**Chapter 1**

**Understanding the Foundations of AWS**

**Associate:** There are several Associate certifications.

**AWS Certified Solutions Architect -** Associate: For individuals working as solutions architects, designing AWS solutions using AWS services

**AWS Certified SysOps Administrator -** Associate: For individuals working as system administrators, managing and operating AWS services.

**AWS Certified Developer -** Associate: For individuals working as developers, deploying and debugging cloud-based applications hosted at AWS.

**The three Associate exams would test different aspects of CloudWatch logs:**

**Architect:**The main focus of this exam is on how CloudWatch logs work and the main design features to consider based on specific needs—that is, design knowledge related to using CloudWatch logs for a variety of solutions.

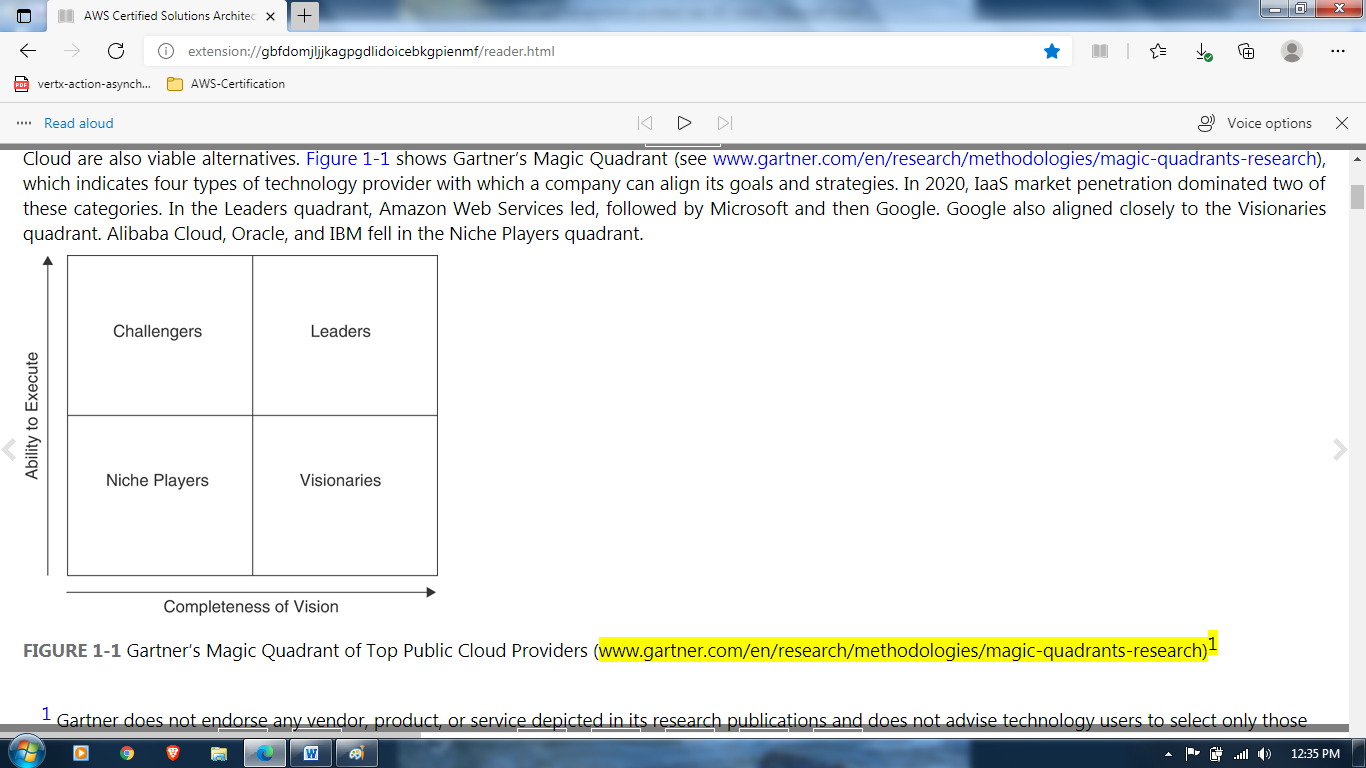
**SysOps:**The main focus of this exam is on how to configure CloudWatch logs based on specific needs—that is, configuration and deployment ofCloudWatch logs using operational knowledge.

