CloudWatch can be configure to send Alarms (SNS Notification) OR to take actions (like EC2 recovery/start/stop/autoscaling) based on the configure condition on collected metric data.
* **CloudWatch Concepts** 
  + **Namespaces:** Namespace helps in isolating cloud watch metric data from mixing up. Each collected metric data has its own namespaces.One can also define custom namespaces for custom metrics.For AWS services AWS/<service-name> namespace is reserved.
  + **Metrics:** time-ordered set of data-points that are published by AWS services/resources to CloudWatch. For example, for EC2 instance CPU utilization data point sends for every 5 min intervals (basic monitoring) OR for EC2 instance CPU utilization data point sends for every 5 min intervals (detailed monitoring). [**CloudWatch are metrics are regional, however the CloudWatch dashboard can global.**]
  + **Timestamp:** each metrics data-point has to be had a specific time stamp attached to it. It can be 2 weeks in past and up to 2 hours in future. If no timestamp is attached to data-points , then the time it been received is set as its timestamp.

**Metrics cannot be deleted** – if now new data point are published on metrics more than 15 months AWS automatically deletes the metrics.

**Data-Point Retention**

**-** Data-Point less than 60 seconds (high resolution data points) are available for 3 hours.

**-** Data-Point equal to 60 seconds (1 minute) are available for 15 days.

**-** Data-Points equal to 300 seconds (5 minute) are available for 63 days.

**-** Data-Point equal to 3600 seconds (1 hour) are available for 455 days (15 months).

**-** Data points older than 15 months automatically rollover by the new data-points.

**-** Shorter Data points can be aggregated to longer data-points to keep it available for long duration.

> After 3 hours all less than 60 sec data-point are aggregated to 1 min data points.

> After 15 days all 1 min data points are aggregate to 5 min data points

> After 63 days all 5 min data points are aggregate to 1 hr data points and retain for 15 months.

* + **Dimensions:** << TBD >>
  + **CloudWatch Alarm:**

- It can be created on a single metrics or on math expression based on CloudWatch metrics.

- CloudWatch Alarm can be configure to perform action base on the configure threshold value over a time period (duration of time period where the alarm is evaluated).

- Possible state of a CloudWatch alarm are – OK / ALARM / INSUFFICIENT DATA

- For a CloudWatch ALRM to trigger three things are required

a) Data-Point: length of time to evaluate the metrics expression to create individual data point for alarm.

b) Evaluation Period: time period for which the data-points needs to be evaluated.

c) Datapoints to Alarm: Number of data points that need to exceeded the threshold value to trigger the alarm.

**[CloudWatch Alarm and CloudWatch Event are two separate things – CloudWatch alarms can trigger EC2 instance recovery/start/stop/autoscaling action/send SNS notifications WHERE CloudWatch Events can trigger many other services then what alarm can].**

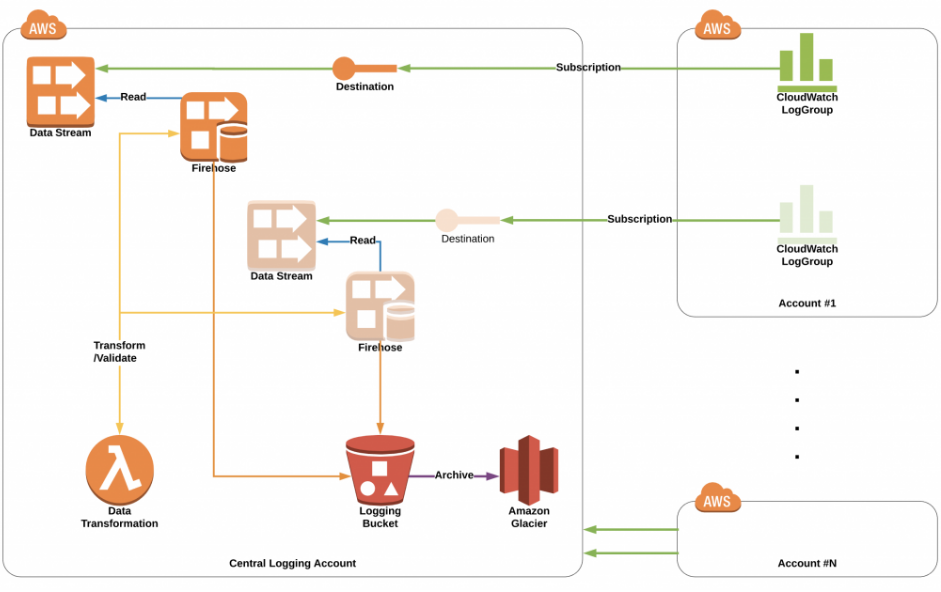
* **CloudWatch Logs: CloudWatch logs can be used for monitor, store, and access log files from AWS and non-AWS sources. CloudWatch logs can be monitored, store, access logs from** 
  + Amazon EC2 instances
  + Containers running on ECS (these required installing of the cloud watch agents).
  + AWS Cloud Trail
  + Route53 DNS Queries logs
  + RDS Aurora, MYSQL and MariaDB
  + Amazon Neptune
  + VPC Flow Logs
  + Elastic BeanStalk for EC2 instance in the EB environment.
  + API Gateway execution logging
  + Lambda function logs and other sources
  + Logs from on-premises instances (The instance needs to install cloud watch agents to sends the logs to cloud watch)

**Note: Logging in the central log repository, help in query filtering and tracing the logs end-to-end.**

* **Log Event** – is the record of some activity recorded by the application or resources been monitored. It consists of two properties timestamp and raw log message.
* **Log Stream** – sequence of log events that share a same source. Empty log steams older than 2 months old, are automatically deleted. One can delete a log stream manually. Log streams need to belong to a Log Group.
* **Log Group** – Group of log streams that shares same retention, monitoring and access control settings.
* **Metric Filter** – metric filter can be used to extract information from the log event and transform them to a data point in a cloudWatch metrics. Metric filters are applied at the log group level and gets applies to all logs stream within that log group. CloudWatch logs send log metric to cloudWatch every minute.
* **Encryption of log data in cloudWatch logs** 
  + **At Transit:** CloudWatch service encrypts the logs during transit using end-to-end encryption of data during transit.
  + **At Rest: CloudWatch** service can be configure to encrypt the log data at rest using server-side encryption, same can also be achieve using server-side-encryption using Amazon KMS – CMK (customer manage key) [*this can be done ONLY from CLI not from the AWS console*]. Encryption can be enabled at the Log Group level, while creating the group or later.
* **CloudWatch Log Insight**: CloudWatch Log Insight enables the interactive search and analysis of the log data in Amazon CloudWatch logs. It consists of few simple but powerful commands to search & perform analysis on the log data stored in the CloudWatch. CloudWatch Log Insight automatically discovers fields in AWS services such as EC2, Route53, AWS Lambda, AWS Cloud Trails, Amazon VPC OR any application/ custom logs that ingest into CloudWatch in JSON format. Up to 20 logs groups can be search a single request. **[CloudWorks insight works for logs that are logged after Nov 5h 2018 – logs sent before this data can’t be search.]**
* Real time monitor of EC2 instances: CloudWatch logs can monitor application logs from the Amazon EC2 instance in real time.
* CloudWatch Logs and CloudTrail:
* Exporting CloudWatch Logs to S3: Depending upon the size of the data , exporting of CloudWatch Logs to S3 bucket can take significant time (up to 12 hour), incase the logs needs to be process in real time then instead of exporting the data into S3 use subscription
* Streaming log data to elastic search
* **Unified CloudWatch Agent:** The newer version of the CloudWatch agent which can be installed on both Lunix and Windows base systems. It can collect metrics and logs from (EC2 or on-premises) systems. Its capable to send both rotating and non-rotating logs to cloudWatch which can be then viewedS on CloudWatch. [*New CloudWatch agent (unified cloudWatch agent) has better performance as compare to old one, and helps to capture advance metrics in addition to the logs for in-guest visibility.*]
* **CloudWatch - Real time processing of log data with subscription:** Use subscription to get access of real time feed of log event from CloudWatch logs to be deliver it to other services for custom processing, analysis, loading of logs through services like - Amazon Kinesis, Amazon Kinesis Data Firehose streams, AWS Lambda etc. [Can have subscription filter, to enable selective subscription of the logs.]
* **CloudWatch logs to be shared across AWS accounts for processing** [*cross account log data sharing*]. (Linked account (member account) can send the log to a designated member account for processing of the logs).

**NOTE: Currently ONLY kinesis data streams is allowed to collect the log from different accounts.** In order to create cross account log data sharing the following steps needs to be performed.

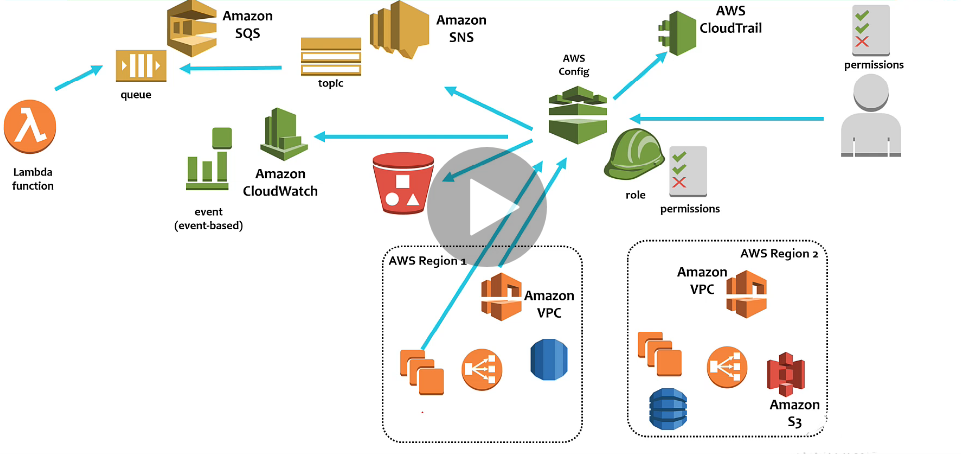
* Create a log group to collect logs streams from different destination – log group and the destination should be in the same region.
* Create a CloudWatch log subscription filter in case one needs to send ONLY the matching parameters to the target Kinesis data stream for processing.
* CloudWatch log destination name, logical name that point to the target (Kinesis data stream).
* Target ARN: ARN of the destination, situated on another AWS account.
* Role ARN: IAM role that grant required permission to enqueue data to the destination Kinesis data stream.
* Access Policy
* Use case: For centrally processing of log across multiple account.



