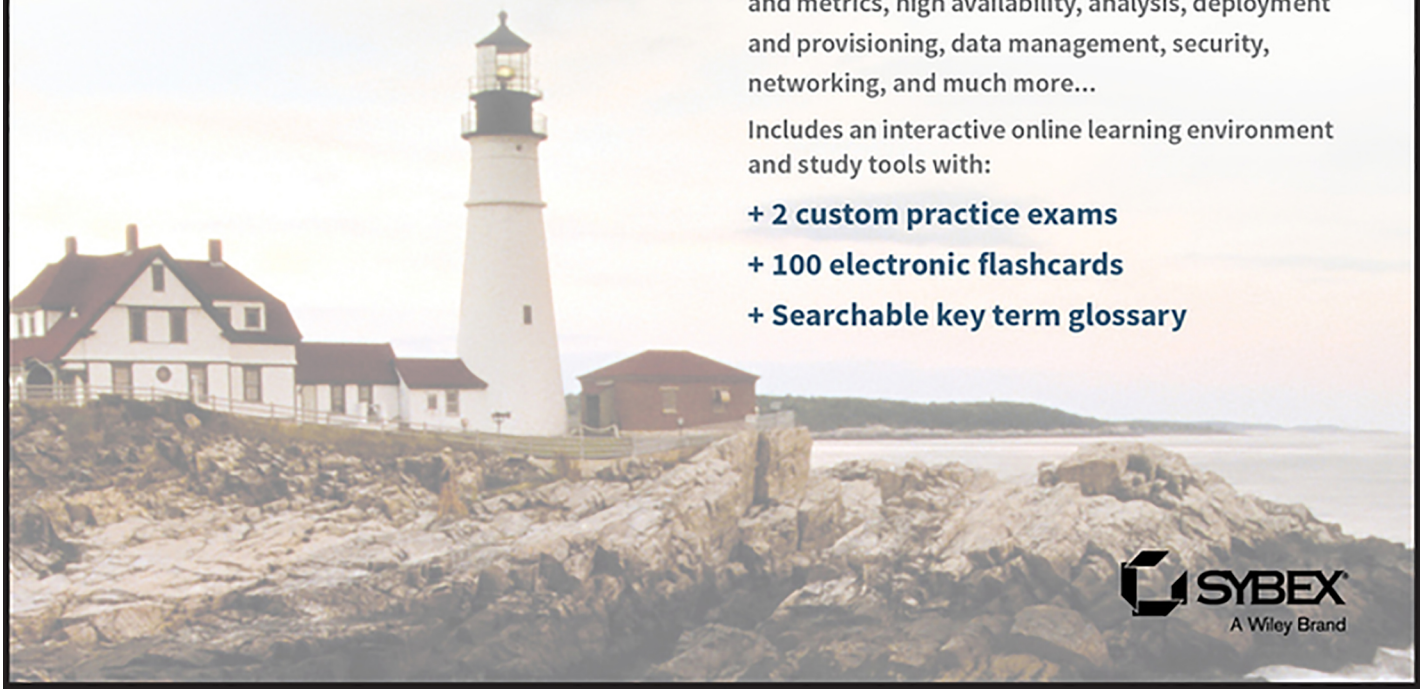
amazon web services

Stephen Cole, Gareth Digby, Chris Fitch,
Steve Friedberg, Shaun Qualheim, Jerry Rhoads,
Michael Roth, Blaine Sundrud

AWS Certified SysOps Administrator

OFFICIAL STUDY GUIDE

ASSOCIATE EXAM



Covers exam objectives, including monitoring and metrics, high availability, analysis, deployment and provisioning, data management, security, networking, and much more...

