Cloud Practitioner Review – Infrastructure and Pricing

Prepared By: Chandra Mohan Lingam, Cloud Wave LLC

Benefits of Cloud Computing

1. Trade capital expense for variable expense
    a. No need to purchase expensive hardware in the cloud
    b. Pay based on usage
2. Benefit from massive economies of scale
    a. AWS buys in bulk as it needs to support 1000s of customers
    b. Shared infrastructure – reduces idle resources
    c. Better utilized
    d. Lower pay as you go pricing
3. Stop Guessing about capacity
    a. Scale up or down only with a few minutes notice
    b. Match infrastructure with actual need
4. Increase speed and Agility
    a. In the cloud, new resources are only a click away
    b. Hourly pricing – try new products at a low cost
5. Stop spending money running and maintaining data centers
    a. Avoid undifferentiated heavy lifting.
    b. In a traditional on-premises data center, you need to spend upfront money on purchasing physical servers, storage, networking equipment, datacenters, and so forth. Whereas, in the cloud, you pay only for what you use
    c. Focus on projects that differentiate your business
6. Go Global in Minutes
    a. Deploy application close to the users
    b. Use AWS regions, Edge Locations

AWS Global Infrastructure

AWS Global Infrastructure consists of multiple regions. There are 24 regions (and expanding) currently

For example, US West (Oregon), US West (Northern California), US East (Ohio), US East (N. Virginia), Europe (Frankfurt), Europe (Ireland), Asia Pacific (Tokyo), Asia Pacific (Mumbai) and so forth

All regions are global/public regions – that means you can spin up EC2 instances and other resources in any of the regions

However, there are two special regions

AWS GovCloud is only accessible to US entities that pass a screening process. Designed to host sensitive and regulated workloads to meet US Government security and compliance requirements

AWS China region complies with China's legal and regulatory requirements. This is a separate account that limits access only to AWS China regions

AWS Region

Each region consists of multiple, isolated, and physically separate Availability Zones within a geographic area.

For example, US West (Oregon) region has four availability zones within 60 miles of each other.

> Oregon Region Name: us-west-2
>
> Oregon Availability Zones: us-west-2a, us-west-2b, us-west-2c, us-west-2d

Availability Zone

An availability zone (AZ) consists of one or more data centers with redundant power, networking, and connectivity in an AWS region

All availability zones (AZ) in a region are interconnected with high speed, high bandwidth, and low latency networking

An architecture best practice is to deploy an application across multiple AZs. This will protect your application from issues such as power outages, lightning strikes, tornadoes, and more

Edge Locations

AWS also has 200+ edge locations worldwide. The Edge locations are used by CloudFront (Content Delivery Network) to cache content close to users

This allows you to deliver content to end-users with low latency

Data and Regions

All your data AWS is stored within a specific region that you choose.

To further increase redundancy and fault tolerance, you can replicate data across AWS regions.

To copy data between regions, you need to explicitly ask AWS service to do so.

For example, you can configure S3 to replicate the content of a bucket to another bucket in a different region

You can configure Relational Database Service (RDS) to maintain a read-replica in a different region

You can configure DynamoDB Global Tables to maintain a copy of data in multiple regions and automatically synchronize changes across regions

Note: CloudFront and Edge Cache also maintain copies of data in multiple edge locations. This is used primarily for performance reasons [and not for redundancy or fault tolerance]. When edge cache expires, it will go to origin to get the latest data

AWS Personal Health Dashboard

AWS personalized health dashboard provides alerts and remediation guidance when AWS is experiencing an outage.

The service health dashboard displays the status of AWS services, and the personalized health dashboard gives insight into your resources that are impacted due to AWS infrastructure issues

AWS Pricing

AWS pricing is based on

1. Compute (No. of hours of compute used per month)
2. How EC2 instance, Compute Capacity was purchased
3. Storage (GB of data stored per month)
4. Data Transfer (GB of data transferred out each month)
5. Region

NOTE: AWS offers a free tier for new customers. This free-tier period is for 12-months and covers some specific services and usage. SageMaker comes with a 2-month trial period. Any use that exceeds the free-tier quota is billed at on-demand rates.