Refer Link : <https://aws.amazon.com/blogs/architecture/central-logging-in-multi-account-environments/>

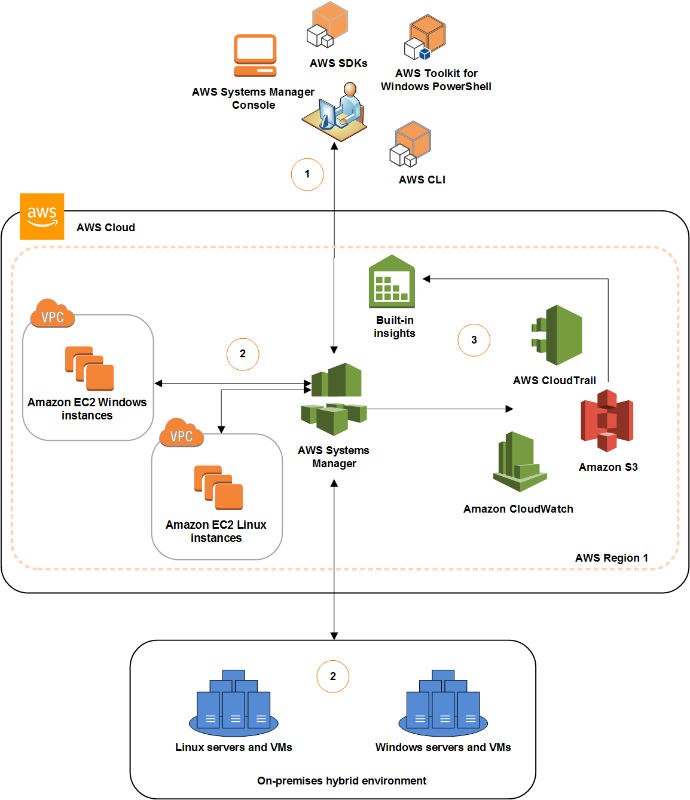
* CloudWatch Event: CloudWatch event delivers near real time streams of system events that describes changes in AWS resources. It can be used for:
* Sending messages to response to the environment.
* Activate functions
* Making changes
* Capturing state information
* CloudWatch **Event**:
  + Indicate change in the AWS environment
  + Custom application that can generate CloudWatch events
  + Schedule CloudWatch event to perform specific actions
* CloudWatch **Rules**:
  + Rule match the incoming CloudWatch event, perform action on them and forward it to the target. Rule can also be use to customized the events, before forwarding it to the target – it can override the JSON content of event with constants OR filter-out certain parts of the event.
* CloudWatch **Target**:
  + CloudWatch Target are the components that receives CloudWatch events in json format and process then. Many AWS services can be configured as AWS CloudWatch targets.

**Note:** *CloudWatch event can be shared across accounts, however the both the sender account and the receiving account needs to be in the same region. Sender account will be charge for sending CloudWatch event to the receiving accounts.*

* **CloudWatch Synthetics** – A fully managed services that enables canaries creation to monitor endpoints and APIs in target environment from outside in. canaries are configurable scripts that follows the same route and perform the same actions as a customer – which enables outside-in view of the customers experience and also the service availability from customer view point. It can also check the availability and the latency of the endpoint and can store load time data and screenshots of the UI. It can be schedule to run once or on a regular schedule intervals. CloudWatch alarms can be configure to notify if *canaries* fails.
* **CloudWatch ServiceLenses** – Enhance the observability of a service/application by integrating, traces, metrics, alarms and logs in one single place. AWS ServiceLense integrates CloudWatch with X-Rays, to provide end-to-end view of the application to help more effectively to pin-point the bottle neck of the service/application. Service Map displays the service endpoints and resources as nodes and highlight the traffics latency and error for each node and its connections. One can be drill down on the nodes to see the detailed insights of the correlated metrics, logs, and traces associated with that part of the service.
* AWS Config: How the resources are connected to other resources & it also helps in know the past configuration (configuration and change history). Its possible to create a private link for AWS config to look **
* Benefit of AWS Config
  + Evaluate AWS resources for desired settings
  + Get a snapshot of the current configuration of the supported resources that are associated with the AWS account.
  + Retrieve configuration of one or more resources that exist in the account.
  + Retrieve historical configuration of one or more resources that exist in the account.
  + Received configuration when a resource is created/modified/deleted.
  + View relationship between resources.
  + AWS Config can help in achieving administrator, auditing, and compliance
    - Achieving governance over the configuration status of your involved resources
    - Get notified when a change/delete/modification of a resources happens.
    - Evaluate auditing compliance of your AWS resources.
    - Perform auditing using the historical configuration data from AWS config.
  + Managing and Troubleshooting Configuration Changes
    - Using AWS config, related resource can be identified – minimizing the risk of cascading failure due to change in one resource.
    - Using historical configuration of the resources provided by AWS Config to troubleshoot issues and access the last know working configuration and the delta between the current configuration and the working configuration.
  + Security Analysis
    - To expose the potential security exposer historical configuration can be used
    - Trace the history of IAM permission granted to user or group.
    - Evaluation the configuration security groups when troubleshooting connectivity or packet dropping issues.
* AWS Config – Concepts
  + **Configuration Items** – a point-in-time view of the various attribute of supported AWS resources that exist in the account. It includes – metadata, attributes relationship, current configuration, and related events. AWS Config create a configuration item whenever it detects a change to a resources type that it is resource.
  + **Configuration History** – is a collection of configuration items for a given resources over given period of time. It helps in finding answers when the resource was created and when it was modified over the due course of time. It automatically delivers to S3 bucket, which can be retain for minimum of 30 days and maximum for 7 years
  + **Configuration Recorder** – When create its stores the configuration of the supported resources in an account as configuration items. By default, a configuration recorder records all the configurations of the supported resources of its region when AWS Config is running. (this can be change).
  + **Configuration Snapshot** – A collection of the configuration items for the supported resources that exist in an account. It’s useful for validating configuration. Configuration Snapshot can be delivered to an S3 bucket that you specify.
  + **Configuration Stream** – A configuration stream is an automatically updated list of all configuration items of the resource that AWS Config is recording. Every time a resource is created/deleted/modified AWS config create a configuration item and adds to the configuration stream. It works using AWS SNS topic. The configuration stream is helpful to monitor configuration changes in real-time as they occur. SNS topic can be use to notify the change in near-real time.
  + **Resource Relationship** – AWS config discovers AWS resources in your account and then create a map of relationship between AWS resources.
* **AWS Config – Delivering Configuration Items**: AWS Config can deliver the configuration items to following channels:
  + **TO Amazon S3 bucket**: AWS Config can track changes in the configuration of the supported AWS resources and sends updated configuration details to a specified Amazon S3 bucket. For each resource type that AWS config records in sends a configuration history file in every 6 hours. Each file consists of configuration details that has change is last 6 hours – if no configuration change AWS config does not manage the S3 bucket lifecycle. One need to create their own lifecycle policy as per their convenience.]
  + **TO Amazon SNS topic**: AWS config can be configured to deliver configuration item into SNS topic. It’s advisable to create an SQS to the SNS topic and process the delivered configuration items programmatically. Under following conditions AWS Config will send notification to SNS topic
    - On change of the configuration item for a AWS resources
    - On successful delivery of a configuration history of the account.
    - On completion of the configuration snapshot of the recoded resources of the account.
    - On execution of the compliant state of the AWS resources of the account.
    - On starting of the evaluation rule for resources of the AWS resources of the account.
    - On failure of the AWS config to send notification.
    - On identification of an non-compliant flag for any AWS resources of the account.
* AWS Config Rule – Evaluation Resources with AWS Config Rules
  + AWS Config rule represent a desired configuration, if the resource configuration changes from the desired configuration then AWS Config flags the resources and mark the rule as non-compliant. The AWS Config rules can be configured to run at pre-define time or when the configuration changes. [There are predefine managed rules for AWS Config and also custom rules can be created.] AWS Config rule are associate with lambda functions which contains the completing it evaluation.
* Permission for AWS Config: AWS Config needs IAM roles to access/evaluate configuration of different AWS resources.
* AWS Config is allowed to aggregated configuration and compliance data from multiple account & different region or organization into a single account – this helps in maintaining a central IT administrator to monitor compliance for multiple account withing an enterprise.
  + **Source Account** – is the AWS account for which configuration and compliance data needs to be collected – it can be a single account or multiple organization within a single account. AND the region where its hosted is known as source region.
  + **The Aggregator** – is a new resource type within AWS Config that collects AWS Config collection and compliance data from the multiple *Source Account* and region AND region where aggregator is created is call Aggregator region. Permission the Source Account grants to Aggregator account for collecting configuration and compliance data is called as “Authorization”, this is NOT REQURIED if source account and aggregator account are part of the same organization.
* AWS Config allows querying of the current configuration state of AWS resources based on configuration properties. It supports Ad-hoc, properties-based queries against current AWS resource state metadata across all supported AWS Config resources. AWS Config supports strong query feature provides a single query endpoint and power query language for obtaining current resource state metadata – this query can be used for:
  + Cost Optimization: For example, identifying list of EBS volume that are NOT attached to any EC2 instance.
  + Security and Operational Intelligence: For example, retrieving a list of all AWS resources within an account / enterprise that have a specific configuration property enabled/ disabled.
  + Inventory Management: for example, retrieving a list of all EC2 instance of a specific size.
* AWS Config monitoring
  + Using SQS queue: Send AWS Config SNS notification to one OR more SQS queue, then programmatically process the message – looking for non-compliant data, any undesired changes, maintain an up-to-date list of AWS resource inventory etc.
  + Using AWS CloudTrail: list out all the action performed on AWS Config by users, role, o by any AWS services.
  + Using AWS CloudWatch Event: Using CloudWatch event to detect and react on any change in the status of AWS Config events.
* **AWS Service Catalogue**: Key concepts
  + **Product** – A product is a service OR an application for the end-user. A Product can comprise of one-or-more AWS resources such as EC2 instance, Storage Volume, DB instance, monitoring configuration OR package AWS marketplace product or network components. It can be a single EC2 instance or a fully configured web-application or anything in between. When a user launches a product that has an IAM role attached to it, AWS Service Catalogue uses that IAM role to launch the products (AWS Resources/Application) using AWS CloudFormation template. By assigning IAM role to the products, one can avoid giving access to user to launch required AWS resources.
  + **Catalogue** – A catalogue is collection of products that the administrator creates, add to the portfolios and provided updates using AWS Service Catalogue. Administrator, creates catalogue of products by importing CloudFormation templates and also defines who can used those products and how they can use them.
  + **Portfolio** – Administrator distribute products to end-user in Portfolio. Administrator can create a customized portfolio for each type of users within the organization and selectively grant access to the users to access those portfolios. Granting access to user to access a portfolio , allows user to browse and launch product from the portfolio.
  + **Publishing Portfolio** – Add products to the portfolios, add IAM users, roles, groups to access the added products.
  + **AWS Catalogue Provision Stack** – The AWS CloudFormation stack that gets provision when an AWS Catalogue get executed is called “AWS Catalogue Provision Stack”.
  + **AWS Catalogue Version** – AWS Catalogue allows to manage multiple version of the same product in the Catalogue. When a new product version is added to the Catalogue its available to all the users to select from. Use can select the new version when creating a product from the catalogue OR can chose to update their running stalk to the newer version. [**AWS Catalogue DOES NOT automatically upgrade the running stalk to the new version – users need to act on it to upgrade the version.]**
  + **AWS Service Catalogue Scope and VPC Access** – Products and Portfolios are regional constructs; they are ONLY available on the region in which they are created – one can restrict in which region the data will be stored. Its possible for AWS Catalogue Provision Stalk resources to privately access VPC resources using VPC endpoints.
  + **AWS Service Catalogue constrains** – rules that limit the users from selecting parameter values from a specific set of values while launching a product from their service catalogue. It’s a means to apply cost control and governance. Constrains can be applied at each product level OR at the portfolio level. When there are multiple constrains at the product level and at the portfolios level – one which is more resistive will be applied. Constrains are applied while launching a new product OR while updating a running stalk with the newer version. Following are the different types of AWS Constrains
    - **Launch Constrains:** IAM Role can be attached to a Product in a portfolio.This role is use for provisioning product at launch, there by user permission can be resisted without impacting the user’s ability to provision product from there portfolio.
    - **Notification Constrains:** Notification constrain enable users to get notification about the stack event using an Amazon SNS topic.
    - **Template Constrains:** Template parameter rules are called as *template constrains* because this constrain are applied to the AWS CloudFormation template for selecting the parameter values while launching a product. One can reuse the same CloudFormation template and apply different template restriction per product basis OR per-portfolio basis.
  + **Sharing of the Portfolios between AWS accounts. Its possible to share portfolios between AWS two accounts, however the control remains with the account who has shared the portfolios with** 
    - The owner account (the account that shares the portfolios) can un-share the portfolio at any time.
    - The owner account can ONLY make changes to the portfolios like – adding new product(s), updating existing products.
    - Any product or stack that are currently in use will continue to RUN till stack owner decides to terminates those products/stacks. [Once the product/stalk is provisioned – portfolio owner account cannot delete them, ONLY the stack owner (account) which has provisioned the product can terminate those products]
    - On top of owner account constrains – provision account can add more constrains to it.
* **AWS System Manager (aka AWS Simple System Manager - AWS SSM):**  These are the collection of capabilities to configure and manage Amazon EC2 instances, on-premises server, virtual machines, and other AWS resources at scale.
  + AWS SSM includes a unified interface that allows to easily centralized operational data and automate task across an AWS client resources across region.
  + AWS SSM helps in effectively detect and resolve operational problem in AWS infrastructure – like finding out instance for which anti-virus are not running/not updated, out-of-sync configuration for instances etc.
  + AWS SSM gives a complete view of the client’s infrastructure performance, and configuration and simplified resources and application management.
  + AWS SSM works with – Windows, Lunix, RaspBian instance and virtual machine.
  + AWS SSM can manage remotely and securely – EC2 instances , on premises servers, virtual machines in cloud or on-premises. IT can also mange virtual machine running on other cloud providers.
  + AWS SSM can remotely and securely manage hybrid workload using same tools and scripts.
  + AWS SSM can be used to use IAM for centralized access control for actions that can be performed on the servers and VMs.
  + AWS SSM can be used to collect auditing information in single location by using AWS CloudTrail.
  + AWS SSM can be configured to use CloudWatch events and SNS notifications for centrally monitoring to send notifications about service execution success & failure.
* How AWS SSM Works?