Includes an interactive online learning environment and study tools with:

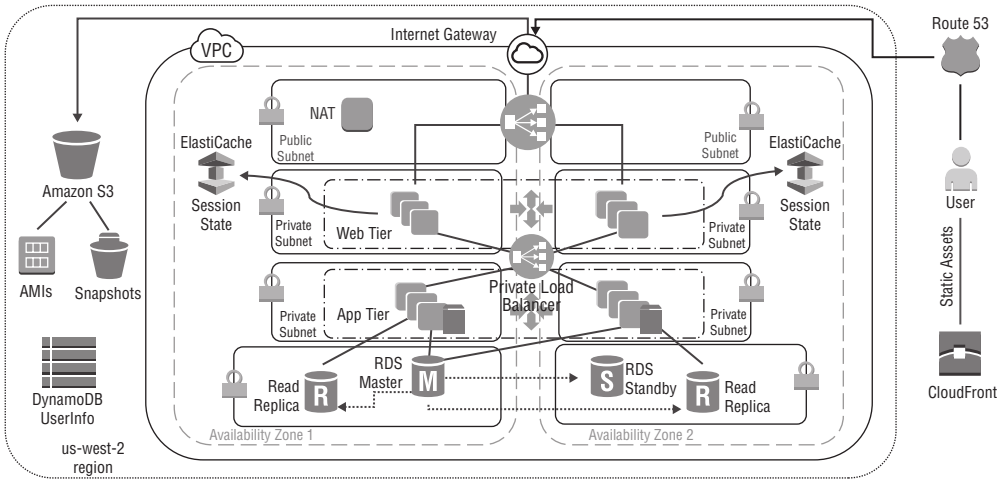
+ 2 custom practice exams

+ 100 electronic flashcards

+ Searchable key term glossary

SYBEX
A Wiley Brand

The Solution



As we examine the pieces of the solution, we start by breaking down the components of the architecture. Then we focus on how systems operators interact with the individual pieces and begin thinking about how those pieces fit into the certification exam.

Environment

Architectures live inside *AWS Regions*; in this scenario, in us-west-2 (Oregon, United States). Regions are made up of multiple *Availability Zones*, which provide the foundation for highly available architectures. Although this is a systems operation exam, it is critical to understand the nature of AWS Regions and Availability Zones.

Each AWS Region is a separate geographic area. Each AWS Region has multiple, isolated locations known as *Availability Zones*. *AWS Regions* and *Availability Zones* are discussed in Chapter 5, "Networking."

Networking

Networking components start inside the AWS Region with Amazon Virtual Private Cloud (Amazon VPC). *Amazon VPC* is a private network in the AWS Region that isolates all traffic from the millions of other applications running in AWS. A deep dive into Amazon VPC (and the rest of its components) is found in Chapter 5.

Amazon VPC is divided into *subnets*; all assets running in your Amazon VPC are assigned to a subnet. Unlike on-premises subnetting decisions that can affect latency between servers, Amazon VPC subnets only affect access. Access between subnets is

controlled through *network Access Control Lists (nACLs)*, and access in and out of Amazon VPC is controlled through attached gateways. In this scenario, the only gateway is the *Internet Gateway (IGW)*, and it allows traffic to and from external (public IP) sources.

By granting route table access to the gateway only to specific subnets, ingress and egress can be tightly controlled. In this scenario, public subnets indicate IGW access. Without IGW access, the subnets become private; that is, they are accessible only to private IP networks.

To learn about the other gateways that could be leveraged to create hybrid or other private architectures, refer to Chapter 5.

Security groups are often part of the networking discussion. They provide stateful firewalls that operate at the Hypervisor levels for all individual*Amazon Elastic Compute Cloud (Amazon EC2)* instances and other Amazon VPC objects. In this scenario, we potentially have seven different security groups:

Public Elastic Load Balancing The only security group that allows full public access

Web Tier Amazon EC2 This accepts traffic only from public Elastic Load Balancing.

Private Elastic Load Balancing This accepts traffic only from Web Tier Amazon EC2.

Application Tier Amazon EC2 This accepts traffic only from private Elastic Load Balancing.

Amazon ElastiCache This accepts traffic only from Application Tier Amazon EC2.

Amazon Relational Database Service (Amazon RDS) This accepts traffic only from Application Tier Amazon EC2.

Network Address Translation (NAT) This is used only for internally initiated outbound traffic.

By specifically stacking security groups in this manner, you can provide layers of network security that surround the database portion of the three-tier design.

Compute

In this scenario, you use traditional compute methods, such as Linux servers running on Amazon EC2. Amazon EC2 comes in many sizes (how many CPUs, how much memory, how much network capacity, and so on), known as *instances*. Based on the Amazon Machine Image (AMI), each Amazon EC2 instance can run a wide range of Linux- or Windows-based operating systems as well as preinstalled software packages. Amazon EC2 instances also support runtime configuration as required.