Per-Second Billing

EC2 usage is billed in per second increments with a minimum of 60 seconds. Per-second billing is available for Linux, Windows, Ubuntu instances.

Compute Pricing

You can purchase compute capacity using

1. On-demand
    a. No upfront payment or long-term commitment
    b. Recommended for short-term, spiky, or unpredictable workloads that cannot be interrupted
    c. Recommended when you are trying out AWS for the first time
2. Reserved Instances
    a. Significant discount when compared to on-demand pricing
    b. Standard Reserved Instances provides a discount of up to 72%, and it is region and instance family-specific
    c. Convertible Reserved Instances provides a discount of up to 54%, and it is region-specific. However, you are free to change instance family
    d. Requires 1-year or 3-year commitment
    e. Recommended for steady-state or predictable usage
    f. Three payment terms: No upfront, Partial upfront, All upfront
    g. No upfront payment - There is no upfront payment, and you pay a reduced hourly rate each month
    h. Partial upfront payment - Provides a higher discount than no upfront. Part of the usage is paid up front and you pay a smaller reduced hourly rate each month
    i. All upfront payment – Highest discount. Usage for the entire period is paid up front and there is no hourly charge for rest of the term
3. Savings Plan
    a. Three types of savings plans: EC2 Instance Savings Plan, Compute Savings Plans, SageMaker Savings Plans
    [My opinion – Compared to Reserved Instances, Savings plans are more flexible. This may be the future of AWS pricing]
    b. Requires 1-year or 3-year commitment

    c. EC2 Instance Savings plan applies to EC2 usage, and it is region and instance family-specific (up to 72% discount). This is similar to Standard reserved instances
    d. With SageMaker Savings Plans, you can reduce your SageMaker machine learning usage costs (up to 64% discount) irrespective of the region, instance family, and size.
    e. With Compute Savings Plan, you can get a significant discount (up to 66%) on any compute service such as EC2, Lambda, Fargate Containers (applies to both server-based and serverless compute)
    f. Compute savings plans automatically lowers cost irrespective of the region, instance family and size
    g. The cost explorer tool (available in the billing dashboard) can analyze your historical on-demand usage over a selected period (7, 30, or 60 days) and recommend a savings plan to maximize savings.
    h. Three payment terms: No upfront, Partial upfront, All upfront
4. Spot Instances
    a. Request for unused AWS capacity at a steep discount (up to 90% off on-demand pricing)
    b. AWS can terminate an instance with 2-minute notice [i.e., your workload can be interrupted any time]
    c. Recommended for applications that have flexible start and end times, flexible to use different instance families and types, and urgent computing needs that require a large amount of compute capacity
    d. Suitable for workloads that are fault-tolerant [i.e., workload that can safely continue in another instance when interrupted]
5. Spot-Block
    a. Spot-Block is for workloads that need to run continuously for 1 to 6 hours
    b. Request for unused AWS capacity (discount up to 45% off on-demand pricing)
    c. When spot instance capacity is available for the requested duration, the request is fulfilled
    d. An instance is automatically terminated at the end of the time block
    e. [My Opinion - It appears spot block is being phased out as new accounts are not eligible for spot blocks]
6. Dedicated Host
    a. Physical EC2 server dedicated for your use
    b. Can be purchased on-demand or as a reserved instance
    c. Useful when you need to bring your own software license to AWS cloud [such as SQL Server, SUSE Enterprise Linux Server, Windows Server]
    d. Useful when your software license is tied to sockets or physical cores