The below diagram shows how AWS SSM works in details – it has three distinct phase **Configuration**, **Verification**, and **Reporting**.

**Configuration**: use AWS SSM console, SDK, AWS CLI, or AWS Toolkit for Windows PowerShell to configure, schedule, automate, execute SSM actions on AWS or On-premises resources. Install SSM agents and configure permissions.

**

**Validation**: AWS SSM verifies the configuration including permissions and send request to SSM agents running on AWS instance or in servers in hybrid environments. SSM agent perform the specific configuration changes.

**Reporting**: SSM Agent reports the status of the configuration changes and actions to AWS SSM. AWS SSM sends (forward) the status to the user and various AWS server if configured.

* For connecting SSM agent to SSM service hosted on AWS infrastructure.
  + For instance, in public subnet, SSM agent can connect to the SSM service directly OR using AWS SSM VPC endpoints.
  + For instance, in private subnets one need to configure NAT gateway to route the SSM agent request to SSM Service through internet OR through SSM VPC endpoints.

NOTE: VPC endpoints are more preferred approach.

* Required permission for AWS SSM
  + IAM User permission – Needed for those users without administrator access.
  + IAM Instance Profile – Needed for the EC2 instance that will be managed by SSM. If the instance profile change, till the change credentials are reflected SSM agents will NOT work. SO, it’s advisable to restart the SSM agent when instance profile change to reflect the credentials change immediately.
  + IAM Service Role – Needed for on-premises servers/VMs

**03/31/2020**

* SSM Capabilities
  + **Resource Group** makes it easy to manage and automate task on large number of resources at one time. Resource group is a collection of AWS resources that are all in the same AWS Region and that matches criteria provided in the query. Queries can be built in the resource group console OR pass as arguments to Resource Group command in the AWS CLI. Resource group can be used to create custom console that organizes and consolidates information based on criteria that are specifies tags.
  + **Insights**
    - **Built-In insights:** Built-in Insight shows detailed information about a resource in the respective AWS Resources Groups such as
      * CloudTrail Logs
      * Result of evaluation against Config Rules
      * AWS Trusted Advisor reports

Single, or multiple resource groups can be selected at a same time.

* + - **AWS CloudWatch Dashboards**: Helps in viewing data of the manage instances.
    - **Inventory Management:** It’s an automated process of collecting software inventory from manage instances – it can be used to gather metadata about an application, files, component, patches, and many more on managed instance.If there are multiple Account spanning across multiple region – then Inventory manager need to use Resource data-sync to send all the inventory related data to a single S3 bucket from were data can be queried using amazon Athena. The data can send further to AWS Quick Sight or third-party tool to visualized the data.

*Inventory Manager store system data for 30 days – once the data is sent to s3 bucket one can configure required Lifecycle policy to persist the data in an cost optimized ways for a longer duration of time.*

* + - Configuration Compliance: System Configuration Compliance can scan the fleet of instances of managed instance for patching compliance and configuration inconsistency. One can collect data from multiple AWS Account and AWS region and then drill down on the data of those instances which are found non-compliant.

Other benefit of using System Configuration Compliance

* View compliance history and change tracking of Patch Manager patching data and State Manager association by AWS Config.
* Remediate issues by using system manger run command, state manager or Amazon CloudWatch event.
* Port data to Amazon Athena and Amazon Quick Sight to generate fleet wide reports.
  + **Actions** 
    - Automation: SSM can perform automated task on mange instance at scales. SSM simplifies the common maintenance and deployment tasks on managed instance like
      * Build automation workflow to configure and manage instances and AWS resources by using custom workflows OR using predefine workflows maintained by AWS.
      * Received notification of the automation workflow tasks using AWS CloudWatch Events.
      * Monitor Automation process and execution details by using the Amazon EC2 instance OR System Manage console.

Some of the example of task that can be automated by SSM Automation capabilities are:

* + - * Create or update Amazon Machine Images.
      * Apply Driver or Agent updates.
      * Reset passwords on windows instances.
      * Reset SSH key on Linux instances.
      * Apply OS patching OR application updates.