**Developer:**The main focus of this exam is on what CloudWatch logs are useful for when developing applications for tracking performance of an application hosted on an EC2 instance—that is, knowledge of how a particular AWS service can help in the development and testing process with applications.  
  
**The Exam Objectives (Domains)**

The AWS Certified Solutions Architect - Associate (SAA-C02) exam is broken down into five major domains. This book covers each of the domains and the subtopics included in them.

The following table lists the breakdown of the domains represented on the exam:

|  |  |
| --- | --- |
| Domain | Percentage of Representation in Exam |
| Domain 1: Design Resilient Architectures | 30% |
| Domain 2: Design High-Performing Architectures | 28% |
| Domain 3: Design Secure Applications and Architectures | 24% |
| Domain 4: Design Cost-Optimized Architectures | 18% |
|  | Total 100% |

  
Today we have the cloud, which is just a collection of data centers. There is no ownership from the customer’s point of view; the cloud provider owns the services, and customers rent services as required.

AWS is also quite happy to allow you to continue to operate your on-premises data centers and allow them to coexist with cloud resources and services operating at AWS. Microsoft Azure will offer to sell you a copy of its complete Azure cloud operating system, called Azure Stack, to install on your servers in your data centers. As you can see, it’s hard to define the public cloud these days other than as a massive collection of compute and storage resources hosted on a network stored in the collection of data centers accessible across the Internet or by using private connections.  
Individual software vendor [ISV], Elastic Cloud Compute [EC2] instancesNational Institute of Standards and Technology (NIST), a branch of the U.S. government, you’re not alone. Around 2010, NIST began documenting the public cloud. After talking to all the major vendors, it released an initial report in June 2011, defining many cloud components that were common across all the public cloud vendors.  
  
According to NIST, five key definitions of the public cloud have morphed into a definitive standard methodology of operating in the public cloud. The following sections describe these definitions.

* **On-Demand Self-Service via AWS Management Console**.
* **Broad Network Access**: At AWS, applications can be publicly available, or they can remain completely private. Virtual private network (VPN) connections from your place of work to AWS are commonplace; in fact, you can order an AWS Direct Connect connection and establish a private fiber connection to AWS running at speeds up to 10 Gbps. Depending on the type of applications you’re using in the cloud, high-speed network access may be essential. It is even possible to administer AWS services from a phone by using AWS apps. Certainly, accessing AWS from any device is possible.
* **Resource Pooling :** AWS has clusters of data centers, known as availability zones (AZs), and each AZ could have more than 80,000 bare-metal servers available and online, allowing customers to host their application services with a high level of resiliency and failover. Having many available online resources also enables AWS to minimize prices. Without a massive pool of resources, AWS would not be able to offer its cloud services on demand and allow customers to scale up and down based on demand. For example, AWS S3 storage is unlimited; that is, it comes with no defined maximum limit.
* **Rapid Elasticity:** With AWS, customers get elasticity in terms of both compute and storage. Elasticity is useful only if it’s automated based on demand. You do not want to have to turn off a virtual server, add RAM, and turn the server back on again; you want horizontal scaling—that is, more application servers, not just a bigger server. Real-time monitoring of a hosted cloud application allows AWS to react almost instantaneously, before the application’s performance is close to degrading. Rapid elasticity based on demand is possible only with real-time monitoring driving automated scaling.
* **Measured Service**: In the cloud, you are billed for only what you use; this is referred to as measured service. Cloud providers make money by charging for everything that you use in their data centers, including data transfer costs. Packet flow inbounds, or ingress to the public cloud, is usually free; outbound packet flow, or traffic between subnets hosted in different data centers, is usually charged an outbound data transfer fee. Charges are per second or per minute, in the case of computer services such as AWS EC2 compute instances; charges may also be per gigabyte per month in the case of storage services such as S3 or virtual hard drives, such as Amazon Elastic Block Store (EBS). AWS charges can be broken down into compute, storage, and data transfer charges. If an AWS service is on, the meter is running. Cost management is one of your most important jobs when operating in the cloud. AWS has many useful tools to help you control your costs, including the AWS Simple Monthly Calculator, AWS Cost Explorer, and AWS Budgets.
* **Moving to AWS**: The reality is that when you move to the cloud, you must give up an element of control. After all, it’s not your data center. At AWS, you cannot get deeper into the infrastructure stack than the subnets that host your applications. With any public cloud provider, the only access allowed is to Layer 3 through Layer 7 in the ISO stack. So customers have control of the network resources at Layer 3, such as subnets and VPN connections, but do not have access to Layer 2 switching resources. Remember that the cloud is a data center, but it’s not *your* data center.