The requirements for the scenario include scalable solutions. AWS provides Auto Scaling as an engine that can take predefined launch configurations and dynamically add or remove instances from the web or the Application Tier based on metrics.

Details on Amazon EC2, Auto Scaling, and other compute resources are found in Chapter 4, "Compute."

Database

Amazon RDS runs in your Amazon VPC on Amazon EC2. You select the database engine and version (MySQL, Oracle, Postgres, and so forth) and the configuration (the size of the Amazon EC2 instance, which subnets to use, how often to take backups, and so on). Amazon RDS takes care of the infrastructure of the instances and the engine; your database administrator (DBA) takes care of the database schema and data.

This scenario also includes *Amazon DynamoDB*, a native NoSQL engine optimized for consistent low latency, high availability, and strongly consistent reads and writes. Unlike Amazon RDS (or do-it-yourself databases running on Amazon EC2), Amazon DynamoDB operates at the regional level through API access only.

For details on how Amazon DynamoDB and other databases function, refer to Chapter 7, "Databases."

Storage

This scenario looks at storage in three different areas: the block storage used by the Amazon EC2 instances, the object storage keeping all of the media as well as backups and AMIs, and the caching storage used by Amazon CloudFront.

Amazon EBS is durable, persistent block storage used by most Amazon EC2 and Amazon RDS instances. It provides drive space for boot volumes and data volumes. Additionally, AWS provides ephemeral storage for many Amazon EC2 instance types through instance storage. Deciding which one to use becomes an operational value judgment, one that compares speed, persistence, and cost.

Object storage is provided by Amazon S3. *Amazon S3*, like Amazon DynamoDB, operates at the regional level outside Amazon VPC. It is only accessed through API commands that your operations team controls with fine-grained precision. Highly cost-effective and massively durable, Amazon S3 provides web-enabled storage for content as well as protected storage for database backups and AMI storage.

Amazon CloudFront is the *AWS content delivery network service (CDN).* This application leverages Amazon CloudFront to cache content close to consumers in order to improve performance (reduce latency) and reduce costs.

Storage systems, including shared file systems, the Amazon Elastic File System (Amazon EFS), and cold storage via Amazon Glacier, are discussed in Chapter 6, "Storage."

User Management

Although not drawn in the sample three-tier architecture diagram, user management becomes one of the critical elements of the AWS operational design. Operator access is controlled through *AWS Identity and Access Management (IAM)*. IAM maintains control over validating authentication methods (passwords, access keys, and so on) and then grants access to authenticated operators.

Because everything in AWS is accessed through APIs, IAM becomes a comprehensive tool for controlling all permissions to AWS services and resources.

For established enterprise customers, IAM can be integrated with existing directory systems via AWS Directory Service.

AWS IAM controls access to AWS services and resources. It does not control access to the Amazon EC2 operating system or application-level authentication. For more details, refer to the shared responsibility model in Chapter 3, "Security and AWS Identity and Access Management (IAM)."

Security, Monitoring, and Deployment

Security is integral to every part of the AWS platform. This means that security is part of each piece of the architecture.

There are some specific AWS security tools, such as Amazon Inspector, Amazon VPC Flow Logs, Amazon CloudWatch Logs, and others which provide a more focused toolset that the AWS operations team can leverage to ensure the security profile of the AWS application. These and many other tools are discussed in Chapter 3.

Monitoring of critical systems is provided by *Amazon CloudWatch*, which provides visibility into metrics that happen on the Customer side of the shared responsibility model. Thousands of metrics across more than 90 services keep track of everything from CPU consumption to latency, queue depths, and so on.

AWS CloudTrail records every API call in the AWS system, including:

- Who made the API call
- When the API call was performed
- Where the API call originated
- The result of the API call

These records and other log files are processed through Amazon CloudWatch Logs, which analyze text data for patterns that trigger alerts and corresponding actions.

Automated deployment methods ensure that human error does not disrupt rollouts or updates to production or sandbox environments. *AWS CloudFormation* turns infrastructure plans into code, allowing your operations team to build and tear down entire systems in a single action. Refer to Chapter 8, "Application Deployment and Management," for more details.