Automation Use case using SSM Automation capabilities

* Request that one or more IAM user approves the instance stop action.
* Automatically stop and start instance on a schedule by using CloudWatch events or by using Maintenance Window task.
* Update resources that are deployed by CloudFormation templates – additionally the workflow can be configure to have one or more IAM users to approve the task before updating the resources.
* Tag large/small groups of EC2 instances using Amazon EC2 tags that helps to role out changes according to the limit defined OR target a AWS resources group that includes multiple instances.
* Create Golden AMI from the source AMIs
* Included or exclude specific package from getting deployed.
* Recover Impaired instances.
  + - **RUN Command:** RUN Command can be used to remotely and securely manage the configuration of the manage instances at scale – it can be used to run administrative task OR ad hoc configuration at scale. RUN Command can be executed from AWS Console, AWS CLI, AWS Toolkit for windows powershell or using AWS SDKs. RUN Command can be configure to send its logs to CloudWatch events, if one desired to maintain the complete log then RUN Command can be configure to send its logs to S3 bucket. RUN Command can be configure to use CloudWatch event – to notify command execution status OR to it can be specify as target for CloudWatch event to execute specific documented commands on receiving a CloudWatch event.

RUN Command can be used for the following use cases to run on-demand changes at scale.

* Update Application
* Running Linux shell scripts OR windows PowerShell
  + For Linux instance use RUN Command with AWS-RunShellScript document
  + For Windows instance, use RUN Command with AWS-RunPowerScript document
* Install OR bootstrap applications
* Build a deployment Pipeline
* Capture the log files when the instance is terminated from the autoscaling group.
* Join instance to a windows domain.

**Troubleshooting RUN Command:**

* *Can’t find the desired instances, to run RUN Command.*

*Possibilities*

*- SSM agent on the instance are NOT running/installed*

*- EC2 instance are attached with IAM role that allows it communicate with the System Manager APIs.*

*- Verify the user account has IAM trust policies that enable account to communicate with the System Manager API.*

*- Ensure that SSM document supports the instance type. (Some of the SSM documents are specific to windows or Linux)*

* + - **Session Manager:** AWS SSM State manger is a secure and scalebale configuration management service that automate the process of keeping the manage instance in a desired state as define by the administrator.

SSM Session Manger helps in complying with the stick enterprise policies and security practices with fully auditable logs while providing an easy means to the end-users with simple one-click access.

**SSM Session Manager – How it works?**

If administrator need to connect to an (private) EC2 instance – instead of creating and configuring baston host for Ad-hoc login, they can quickly and securely connect to the EC instance using SSM Session Manager console or through AWS SSM Session Manger CLI behind the scene – AWS SSM Session Manager on receiving the first command to start a new session performs

1. Authenticate the login ID
2. Verify the action granted by the IAM policies
3. Check the configuration settings (such as verify the allowed limit of the sessions etc.)
4. Send a request to SSM Agent deployed on the EC2 instance to open a 2-way communication connection for the use to interact with the EC2 instance.

NOTE: As SSM Session manager used SSM Agent to open 2-way communication connection, it does NOT require inbound ports to be open on the managed instances.   
**AWS SSM Session Manager can be integrated with**

* *AWS CloudTrail*
* *Amazon S3*
* *Amazon CloudWatch Logs*
* *Amazon CloudWatch events and Amazon SNS.*
  + - **Distributor:** SSM Distributor capability an be used for deploying/publishing software packages to SSM Managed instances.Package can be custom packages OR AWS provided agent software packages.

*-Package can be delivered to specified mange instance using – RUN Command/State Manger based on - AWS AccountNumber/instanceID/tags/AWS Region*

-Package deployment can be automated using State Manger to schedule package for automatic deployment on target instances when those instance are first launched.

* + - **Patch Manager:** SSM Patch Manage can be used for automating the process of scanning managed instances for missing patches and to apply the patches. It can be used for scheduling recurring patching activities using System Manager Maintenance Window
      * **For Windows based instance:** ONLY security patches can be delivered using SSM patch manager.
      * **For Linux based instance:** Along with security patch, it can also be used to install/update package on the managed instance.
    - **Maintenance Window:** Use to setup recurring schedule for managed instances to execute administrative tasks like installing packages and updates without interrupting business critical operations.
    - **State Manager:** SSM State manage is a fully managed service by AWS, that allows maintaining fleet of EC2 instance through an interactive one-click browser base shell OR through AWS CLI – it allows accessing of the EC2 instance without opening the inbound ports OR manage baston host OR maintain SSH keys.

**AWS SSM State Manager can be used to do**

* *Instance are patched with specific software update.*
* *Bootstrap instances with specific software at the start-up*
* *Download OR update agents at define schedules*
* *Configure Network settings*
* *Join instance in windows domain*
* *Patch instance with software update throughout their lifecycle*
* *Run scripts on windows / Linux instances throughout their lifecycle [if one desired to run script in recurring basis thought the lifecycle of the instance then – State Manager is the right approach NOT RUN Command].*
* *AWS SSM - State Manager can be integrated with the AWS CloudTrail to provide all the logs of all execution for auditing purpose.*
* *AWS SSM – State Manage can be configure to save detailed command output into S3 bucket for viewing later*

**SSM State Manager Association with manage instances**

* *State manage uses SSM document to create an association*
* *Configuration defines the state that needs to be maintained on the instance. Example of an SSM State Manage State – antivirus is installed & running on the instance sand specific ports are close.*
* *State Manage association can also specify a schedule for when the configuration need to be (re)applied (Example: like every Thursday 3 PM)*
* *State Manage association can specify action to be taken when applying an configuration. Example – if the antivirus is NOT installed then state manage installed the antivirus – if the antivirus is installed but not in running state then it can turn-on the antivirus.*
  + Share Resources
    - **Managed Instance:** Instances that are managed by the AWS SSM – there are two types of managed instance – Standard and Advance Instances. Advance Instance enable AWS SSM to connect to hybrid machine by using AWS SSM Session Manger – Session manager also provides interactive (2-way-communication) shell access to the instances.
    - **Activation**:
    - **System Manager (SSM) Document**:
      * These are list if actions that are performed by SSM on managed instances.
      * It includes steps and parameters that one can specify for SSM to act upon.
      * SSM documents are written in JSON or in YML.
      * There are two types of SSM Documents
        + Command Documents: These documents are used by SSM and RUN Command.
        + Automation Document: These documents are used by SSM Automation.

SSM offers large sets of pre-configured documents that one can use by specifying parameters at runtime.

* + - **Parameter Store**: Parameter store secure credentials, passwords, security strings, software licenses that one didn’t want to add in the code (document) level, which can be shared securely with SSM components. There are two types of parameter store – Standard and Advance Parameter store. Standard parameter can store up to 4K and can be use free of cost. Advance Parameter can store up to 8K of data and come with additional cost. Advance parameter are chargeable based on the number of advance parameters store each month AND per API interaction.

AWS SSM Parameter Store provides a serverless, secure, hierarchical storage for configuration data management and secret management. Parameter store can be used for manage instance as well as for on premises instances. Parameter stored in Parameter store can be referenced using UNIQUE string value – the parameter store values can be stored in plain-text as well as encrypted text using KMS encryption (in case the parameter values is encrypted – instance accessing those values should have access to the encryption key.

**Benefits of using Parameter store**

* A secure, scalable, AWS managed secret management service.
* It stores configuration and secure string in hierarchical and trackable version with control and audit access at granular level.
* It can refer Secret Manager.
* Individual parameters can be tag and its access can be secure at different levels – operational, parameter, Amazon EC2 tags, or at path level.
* It can be integrated with other SSM capabilities and AWS services to retrieve secret and configuration data from a central store.
* Parameter store is supported by – Amazon EC2, Amazon ECS, AWS Lambda, AWS CloudFormation, AWS CodeBuild , AWS CodeDeploy, AWS KMS, AWS SNS, AWS CloudWatch, AWS CloudTrail.

**Parameters Store can be referred from**

* SSM Scripts, commands, SSM documents, configuration and automation workflow.
* SSM RUN Command, State Manager and Automation.
* Form AWS ECS and AWS Lambda.
* AWS CloudWatch Agent can store agent configuration in Parameter Store.

AWS Parameter store secret using secret manager, can be leverage to store & organized secret data like credentials, passwords, security strings, software licenses etc. AWS Service access Parameter Store – Parameter Store can refer Secret Manager for accessing secret. (Parameter store acts like a proxy for secret manager, and it doesn’t retain secret data or metadata.)

* **AWS SSM Inventory and other AWS Service**s: AWS SSM inventory provides a snapshot of the current inventory which can be use to mange software policies and improve the security of the entire fleet. AWS SSM Inventory management and migration capability can be enhance using following AWS services
  + **AWS Config:** It can provide historical record of changes to the inventory, along with the ability to create rules to generate notification when a configuration item is changed.
  + **AWS Application Discovery Services**: It is designed to collect inventory on OS type , application inventory, processes , connections and server performance metrics from on premises VMs for an successful migration to AWS infrastructure.
  + **AWS System Manager Resource Data Sync** can be used to send data from different AWS Service Manger gathering information from all managed instances into a single S3 bucket. Additional services like KMS can be used to encrypt the data store in the S3 bucket , which then can be used by AWS Services like AWS Athena and AWS Quick Sight to query and analyze the aggregated data.
* **AWS SSM Pricing** 
  + **Parameter Store Charges**: For using advance Parameter storages there is additional charges.
  + **Distributer Charges:** For AWS Software Packages there is NO CHARGES (FREE) for distribution and update. For Custom packages its CHARGEABLE.
  + **Mange instance Charges:** For standard instance up to 1000 on premises instances per account / per region are FREE. For more than 1000 instances – one need to covert all the instances in to “Advance” instance which will be CHARGEABLE
  + **Automation Charges**: Based on its usages, number of steps & types of steps its chargeable.