**Infrastructure as a Service (IaaS):**

Most of the services AWS offers fall into the infrastructure as a service (IaaS) category. With this model, which is certainly the most mature cloud model offering, virtualized servers and virtualized storage arrays are hosted on a software-defined network, and each customer’s infrastructure is completely isolated as a private resource. Creating resources at AWS typically starts with the creation of a virtual private cloud (VPC). Virtual servers, virtual hard drive volumes, and indeed complete managed services and products can be hosted on an isolated private network. You have the flexibility to create whatever architectural stack you desire at AWS, using the vast number of services and utilities available in the IaaS toolbox. Companies moving to the AWS public cloud typically start with IaaS because the compute and storage services closely mirror their current on-premises virtual environment. IaaS cloud services at AWS are bundled with managed services. A managed service is built using compute, storage, and networking services and customized software to provide something you want AWS to manage and maintain for you. Here are a few examples of managed services at AWS.

* **Database services:**AWS offers a managed service called **Amazon Relational Database Service (RDS).** With this service, AWS builds, hosts, maintains backups, handles failover, synchronizes, and monitors a pair of primary/standby database servers for you, leaving you the single task of managing your data records. Many other managed database services are also available at AWS, including Dynamo DB, a NoSQL database offering.
* **Automating infrastructure:**An automation service called **CloudFormation** enables you to automate the process of building your two- and three-tier infrastructure stacks, complete with the required compute, storage, networks, and load balancers required for your application. In fact, practically anything to do with building, updating, or deleting your infrastructure at AWS can be automated with CloudFormation. Template files are created using either JSON or YAML declarative code**.**
* **Auditing:**A handy service called CloudTrail is enabled in every AWS account and provided free of charge for 90 days. CloudTrail tracks and records all application programming interfaces (APIs) and authentication calls that are carried out in every AWS account. You can configure CloudTrail to store your API calls in S3 storage forever**.**
* At AWS, infrastructure resources are spread across the world in 24 different regions, and more regions are being added on an ongoing basis. If you are in a large population center, the odds are that Amazon is close by. If Amazon is not close by, you still may be able to connect into it through an edge location or a local point of presence.

**Platform as a Service (PaaS):**

Platform as a service (PaaS) cloud providers enable your company’s developers to create custom applications on a variety of popular development platforms, such as Java, PHP, and Python. The developers don’t have to manually build the infrastructure components required for each application per se; the required infrastructure resources are defined at the beginning of the development cycle and are created and managed by the PaaS cloud provider.

After an application has been developed and tested and is ready for prime time, the application is made available to end users using public URLs. The PaaS cloud provider hosts and scales the hosted application based on demand. As the number of users using the application changes, the infrastructure resources scale out or in as required. PaaS environments are installed on the IaaS resources of the PaaS cloud provider.

In fact, IaaS is always behind all “as a service”. **AWS can integrate with many PaaS solutions by offering several serverless solutions, such as AWS Lambda, the API Gateway, and code deployment tools such as** **CodeBuild, CodeCommit, and CodeDeploy**. It has also recently purchased Cloud9, a cloud-hosted integrated development environment (IDE) that supports more than 40 programming languages.

For the purpose of the AWS Certified Solutions Architect - Associate (SAA-C02) exam, you do not really need to be worried about these PaaS components with the exception of some of the application integration tools (see Chapter 4, “Decoupling Mechanisms Using AWS Services”).