Data Transfer Pricing

1. Same Availability Zone – free when using Private IP
2. Same Availability Zone – USD 0.01/GB when using Public or Elastic IP
3. Same Region – USD 0.01/GB
4. Another Region – USD 0.02/GB
5. From Internet to AWS – free
6. From AWS to the Internet – USD 0.09/GB

Pricing Calculators

1. Simple Monthly Calculator – Helps estimate monthly AWS bills. Explore various AWS services, and create an estimate for your use case on AWS
2. AWS Pricing Calculator – Same capability as Simple Month Calculator [replacement for a simple monthly calculator with better User interface]
3. TCO Calculator – Total Cost of Ownership Calculator compares the cost of running your workload in an on-premises data center and AWS cloud. Useful when you need to plan for cloud migration

AWS Organizations

A lot of companies use several AWS accounts (some even have 100s of accounts). AWS Organizations helps you centrally manage all your accounts

1. You can organize various account as hierarchies
2. Centrally secure, control, and audit your accounts. For example, which region to use, what AWS services are allowed
3. Simplify billing by consolidating expenses across all accounts – single payment for all accounts
4. Aggregate usage across all accounts for volume discounts [AWS has a tiered pricing band. For example, if you store more data, you will get a better per GB storage pricing]
5. Share reserved instance discounts and Savings plan discounts across the entire organization

Cost Explorer and Usage Alerts

You can gain visibility of charges using these tools

1. AWS Budgets – Set your monthly budget and configure alerts when actual usage or forecasted usage exceeds your budget threshold
2. AWS Bills – At the end of each billing cycle, you will receive a monthly invoice of usage charges. In addition, you can also view the past bills and current month bill from the Billing dashboard
3. AWS Cost Explorer – Interactive tool to explore, understand, visualize, and manage costs over a time period. Useful to detect trends, pinpoint where you are spending money, and forecast future costs. Identify cost optimization opportunities using Savings plan
4. AWS Cost and Usage Report – Detailed, low-level cost and usage data in CSV format. Useful when you want to do your own analysis
5. Cost-allocation tags – Organize your resources using tags and track AWS usage costs at detailed level (by cost center, by environment-development/test/production, by application and so forth)
6. CloudWatch Alarm – Estimated Charges are calculated several times daily and published to CloudWatch as a metric. You can also configure alarms based on this metric
7. Simple Notification Service (SNS) – SNS Topic allows you to create an Email, SMS distribution list. You can configure Budget alerts and CloudWatch Alarm to notify an SNS topic whenever the usage exceeds a configured threshold. SNS also allows you to invoke other services such as the Lambda function in response to an alert or event.

Trusted Advisor

AWS Trusted Advisor – Compares your AWS environment against AWS best practices and provides recommendations across five categories:

1. Cost optimization – eliminate unused and idle resources, analyze on-demand usage costs and recommend reserved instances and Savings plan (data sourced from cost explorer)
2. Security – close security gaps, examine permissions, enable security features
3. Fault tolerance – increase the availability and redundancy using auto scaling, health checks, multi-AZ and backup capabilities
4. Performance – monitor for overutilized instances, ensure usage of provisioned throughput, verify service limits that can cause performance issues
5. Service Limits – checks for service usage that is more than 80% of the service limit
6. All customers have access to seven core trusted advisor checks. Business/Enterprise customers have access to the full set of checks

Seven Core Trusted Advisor Checks

- Checks security groups for rules that allow unrestricted access (0.0.0.0/0) to specific ports. Unrestricted access increases opportunities for malicious activity
- IAM Use - This check is intended to discourage the use of root access by checking for existence of at least one IAM user
- S3 Bucket Permissions - Checks buckets that have open access permissions or allow access to any authenticated AWS user
- MFA on Root Account - Checks the root account and warns if multi-factor authentication (MFA) is not enabled
- EBS Public Snapshots - Checks the permission settings for your Amazon Elastic Block Store (Amazon EBS) volume snapshots and alerts you if any snapshots are marked as public. When you make a snapshot public, you give all AWS accounts and users access to all the data on the snapshot
- RDS Public Snapshots - Checks the permission settings for your Amazon Relational Database Service (Amazon RDS) DB snapshots and alerts you if any snapshots are marked as public. When you make a snapshot public, you give all AWS accounts and users access to all the data on the snapshot
- Service Limits – Alerts you about services that use more than 80 percent of a service quota. If the number of servers is approaching your account limit, you cannot launch any more servers, which will prevent autoscaling from responding to traffic.