# **Section 6: Designing Network on AWS for Complex Organization and Hybrid Architecture**

Network refresher

* For a VPC CIDR once selected can’t be edited once created – possible CIDR block size are /28 (232-28 - 5 = 11 IP addresses) to /16 (232-16 - 5=65531 IP addresses). However, one can add a non-overlapping secondary CIDR block to the existing VPC to expand the size of the VPC.
* VPC Wizard options
  + VPC with single public subnet
  + VPC with public and private subnet
    - Custom route table will be associate with the public subnet with a route entry with IGW
    - Default route table will be associate with the private subnet with a route to NAT Gateway.
  + VPC with public and private subnet and Hardware VPN access
    - Custom route table will be associate with the public subnet with a route entry with IGW
    - Default route table will be associate with the private subnet with a route virtual gateway, which will have a connection to VPN or a Direct Connect (Direct Connection) to HQ.
    - Route propagation will be set to the default route table – so that it need not to be updated manually for the know route from the virtual gate to the default route table.
  + VPC with Private Subnet ONLY and Hardware VPN access
    - Create a Virtual Private Gateway with NO Elastic IP configured
    - Create a default route table with the route to virtual private gateway.
    - It will create a VPN Connection (user can configure Direct Connection if needed).

link : <https://docs.aws.amazon.com/vpc/latest/adminguide/Introduction.html>

**VPN Based connection**

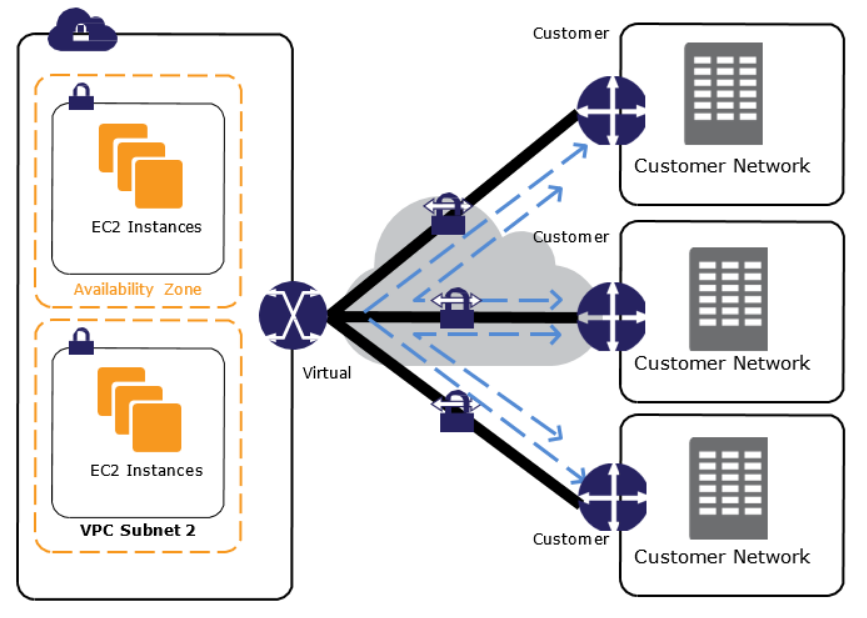
|  |  |
| --- | --- |
|  | * + Mostly desirable as it cost effective and takes less time to provision as no dedicated line needs to be created.   + On customer side Customer Gateway needs to be configured – on VPC side Virtual Private Gateway (VGW) needs to be configured.   + Customer Gateway can be physical devise or software application with internet routable IP address.   + The connection between Customer Gateway and Virtual Private Gateway (VGW) is established over IPsec Tunnels. Two connection are configured for each VPN Connection, to increase availability. Multiple customer gateway can be connected to a to single VGW   + NAT gateway can be reached through the VPN Connection. Connection initiated from the customer gateway cannot be route through NAT gateway. |

* + *How route table connected to the Virtual Gateway knows about the subnets behind the customer gateway?*

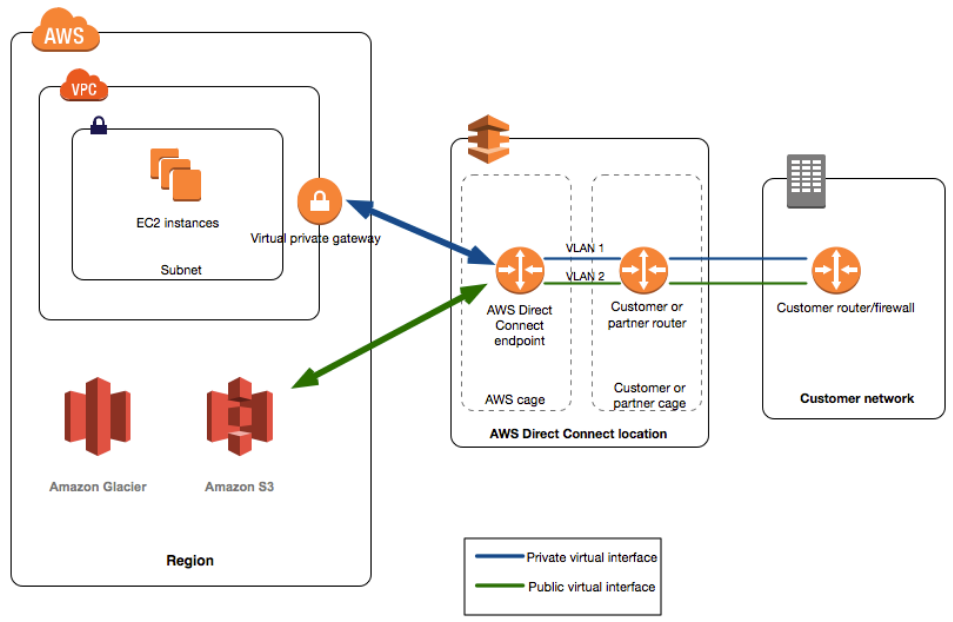
In order to connect VPN only subnets to the subnets from the customer datacentre one needs to updates the vpn-only subnet route tables to point it to the customer subnets and in customer side route table need to updated with the VPN only subnets information. This can be done by Manually updating the Route Tables Or by enable dynamic route propagation.

* *AWS VPN CloudHub:* AWS VPN CloudHub operates as simple hub-spoke model that can be use with or without VPC. *It’s a cost-effective option to connecting remote sites with AWS network.* It also be used as a backup connection to connect remote sites. AWS VPN CloudHub leverage AWS Virtual Private Gateway with multiple gateway with each using unique BGP autonomous system number (ASN). Customer gateways advertise appropriate routes over their VPN Connections, these routes are received and re-advertised to each BGP peer connected to the AWS Virtual Private Gateway. This allows connected remote sites to communicate with each other through VPN connection and also send & received data from the VPC. This option can be combined with AWS Direct Connect or other VPN options.

NOTE: The remote network prefixes for each spoke must have unique ASNs, and the sites must not have overlapping IP ranges.

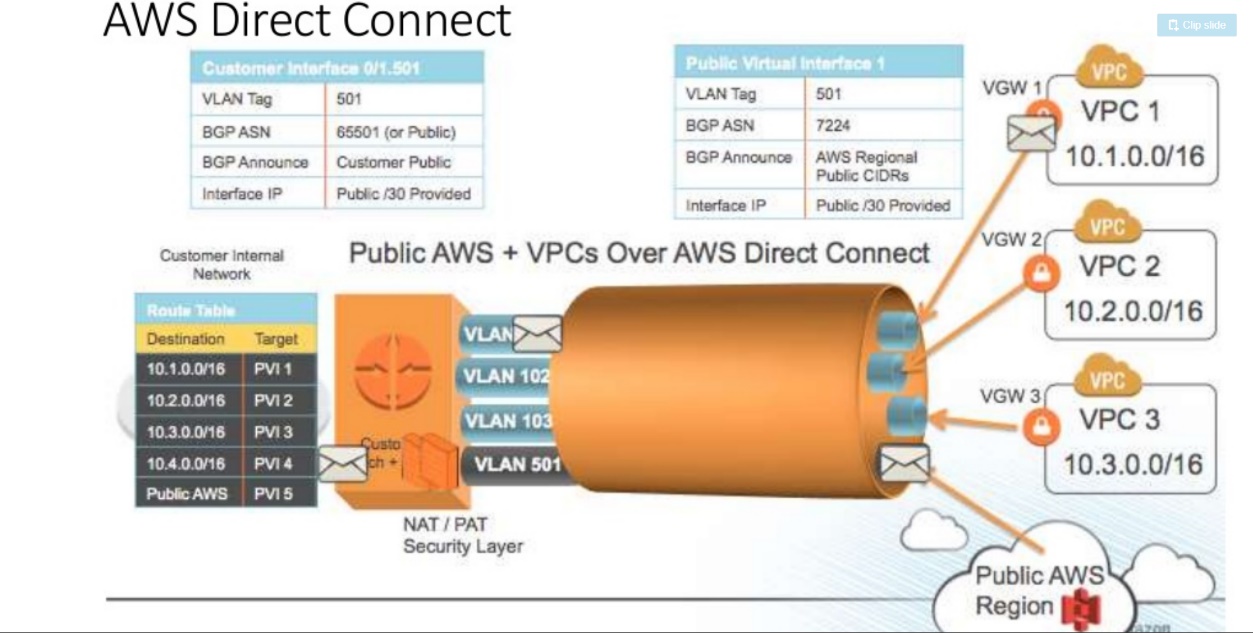
**

* ***CloudHub Pricing:*** *For the VPN uses hourly rate are applicable – and the data charges are applicable for the data send from the AWS network to the VPN connection.*
* For IP prefix to send and received traffic through VPN Connection –IP behind the customer gateway need to advertise the IPs by the customer gateway & the VPC IPs that allow to send/received traffic need to be advertised at the Virtual Private Gateway. ONLY IP prefix that are known to the virtual private gateway are allowed to communicate. VGW learn about the IP prefixes through or BGP routing.
* Direct Connect (AWS DX): AWS Direct Connect links internal network to a Direct Connect location over a standard Ethernet fibre-optic cable. One end of the cable is connected to the customer route and the other end of the cable is connect to the AWS Direct Connect Route. With this connection one can create virtual interface for Public AWS Services like S3 and bypass the internet while communicating with these services – make it fast and more secure. An AWS Direct Connect location provides access to AWS in the region with which its associated. The same connection can be used to connect to other AWS services in another public region.