However, you will be expected to know about certain PaaS services, such as RDS, the relational database service that is offered by AWS for a variety of popular database engines. Another example of a PaaS service that you will be expected to know about for the exam is AWS Lambda, a serverless service that allows you to create and host custom functions written in many popular programming languages.

Amazon has several development solutions, shown in Figure 1-7, including CodeBuild, CodeCommit, Elastic Beanstalk, and CodeDeploy, that can be key components in your application deployment workflow and deployment at AWS.

|  |  |
| --- | --- |
|  |  |

**Operational Benefits of AWS:**

**Servers**: Underutilized servers in your data center are expensive to run and maintain. Moving applications to the public cloud will reduce the size of your on-premises data center. When you no longer host as many physical servers, your total hosting costs (heating, cooling, and so on) are lower as well. You also don’t have to pay for as many software licenses at the processer level because you’re not responsible for running hypervisor services; that’s Amazon’s job. You may think that moving to the AWS cloud means virtualized resources and only virtualization.

**Storage:**For storage area network solutions, Amazon has shareable file solutions, including **Amazon Elastic File System (EFS) for Linux workloads and FSx, a shared file service specifically for Windows File Server workloads.** Virtual hard disks are available using EBS to create the required volumes. Unlimited storage and longer-term archive storage are provided by S3 and S3 Glacier/S3 Glacier Deep Archive.

**Managed Services:**

|  |  |  |
| --- | --- | --- |
| IT Operation | On-Premises | AWS Cloud |
| Monitoring | Nagios, SolarWinds | CloudWatch monitoring provides metrics for every AWS service. All monitoring and logging data can be stored in S3. All third-party monitoring solutions can access S3 to perform their own custom analysis of log data. |
| Data backup | Backup tools such as Commvault and NetBackup | Any third-party vendor that wants to stay in business supports AWS; both Veritas and Commvault have AWS solutions. AWS Storage Gateway can also be installed to cache required content locally, and local disk volumes can be backed up to an S3 bucket. Backups of EBS volumes are called snapshots. There is also a management tool to help you automate the backup and retention of your EBS snapshots, as well as a backup tool called AWS Backup that allows you to centrally control backup of all your data records at AWS, using one management service. |
| Scale | Automation for increasing/decreasing the size of each virtual machine’s RAM and CPU cores | It is possible to scale horizontally by placing multiple virtual machines (instances) behind a load balancer and adding automated scaling based on demand to increase and decrease the required amount of compute power required by your application using EC2 Auto Scaling. |
| Testing | Provisioning hardware for testing is expensive | Provisioning resources for short-term testing at AWS is incredibly inexpensive. Signing up for the AWS free tier allows you to test a variety of AWS services for one year completely free of charge. |
| Identity Management | Active Directory Domain Services for accessing corporate resources | It is possible to extend on-premises Active Directory to the AWS cloud with hosted directory services. You can use AWS single sign-on (SSO) services to manage access to popular business applications that third-party cloud providers are hosting. |

**Security at AWS:**

You can enable **multifactor authentication (MFA) as an additional security control on S3 buckets to control when deletion of data records is performed.**  
**Network Security at AWS:**

Each subnet’s ingress and egress traffic can be controlled by subnet firewalls called *network ACLs* that define separate stateless rules for inbound and outbound packet flow. Each EC2 instance hosted on a subnet is further protected by an additional firewall called a security group, which defines what traffic is allowed into the instance and where outbound traffic is directed. VPC flow logs can be enabled to capture network traffic for the entire VPC, for a single subnet, or for a network interface.