Here are some additional trusted advisor findings,

- Identify idle and underutilized resources
- Check for age of snapshots (ensure backups are up to date)
- Check for database instances deployed in a single-AZ (risk of AZ failure)
- High utilization EC2 instances
- Large number of security group rules that can cause performance degradation

Support Plans

AWS offers the following support plans

1. Basic

    a. Free and included for all customers
    b. You can contact AWS support for questions related to billing, account, and for increasing service limits
    c. Use the web-form to open a case
    d. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
2. Developer
    a. Recommended if you are experimenting or testing in AWS
    b. Pricing starts at $29/month or 3% of monthly AWS usage (whichever is greater)
    c. Email-based tech support during business hours
    d. Response time is 24 hours. For system impaired issues, less than 12 hours
    e. Access to 7-core Trusted Advisor Checks and Personal Health dashboard
3. Business
    a. Recommended if you have production workloads in AWS
    b. Pricing starts at $100/month or 10% of monthly AWS usage (whichever is greater)
    c. Contact Tech support - 24x7 Phone, email, and chat
    d. Response time is less than 1 hour for Production system down
    e. Access to the full set of Trusted Advisor Checks
    f. Paid access to Infrastructure Event Management (additional planning and support for special high traffic events such as your new product launches, sports broadcast)
4. Enterprise
    a. Recommended if you have a business or mission-critical workload in AWS
    b. Pricing starts at $15,000/month or 10% of monthly AWS usage (whichever is greater)
    c. Contact Tech support - 24x7 Phone, email, and chat
    d. Response time is less than 15 minutes for Business-critical system down
    e. Access to the full set of Trusted Advisor Checks
    f. Designated Technical Account Manager (TAM) – your single point of contact for AWS
    g. Infrastructure Event Management is included for free
    h. Concierge Support Team
    i. Well architected reviews, consultations, and guidance

Shared Responsibility Model

Security and compliance is a shared responsibility between AWS and the Customer. [Customer here refers to one who created an AWS account]

AWS is responsible for “security OF the cloud.”

Customer is responsible for “security IN the cloud” – another way to think about this is anything that you store and do in the cloud is your responsibility.

AWS Responsibility – Security “OF” the cloud

AWS is responsible for protecting the infrastructure. For example, AWS is responsible for the physical security of the facilities. AWS is also responsible for protecting the cloud from infrastructure level attacks such as a distributed denial of service (DDOS).

The infrastructure is composed of hardware, software, networking, and facilities that run AWS cloud services

AWS is responsible for patching and fixing flaws in the infrastructure - Such as physical server security, patching the physical host operating system, and so forth

AWS also has increased responsibility in managed services such as S3 and DynamoDB. They are responsible for infrastructure, operating system, software

Customer Responsibility – Security "IN" the cloud

Customer responsibility varies based on specific AWS Services that the customer selects

For example, if you spin up an EC2 instance, you are responsible for the guest operating system, patching the guest OS, any application or utilities installed in your instance, configuration of security group firewall, network ACL firewall, and so forth

Whereas AWS is responsible for an underlying physical server, virtualization software, and infrastructure

If you use a managed service such as S3 or DynamoDB, you are responsible for selecting the appropriate region, managing the data, encryption options, classifying assets, and using IAM to apply appropriate access controls [i.e., who can read, write, update, delete]

AWS trains AWS employees. Similarly, a customer must train their own employees