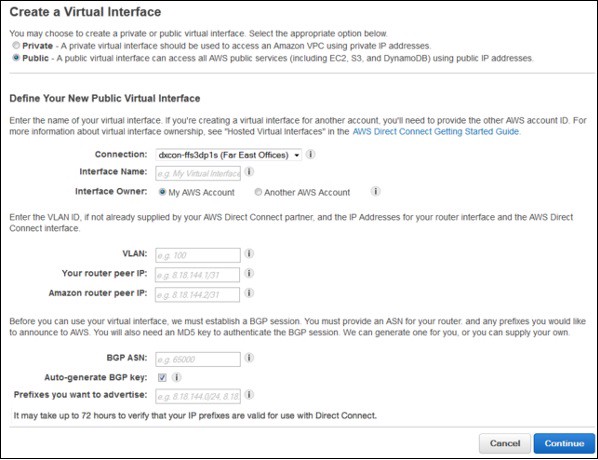


* Benefits of Direct Connect
  + Low latency and jitter (using direct connect one can bypass internet to connect to AWS services or VPCs).
  + Consistent performance and dedicated bandwidth.
  + Direct Connect enable customer to lower their Internet service provided bandwidth commitment.
  + AWS Charges lower data rate on Direct Connect then VPN connections.
  + Single Direct Connect Connection to build multi-region services - AWS Direct Connect to a public Regions or AWS GovCloud (US) can access public services in any other public Region, as wells as can access VPC in any public region.
* Requirement for establishing Direct Connect
  + Connections to AWS Direct Connect requires single mode fiber, 1000BASE-LX (1310nm) for 1 gigabit Ethernet, or 10GBASE-LR (1310nm) for 10 gigabit Ethernet. Auto Negotiation for the port must be disabled. Support for 802.1Q VLANs across these connections should be available.
  + Network must support Border Gateway Protocol (BGP) and BGP MD5 authentication. Optionally, we may configure Bidirectional Forwarding Detection (BFD). Border Gateway Routing (BGP) Connection must be used to connect between the two ends of the Direct Connect.
  + Customer Routing must be capable of 802.1Q and BGP Routing (in case of VPN connection Static routing is allowed along with BGP routing – in case of DirectX ONLY BGP is allowed).
  + Direct Connection can be used to established Private Virtual Interface (VIF) to connect to VPC in the same region (via the VPC’s VGW).
  + One can establish multiple private VIF to connect to multiple VPC within the same region.
  + One can also establish multiple public VIF to connect to any public AWS Service in any Region.
  + Direct Connect support both IPV4 and IPV6, but IPV6 needs public VIF.
  + All network traffic remains on the AWS global network backbone, whether one access public AWS service OR any VPC in another region. NO TRAFFIC LEAVES AWS NETWORK.
  + Any data transfer out of the remote region is billed at the remote region data transfer rate.
* 802.1q Link and Virtual Interface: *802.1q is industry standard Layer 2 protocol to segregate and isolate chunks of packets into separate VLAN channels. Each of the VLAN is configure to a separate Virtual Interface (VIF). This allows users to use the same connection to access public resources such as objects stored in Amazon S3 using public IP address space, and private resources such as Amazon EC2 instances running within an Amazon Virtual Private Cloud (VPC) using private IP space, while maintaining network separation between the public and private environments.*

*Up to 4019 channels (VLAN) can be possible in a standard connection. Virtual interfaces (VIF) mapped to each VLAN can be reconfigured at any time to meet your changing needs.*



* **Static Vs Dynamic Routing**: Routing is needed to correctly forward the traffic correctly. Static Routing - For simple network (with few routers) one can manually define the routes that is called static routing like in the case of AWS Route table. As the network grows (large number of routers gets added) maintaining static routing can be cumbersome, to overcome such administrative challenges Dynamic Routing (for example enabling route propagation between customer gateway router and AWS Virtual Gateway (VGW) when connecting to AWS network using VPN or Direct Connect.
* Direct Connect configuration page:



* BGP (Border Gateway Protocol):
  + Currently, BGP is the main internet routing protocol.
  + It can carry a large number of routes with granular route control.it has many route/path attributes to control routing.
  + BGP it connects different Autonomous Systems (AS) together.
  + Each of the Autonomous System (AS) should have an Autonomous System Number (ASN).
  + BGP creates a TCP based session between two peers on port 179.
  + BGP has two flavours – IBGP (Internal BGP – within a same Autonomous System) and EBGP (External BGP – within a different Autonomous Systems).
  + BGP can also have Transit AS capability, where peers connecting to a same AS can also communicate without establishing a peering connection between them.
* BGP Communities & Attributes:

**TO be added later.**

* AWS Direct Connect Inbound routing policies
  + Client should own the prefix that its advertising to AWS VGW.
  + Transitive routing through AWS network is NOT allowed – Traffic must be destined to the AWS prefix ONLY.
  + AWS Direct Connect perform inbound packet filtering to validate that the source of the traffic originated from clients advertise prefix. (Client should own those prefix).
* AWS Direct Connect Outbound routing policies
  + AWS Direct Connect advertise all local and remote AWS prefixes where available, including on-net prefix from other AWS non-region Point-Of-Presence (PoP) where available – example route53 / CloudFront etc. (NO ACCESS TO INTERNET OVER DIRECT CONNECT).
  + Inbound traffic Load sharing between multiple AWS Direct Connect can be achieved by advertising prefix with similar attribute (*AS\_PATH attribute should have equal path length for all the available connections* – AWS Direct Connect advertise prefix with minimum path length of 3).
  + To create a failover connection – secondary connection advertise prefix with higher path length then that of the primary connection.
  + AWS advertise all its public prefixes with well know NO\_EXPORT community. This prevent client routers to share AWS prefixes with other connected AS (Autonomous System) beyond the network boundaries of the client connection.

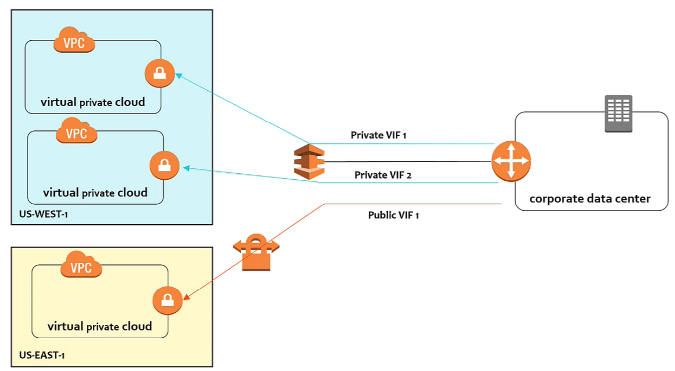
**AWS ROUTE TABLE NEEDS TO BE MAPPED TO VGW NOT TO THE DIRECT CONNECT OR THE VPN CONNECTION.**

* Link Aggregation Group (LAG)
  + Combining multiple direct connect links into a logical group using Link Aggregated Control Protocol (LACP) is called LAG.
  + Link Aggregation group is a logic interface that used Link Aggregation Control Protocol to aggregate multiple connection at a single AWS Direct Connect endpoint allowing it connect as single managed connection. All connection in the LAG are in active/active mode.
  + LAG can be created from the existing connection or can be provisioned from new connections – standalone connections can be added to the LAG at later point of time.
  + Rule to add connection to a LAG
    - The connection should be of same bandwidth.
    - Up to 4 (four) connections can be added to a LAG.
    - All connection LAG should be terminating at the same AWS Direct Connect Endpoint.
  + One can define number of active connections for a LAG to be operational - default value is set as ‘0’ however one can override the default value – in doing so, if the minimum number of connections is NOT available the LAG will become NON operational. This LAG attribute can be used to prevent overutilization of the remaining connection.
* AWS Direct Connect HA
  + By having a secondary connection build on VPN – which will allow traffic to reach to/from AWS cloud to the data-centre using VPN connection over the internet – this option can only be use as Failover scenario as VPN connectivity over internet is non-reliable.
  + By having a secondary connection build on Direct Connect – this option can be used with both Failover and Active/Active scenario.
* ***Longest prefix match*** – there are two route table entries 10.0.0.0/16 and 10.0.0.0/24. Second cidr block is part of the first cidr block in that case longest prefix match is applied. ONE with longest prefix in this case /24 will be selected. It will go for the specific.
* **Route Table priorities – From AWS to On-Premises** 
  + Route table on AWS side determines where the network traffic is directed.
  + For AWS to route traffic from VPC to On-Premises CIDR block - it needs a route entry for On Premises CIDR block destined to VGW (Virtual Gateway). [THIS IS TRUE FOR BOTH DIRECT CONNECT AND FOR VPN CONNECTIONS.] This can be achieved using Static Routing OR by enabling Route propagation where CIDR block entries are auto populated based on the route received from the BGP.
  + In case of the overlapping route entry, and Longest Prefix match can’t determine the preferred route, the following rules gets applied – from most preferred to least preferred
    - Local route will be preferred most.
    - Static route to Internal Gateway, Virtual Private Gateway, Network Interface of a instance , VPC peering Connection, A NAT gateway, VPC Endpoint will be the 2nd (second) preference.
    - BGP propagated route from the AWS Direct Connection is always preferred.
    - Manually add static routes for site-to-site VPN Connection.
    - BGP propagated routes from site-to-site VPN Connection.