**Application Security at AWS:**

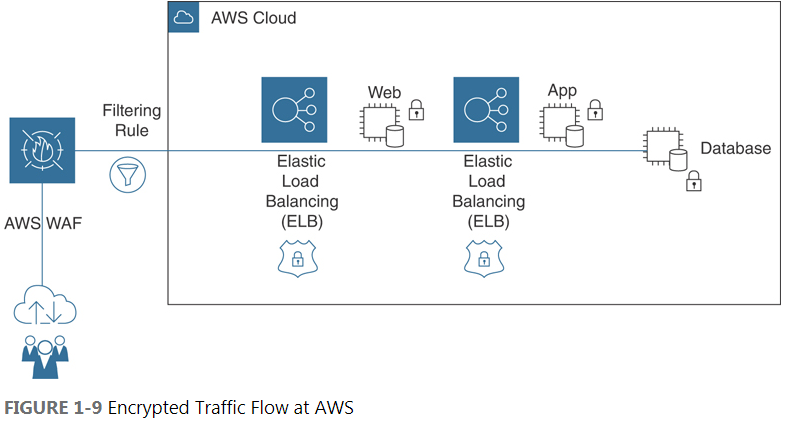
One load balancer type offered by AWS is Application Load Balancer, which can perform authentication and SSL offloading services. The end-to-end traffic pattern for a three-tier web application can be designed using many encryption/decryption points on its path from source to destination, as described in the list that follows and as shown in Figure 1-9:

**Web Application Firewall (WAF)**: WAF is a custom traffic filter in front of Application Load Balancer that protects against malicious traffic.

**Elastic Load Balancer (ELB)**: ELB accepts only encrypted HTTPS traffic on port 443 and provides Secure Sockets Layer/Transport Layer Security (SSL/TLS) decryption and, optionally, user authentication.

**EC2 instance hosting a web application**: EBS (**Elastic Block Store**) boot and data drives can be encrypted.

**EC2 instance hosting an application server**: EBS boot and data drives can be encrypted.

**Database server**: EBS boot and data drives can be encrypted.  


AWS has several free hands-on labs. You can sign up for **QwikLabs at**[**https://qwiklabs.com/**](https://qwiklabs.com/) to carry out a variety of AWS tasks in the AWS cloud.

**The AWS Well-Architected Framework:**

The documentation for the Well-Architected Framework (see [**https://d1.awsstatic.com/whitepapers/architecture/AWS\_Well-Architected\_Framework.pdf**](https://d1.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf)) also presents many key questions you should ponder. It is useful to discuss these questions with other technical folks in your company in order to make key decisions about your infrastructure and applications hosted at AWS. The main document for the framework can be found here: [**https://d1.awsstatic.com/whitepapers/architecture/AWS\_Well-Architected\_Framework.pdf**](https://d1.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf). The AWS Certified Solutions Architect - Associate (SAA-C02) exam is designed around the Well-Architected Framework. It is well worth downloading and reading five PDFs related to the framework as your study for the exam from [**https://aws.amazon.com/architecture/well-architected/?wa-lens-whitepapers.sort-by=item.additionalFields.sortDate&wa-lens-whitepapers.sort-order=desc**](https://aws.amazon.com/architecture/well-architected/?wa-lens-whitepapers.sort-by=item.additionalFields.sortDate&wa-lens-whitepapers.sort-order=desc). You can view each application to be deployed at AWS through the lens of the Well-Architected Framework by following these five principles:

**Operational excellence:**

This principle relates to how best to execute, deploy, and monitor applications running at AWS using automated deployment monitoring procedures, continuous improvement, and automated solutions for recovering from failures. Key AWS operational excellence–related services include CloudWatch events and alarms, CloudTrail, EC2 Auto Scaling, AWS Config, and Trusted Advisor. Operational excellence questions to consider include these:

* + How are disruptions to applications handled—manually or automatically?
  + How can you analyze the ongoing health of your applications and infrastructure components hosted at AWS?

**Security:**

This principle relates to how to best design systems that will operate reliably and securely while protecting customer information and data records. Key AWS security-related services include IAM, AWS Organizations, CloudWatch logs, CloudTrail events, S3 and S3 Glacier, and VPC flow logs. Security questions to consider include these:

* + How are security credentials and authentication managed at AWS?
  + How are automated procedures secured?

**Reliability:**

This principle relates to how systems and applications hosted at AWS recover from disruption with minimal downtime and how applications meet escalating demands. Key AWS reliability-related services include ELB, EC2 Auto Scaling, and CloudWatch alarms. Reliability questions to consider include these:

* + How do you monitor resources hosted at AWS?
  + How do applications hosted at AWS adapt to changes in demand by end users?

**Performance efficiency:**

This principle relates to how to use compute resources to meet and maintain your application requirements on an ongoing basis. If you need to change a solution from EC2 instance to containers or serverless then Key performance-related services include EC2 Auto Scaling, EBS volumes, and RDS. Performance efficiency questions to consider include these:

* + Why did you select your database?
  + Why did you select your current compute infrastructure?

**Cost optimization:**

This principle relates to how to design systems that meet your needs at the lowest price point. Key AWS cost optimization–related services include Cost Explorer, Budgets, EC2 Auto Scaling, Trusted Advisor, and Simple Monthly Calculator. Cost optimization questions to consider are as follows:

* + How do you oversee usage and cost?
  + How do you meet cost targets?
  + Are you aware of current data transfer charges based on your AWS designs?

**Chapter 2**

**Designing Multi-Tier Architecture Solutions**   
 **Domain 1: Design Resilient Architectures**

**High availability (HA) refers to making hosted applications always available, regardless of the situation or circumstances that happen from time to time in the cloud. After all, everything fails sometime.** How does an application respond to failure? If one web server fails there is a backup web server or two to take its place. **Resiliency is mostly used to describe data storage in the cloud. Any data storage service available in the cloud maintains multiple copies of all records and files stored. Resiliency is also present because of the high availability. If you have multiple web servers available, you have high availability and also a higher level of resiliency. High-availability and resiliency are keys when hosting application stacks in the cloud**.

**Availability and Reliability**

When you host an application at AWS, of course you expect the application to be available at all times. However, there are a number of moving parts to consider with each application stack that you design. **A typical hosted application uses a number of managed services provided by AWS for your compute, networking, and storage services. A managed service is an AWS service that is built and maintained by AWS; every single offering shown in the AWS Management Console is a managed service that can be ordered and used as a component of an application stack. Also, at a minimum, you will be using a monitoring service to make sure that the services that you have ordered are working—and determine when they are not working. The monitoring service that is embedded into each service offered by AWS is called CloudWatch. Your application(s) will be hosted on either virtual machines (that is, EC2 instances) or using one of the elastic container service, such as Amazon Elastic Container Service (ECS). You will also be using virtual hard drives for storage, which means you will be using Amazon Elastic Block Store (EBS). Each of these management services has built-in reliability, availability, and scalability by design.**

Amazon expects you to design a hosted application to be secure, reliable, and available. The Amazon’s key to design with security in mind and then to make the system as reliable and dependable as possible and finally to make the application as responsive and as fast as it can be without sacrificing security or reliability.   
  
**Availability in the Cloud:**

**Availability is the percentage of time that an application remains operational—that is, the percentage of time that it can be accessed in the cloud from any location on any device. AWS publishes a compute service-level agreement that covers Amazon Elastic Compute Cloud (EC2), Elastic Block Store (EBS), and Elastic Container Service (ECS) with a stated availability of 99.99%, which means that the total downtime expected per year utilizing any of these services is a total of 4.38 hours. The downtime per week is calculated as a scant 5 minutes**.

Most of the time you are hosting application on multiple EC2 instances hosted on separate subnets in separate data centers. In addition, each subnet is hosted in a separate data center, which is also located in a separate availability zone within the AWS region.

If you look at your virtual hard drive that stores your boot drives and data volumes, you can see that your EBS storage volumes are created and maintained as replicated storage volumes. Multiple copies of each virtual hard are created and stored within the data center where the EBS volumes were first created and attached. Fault tolerance using multiple physical locations allows you to define your application availability as highly available. And if you have high availability, you have also designed a level of reliability for your application as well.  
**Reliability**

Reliability is the amount of time during which an application operates at a certain level of performance. Under Amazon’s compute SLA offering, you may still expect that there will just be 4.38 hours of downtime per year. However, customers at Amazon need to design their own SLAs and their own expectations for their hosted applications. How much application time loss is acceptable to you and your company? How many failures can occur before the entire application cannot be trusted to operate at an acceptable level? When designing for reliability, you need to consider for a hosted application an acceptable trade-off between the desired service level based on your requirements and the true cost of the application.