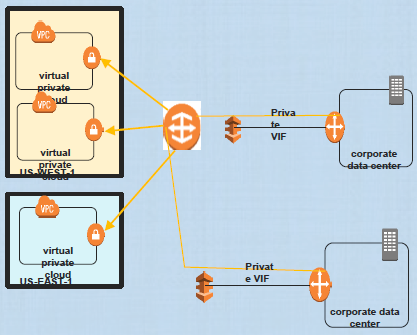
**[VIRTUAL PRIVATE GATEWAY ONLY ROUTES TRAFFIC TO KNOW PREFIXES THROUGH BGP ROUTE PROPAGATION OR THROUGH STATIC ROUTE ENTRY.]**

* **Direct Connect Gateway**

**Prior to using Direct Connect Gateway -** In case one needs to connect to different VPC (same or different region) using Direct Connect then – for the region where Direct Connection is connected (associated) one can create Private VIF (Virtual Interface) per VPC and connect. For other region one need to create a Public VIF (Virtual Interface) and connect it over VPN Connection (IPSec tunnels) OR work with AWS Direct Connect Partners to establish a new connection to AWS region of interest. This solution is expensive and cannot be scaled – as it required to maintain large number of VPN Connection (IPSec Tunnel).

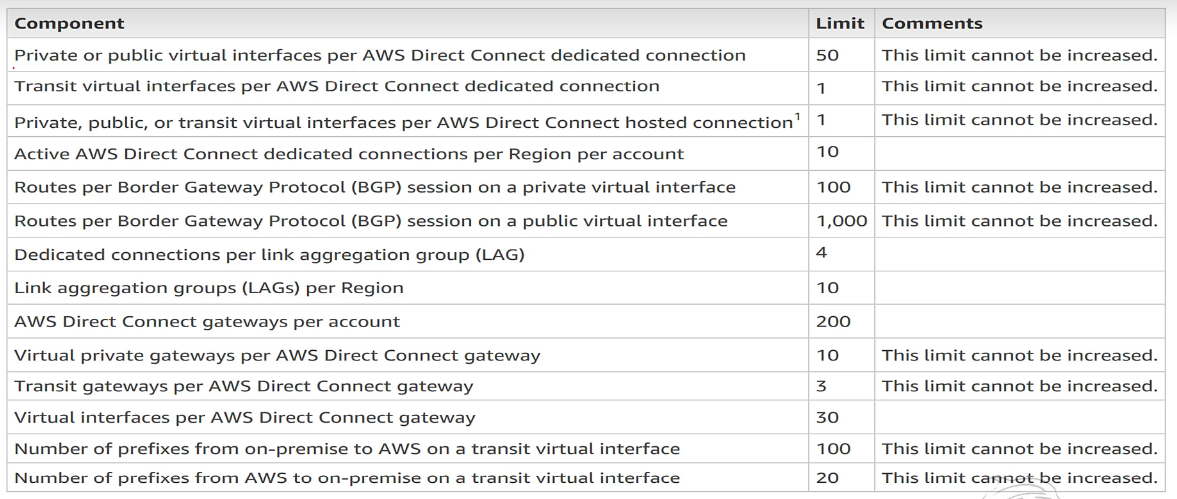


**When using Direct Connect Gateway:** The above mention problem was overcome after the introduction of the Direct Connect Gateway. A Direct Connect gateway is a globally available resource. One can create the Direct Connect gateway in any public Region and access it from all other public Regions (except China). It also possible to connect multiple direct connection to Direct Connect Gateway.



**The limitation of Direct Connect Gateway:**

* The VPCs that are connected through a Direct Connect Gateway can’t have overlapping CIDR block.
* Public VIF can’t connect to a Direct Connect Gateway.
* A Direct Connect gateway supports communication between attached private virtual interfaces and associated virtual private gateways only.
* Following connections are NOT SUPPORTED
  + Direct communication between the VPCs that are associated with the Direct Connect gateway
  + Direct communication between the virtual interfaces that are attached to the Direct Connect gateway
  + Direct communication between a virtual interface attached to a Direct Connect gateway and a VPN connection on a virtual private gateway that's associated with the same Direct Connection.
* One cannot associate a virtual private gateway with more than one Direct Connect gateway. Similarly, cannot attach a private virtual interface to more than one Direct Connect gateway.
* A virtual private gateway associated with a Direct Connect gateway must be attached to a VPC.
* **Direct Connect Limits**



* **VPC Peering**
* Inter region VPC peering is allowed – traffic between different VPC from different region travels over AWS backbone.
* There is NO SINGLE POINT OF FAILURE in case of VPC peering.
* Edge-to-Edge to peering is NOT allowed. For example, VPC A is peered with VPC B - VPC A is connected to corporate HQ using direct connect, the same connection cannot be used for connecting to VPC B.

Note: for connecting multiple VPC over VPC peering – one need to create a full mesh. Number of peering connection need to connect all the VPCs are n (n-1)/2 for example for 3 VPCs to connected 3 \* ((3-1)/2) = 3 peering connection.

* **AWS Transit Gateway**
  + Transit Gateway is a network transit hub that can be used to interconnect VPCs and Corporate HQ.
  + It’s a regional construct – only able to connect VPCs within the same region.
  + Though Transit Gateway, connected VPCs can communicate with each other and also with the corporate HQ – however this behaviour can be alter by creating multiple route tables , to create a control access within the connected VPCs and Corporate Data Centre.
  + Transit gateway can be use to connect to VPCs within across accounts.

Ref Link : <https://docs.aws.amazon.com/vpc/latest/tgw/tgw-transit-gateways.html>

**Transit Gateway important concepts**

* + - VPCs / Direct Connect / VPN Connection – all these can be attached to a transit gateway.
    - Default – transit gateway is connected to a single route table. Additional route tables can be attached to segregate / control the route.
    - By default, VPN and Direct Connect are attached to the default table.
    - Each association are attached to single route table – route define in the route table are the ONLY association where communication is possible.
    - Route Propagation can be enabled for the Transit Gateway and the VPN / Direct Connect Connections allowing corporate data centre to connect to all connected VPCs.

NOTE: Transit Gateway cannot be connected to a Direct Connect – it can ONLY be connected to the through a Direct Connect Gateway.

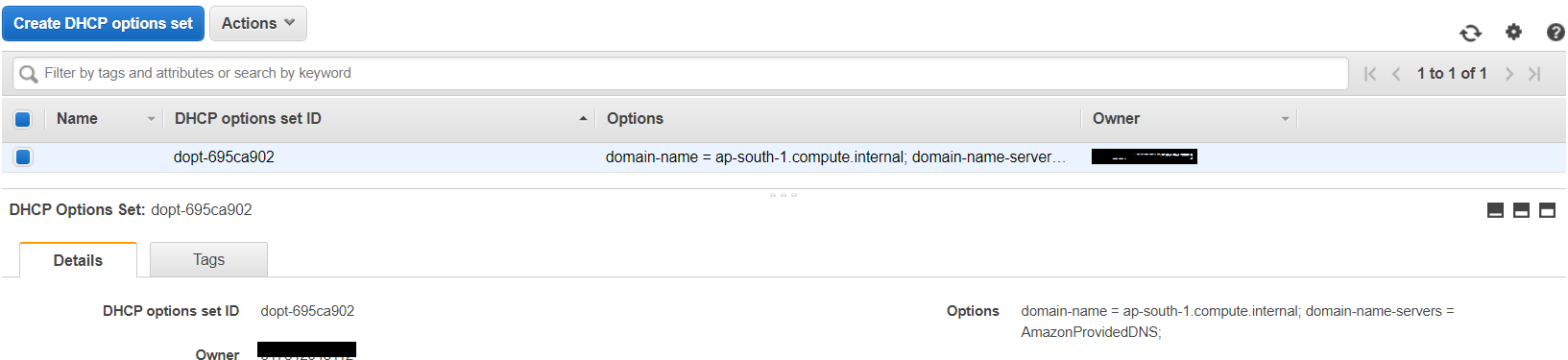
**[Revision Topics: Revision Topics EC2: SR I/OV – Enhance Networking – Spread / Cluster Placement Group.]**

**[Revision Topics: Difference between NAT Gateway and NAT instance.]**

**[Revision Topics: Baston Host – HA: Create Baston Host behind an autoscaling group – set desired size as 2 (to ensure always 1 Baston Host availability) & select multiple AZ, for cost optimization set desired size as 1 and select multiple AZ.]**

**[Revision Topics: EC2 – ENI - eth0/eth1 – Attaching and De-attaching eth1 from/to EC2 instance retains its MAC Address/Elastic IP/IPv4-IPv6 addresses, this feature can be helpful in installing software that tied to MAC/IP address. (ENI can be re-attached to another instance within the same AZ but can be in a different subnet. OTHER usages of having a secondary ENI attached to the EC2 instance – to have two distinct IP for client and for management request connected to two different subnet – public for client and private for management request.)]**

* **VPC Endpoint** 
  + For accessing the AWS Public Services from public subnet, one need to add a NAT Gateway to access all these services OR they can be accessed using VPC Endpoints.
  + When communicating over the VPC endpoint the traffic remains on the AWS network backbone and it does not travel through open internet.
  + There are two types of VPC Endpoints – Gateway Endpoint and Interface Endpoint.
  + VPC Endpoint can be ONLY used within the region.
  + ONLY one is required per VPC.
  + VPC Endpoint Gateway – is ONLY for S3 and Dynamo services
  + VPC Endpoint Interface – For most of the services we have VPC Endpoint Interfaces.
  + VPC Endpoint Private DNS Name -
* **VPC Endpoint Flow Logs** 
  + VPC Flow logs helps to capture the IP ingress/egress traffic to the VPC.
  + The VPC Flow logs can be created at the VPC Level, at subnet level and at the ENI level. Depending upon the desired granularity one can decide the where to enable the VPC Flow log. Additional one can select the type of traffic it needs to be logged – rejected traffic, accepted traffic or both.
  + Flow logs can be sent to Amazon CloudWatch Logs OR to AWS S3 bucket.
* **VPC Endpoint DHCP Option Sets** 
  + One can use on-premises DNS host for AWS instances; however, AWS DNS host are NOT allowed to use for on-premises instances.
  + Dynamic Host Configuration protocol provide a standard for passing configuration information to host on TCP/IP network. It’s a standard to share information which DNS servers to be used for resolving the DNS names.
  + BY default, DHCP option set is created within the VPC pointing to the Amazon DNS server, with domain-name and domain-name-servers option populated.

****

* + One can create custom DHCP (cannot edit the default DHCP option) connect and attached to the VPC, at a given time ONLY one DHCP option can be attached to the VPC.
  + Once a new DHCP Option is connected to the VPC, instance need not to be restarted, after some time (4-5 hours) the news DNS will be automatically picked up.

# **Section 7: Introduction to Digital Certificate and Public and Private Share Encryption**

* X.509 certificate, SSL, TSL are synonymous.
* X.509 certificates are based on – ITU-T-X.509 format.
* Different Types of Digital Certificates are
  + Personal: Used by individuals requiring secure email, digital signature, and web base transactions.
  + Organizations: Used by corporate to identify employee for secure email and web base transactions.
  + Server: To prove domain name ownership and establishing SSL/TSL base encrypted session between their webserver and a client.
  + Developer: To prove code authorship and retain integrity of distributed software.
* CA Authority actions
  + Accept Certificate Request (issuing) application from the entity.
  + Authenticate issuing application to validate its correctness.
  + Issues Certificate.
  + Maintain Details and status about the issuing certificate.
* CA Public key are wide share – even share within web browser to evaluate other certificates without any delay.
* Asymmetric Key / Symmetric Key encryption:
  + Asymmetric Key are less preferred as compare to Symmetric Keys.
  + In case of Asymmetric Keys there are pairs of keys available – Private/Public Key. Client encrypts the data (plain text) using Public Key and Server decrypts the cypher text using the Private Key.
  + Symmetric Key encryption is more efficient then Asymmetric Key encryption which is also knows as Shared Secret key.
  + **AWS uses Symmetric Key encryption for KMS – for sharing the keys its used Asymmetric key.**
  + In case of Symmetric Key encryption there is ONLY ONE key – Client uses the key to encrypt the plaintext – server use the same algorithm (encryption-decryption) algorithm to decrypt the cypher text message.
  + CA and Intermediators for the certificate to be authenticate its chain of trust needs to be authenticated.

# **Section 8: AWS Elastic Load balancer – Network / Application / Network**

There are three types of Elastic Load balancer available – Application Load Balancer (Layer 7 Load balancer)/ Network Load Balancer (Layer 4), AND Classic Load Balancer (Operates at both layers 7 & 4 but now delicate by AWS).

Enabling Proxy Protocol for NLB (TSL/SSL connection – layer 4 ) – As NLB terminates the incoming request and start a new request to the backend services, for backend services it shows source as NLB instead of the actual client – Enabling proxy protocol ensure that original source is maintain however if the inbound request has pass through an proxy there will be a double proxy header added to the original request.

Enabling Proxy Protocol for ALB (HTTP/HTTPS connection – layer 7) – As ALB terminates the inbound request on receiving, and initiates a new request to the backend services. For backend services to know the actual client address it creates and add original client IP into x-forward-for header custom header. Backend services that desires to filter request based on the actual originator IP addresses, then it needs to process the information from x-forwarded-for header.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case** | **Front-End Protocol** | **Front-End Options** | **Back-End Protocol** | **Back-End Options** |  |
| Basic TCP Loadbalancer | TCP | NA | TCP | NA | Supports Proxy Protocol header |
| Secure website or application offloading SSL Decryption | SSL | SSL Negotiation | SSL | NA | Required SSL certificate to be deployed to Elastic Loadbalancer  Support Proxy Protocol header |
| Secure website or application using end-to-end encryption with elastic Loadbalancer | SSL | SSL Negotiation | SSL | Backend validation | Required SSL certificate to be deployed to Elastic Loadbalancer.  Doesn’t support insertion of |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case** | **Front-End Protocol** | **Front-End Options** | **Back-End Protocol** | **Back-End Options** |  |
| Basic HTTP Loadbalancer | HTTP | NA | HTTP | NA | Support x-forward-for header. |
| Secure website / Application using Elastic Loadbalancer to offload SSL decryption | HTTPS | SSL Negotiation | HTTP | NA | * Required loading of the certificates to the Elastic Loadbalancer. * Supports x-forward-for header. |
| Secure website or application using end-to-end encryption | HTTPS | SSL Negotiation | HTTPS | Backend validation. | * Required loading of the certificates to the Elastic Loadbalancer. * Supports x-forward-for header. |

[Revision topic]

ELB – Cross Zone Load Balancing

**ELB health check** – there are two type of health checks – TCP (layer 4 response) and HTTP (layer 7 response) checks. For the legacy application, if it doesn’t provide standard HTTP response 200 then ELB will mark the instance as unhealthy, in-order to overcome the same configure the check to TPC port. For nonstandard/custom/legacy application which host its application in other port apart from standard HTTP port 80 then it’s advisable to use TCP based health check instead of HTTP based health check.

**Sticky Session (server affinity)** Best practice to store the session information in persistence store. Server Affinity needs SSL certificate offloading at the Loadbalancer side.

loading **X.509 certificate** to ELB or to IAM or managed by ACM. Certificate should be in the same region as that of the ELB.

**ELB** can **managed cookies** if application is not managing the cookies.

ELB Doesn’t supports multiple SSL certificates. ELB Cannot be used for two-way mutual certification. Does not supports client side x.509 certificate authentication (two way SSL authentication). One needs to enable proxy protocol over Layer 4 (TSL) connection and the backend EC2 server should handle the 2-way authentication.

Server Name Indication (SNI) is an extension to the Transport Layer Security (TLS) computer networking protocol by which a client indicates which hostname it is attempting to connect to at the start of the handshaking process. ELB does not supports SNIS certificates. [**Application Load Balancers Now Support Multiple TLS Certificates With Smart Selection Using SNI**] : <https://aws.amazon.com/blogs/aws/new-application-load-balancer-sni/>

|  |  |
| --- | --- |
|  | SSL and TSL are not interchangeable terms – TSL is applicable to the Layer 4 (network layer) while HTTP is applicable to Layer 7 (application layer).  TLS is a protocol for securely transmitting data like passwords, cookies, and credit card numbers. It enables privacy, authentication, and integrity of the data being transmitted. TLS uses certificate-based authentication where certificates are like ID cards for the websites. HTTPS is basically HTTP connection which is delivering the data secured using SSL/TLS.  With SNI support we’re making it easy to use more than one certificate with the same ALB. The most common reason you might want to use multiple certificates is to handle different domains with the same load balancer. It’s always been possible to use wildcard and subject-alternate-name (SAN) certificates with ALB, but these come with limitations. Wildcard certificates only work for related subdomains that match a simple pattern and while SAN certificates can support many different domains, the same certificate authority has to authenticate each one. That means one have to reauthenticate and reprovision your certificate everytime you add a new domain.  ALB Access Logs now include the client’s requested hostname and the certificate ARN used. If the “hostname” field is empty (represented by a “-“) the client did not use the SNI extension in their request.  Certificates can be stored in ACM or IAM. The certificates created by ACM are of 4096 key length and EC certificates. One cannot use the certificates directly on the ELB – alternatively one can create and get CA Signed uploaded to ACM which can be used in ELB.  ALB allows binding multiple certificates for the same domain(s) to a secure listener. Your ALB will choose the optimal certificate based on following factors including the capabilities of the client. If no SNI is used, ALB selects the default certificate (defined during ALB creation). One can bind up to 25 certificates per load balancer (not counting the default certificate).   * Public key algorithm (prefer ECDSA over RSA) * Hashing algorithm (prefer SHA over MD5) * Key length (prefer the largest) * Validity period. |

When testing ELB – its important to consider global location. If one uses same location (cached dns record) the entire load testing request will be routed to ONLY one single instance which address has been cached and the load will NOT be distributed equally. To overcome this – ensure that test client does DNS resolving in every request.

ELB request can be log using

* **AWS CloudWatch**:(enable by default) ELB service sends metrics to AWS CloudWatch in every 1 minute. THIS does not include client information like source IP address etc. Collected metrics can be used to send alerts or notification if a certain limit is reached.
* **ELB Access logs**: (disable by default) Information about the requester like – requester IP address, request time, request type, etc. Logs can be redirected to S3 bucket an retain it longer as required. There is no cost of enabling ELB access logs, and the data transferred to S3 bucket – ONLY S3 bucket storage needs to be paid.
* **AWS CloudTrail**: capture all the API related call made to the ELB, like enabling and disabling of the access logs.

Classic Load Balancer shortcoming – Benefit of using Application Loadbalancer.