**AWS Regions and Availability Zones**

Amazon cloud services are designed to be available where there are lots of people; therefore, Amazon services are hosted in regions in populous areas of the world. **An AWS region, in simple terms, comprises several data centers and a plethora of managed services located in a specific geographic area of the world.** Each AWS region is completely isolated from the other regions by at least several hundred kilometers so that problems within one region remain localized to that one region and do not affect any other AWS region you also may be operating in. Each administrator working within a separate AWS account can initially choose to work in any AWS region except for the GovCloud regions and several regions in China that require special AWS accounts. For information on current regions and availability zones, see <https://aws.amazon.com/about-aws/global-infrastructure/regions_az/>

**Isolating regions allows AWS to guarantee a level of operational compliance and data sovereignty**. All AWS cloud services have been designed to be durable and redundant and to retain the ability to fail over to other backup services within the same AWS region when disaster strikes.

**The AWS networking campus spread across each AWS region is designed for fault tolerance and failover**. If you wish to host application on multiple AWS regions, then you can use the service **Route53 for geographically load balancing your application stacks that are hosted on different regions**.

For a database solution with high reliability and availability, you might decide to use a database service such as DynamoDB (AWS’s managed NoSQL database solution) and a global table, which is a DynamoDB table that is hosted and replicated/synchronized across multiple regions. Or you could choose to deploy Amazon’s MySQL Relational Database Service (RDS), which typically is deployed using a primary and standby database design with multiple database instances. Most RDS solutions for production have this built-in level of redundancy, although not at the level of a horizontally scaled DynamoDB table across multiple AWS regions. However, there is an RDS solution called Amazon Aurora that is MySQL and PostgreSQL compatible and stores data records in clustered shared storage that can be stored across multiple regions.

**Availability Zones**

An *availability zone (AZ)* is typically defined as one or more data centers geographically located in an AWS region. Each AZ is linked to the other AZs in the region through private dedicated, redundant; low-latency fiber network connections that Amazon owns (see Figure 2-4). Internal AWS private network speeds are in the 40 Gbps range and up. To offer cloud services that are designed to adhere to the many compliance programs that AWS supports, including HIPAA and PDI DSS, Amazon needs to own the equipment so that it can control the operation and setup of the equipment.

|  |  |
| --- | --- |
|  |  |

**An availability zone is basically a collection of subnets. In fact, you’re not going to get any deeper into the weeds networking-wise than at the subnet level**.

The failure of a data center within an AZ does not affect and derail the operation of the AWS managed services that are specifically designed to operate within each region as they are located outside the data centers in each AZ. For example, Amazon’s S3 storage and Elastic Load Balancing (ELB) services are specifically designed to integrate with the data centers within each AZ; however, each of these services functions as a standalone service that continues to function even if one or all the data centers located within an AZ fail. Even though each AWS region is backed by a cluster of multiple data centers, it is important to reiterate that no two AZs share the same single data center. Each data center in each AZ is a separate entity that is isolated from other data centers within the same and different availability zones.

**Availability Zone Distribution**

AWS carries out balancing and distribution of resources hosted in AZs in the background.

|  |
| --- |
|  |

**Average latency between AZs within a region is around 3 ms**. certainly, if you’re concerned about network speeds, you can perform latency testing from one instance hosted in one AZ to an instance in another AZ by using ping because local routing is enabled within a virtual private cloud (VPC). However, you could go a step further and use **AWS Systems Manager to create a monitoring instance that would perform detailed network analysis using ping**, **TCP TRACEROUTE**, **MTR**, and **TRACEPATH**.

**Each AZ within a region has a name based on the region it is hosted within as well as a letter code, such as US-West-1A**.

**Within a single AZ, the data center private network connections are defined as intra-AZ connections, as they are local to the AZ**, as shown in Figure 2-6. **The wiring between the AZs is defined as inter-AZ connections, with private links connecting the regions**. The primary reasons for using AZs in your infrastructure design are for application failover and primary/standby database replication.

|  |
| --- |
|  |

**Multiple Availability Zones**