Key Products: Three-Tier Design

As described above, the three-tier architecture consists of a web front end, an application layer, and database layer. In addition to the compute, storage, and database resources, additional AWS infrastructure may need to be deployed. Refer to Table 1.1 for a list of key products.

TABLE 1.1 Key Products: Three-Tier Architecture

Tools to Enable Hybrid Cloud Architectures	Description
AWS Regions and Availability Zones	Amazon EC2 is hosted in multiple locations world-wide. These locations are composed of regions and Availability Zones. Each region is a separate geographic area. Amazon EC2 provides you with the ability to place resources, such as instances and data, in multiple locations.
Availability Zones	Within a region are two or more Avalability Zones. The Availability Zone is isolated, but the Availability Zones in a region are connected through low-latency links.
Edge Locations	Edge Locations = AWS Lambda@Edge Amazon CloudFront, Amazon Route 53, AWS Shield, and AWS WAF services that are offered at AWS Edge Locations.
Hybrid cloud architecture	Integration of on-premises resources with cloud resources
Amazon Route 53	Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service.
Amazon CloudFront	Amazon CloudFront is a web service that speeds up distribution of your static and dynamic web content to your users. CloudFront delivers your content through a worldwide network of data centers called *edge locations*.
Amazon VPC	A virtual private cloud (VPC) is a virtual network dedicated to your AWS account. It is logically isolated from other virtual networks in the AWS cloud. You can launch your AWS resources, such as Amazon EC2 instances, into your VPC.
Internet Gateways	An Internet gateway is a horizontally-scaled, redundant, and highly available VPC component that allows communication between instances in your VPC and the Internet.
Subnets	A subnetwork or subnet is a logical subdivision of an IP network.
Route tables	A route table is a set of rules that is used to determine where data packets traveling over an Internet Protocol (IP) network will be directed.

5. Use groups to assign permissions to IAM users.
6. Enable MFA for privileged users.
7. Use IAM roles for applications that run on Amazon EC2 instances.
8. Delegate by using IAM roles instead of sharing credentials.
9. Rotate credentials regularly.
10. Remove unnecessary credentials.
11. Use policy conditions for extra security.
12. Monitor activity on your AWS account.
13. Remove root credentials.
14. Use access levels to review IAM permissions.
15. Use AWS-defined policies to assign permissions whenever possible.
16. Use IAM roles to provide cross-account access.

So far, you have learned about the shared responsibility model and how to secure your IAM account. Now let's talk about securing your AWS resources.

Securing Your AWS Cloud Services

In this section, we discuss Amazon EC2 key pairs, X.509 certificates, AWS Key Management Services (AWS KMS), and AWS CloudHSM.

Key Pairs

Amazon EC2 instances created from a public AMI use a public/private key pair instead of a password for signing in via SSH. The public key is embedded in your instance, and you use the private key to sign in securely without a password. After you create your own AMIs, you can choose other mechanisms to log in securely to your new instances. You can have a key pair generated automatically for you when you launch the instance, or you can upload your own. Save the private key in a safe place on your system, and record the location where you saved it. For Amazon CloudFront, you use key pairs to create signed URLs for private content, such as when you want to distribute restricted content that someone paid for.

Amazon CloudFront key pairs can be created only by the root account and cannot be created by IAM users.

X.509 Certificates

X.509 certificates are used to sign SOAP-based requests. X.509 certificates contain a public key and additional metadata (for example, an expiration date that AWS verifies when

Amazon Inspector

Amazon Inspector is an automated security assessment service that helps improve the security and compliance of applications deployed on AWS. Amazon Inspector automatically assesses applications for vulnerabilities or deviations from best practices. After performing an assessment, Amazon Inspector produces a detailed list of security findings prioritized by level of severity.

Amazon Inspector includes a knowledge base of hundreds of rules mapped to common security best practices and vulnerability definitions. Examples of built-in rules include checking for remote root login being enabled or vulnerable software versions being installed. These rules are regularly updated by AWS security researchers. More information about AWS Inspector can be found in Chapters 4 and 9.

AWS Certificate Manager

AWS Certificate Manager is a service that lets you provision, manage, and deploy SSL/TLS certificates for use with AWS Cloud services. SSL/TLS certificates are used to secure network communications and establish the identity of websites over the Internet. With AWS Certificate Manager, you can request a certificate and deploy it on AWS resources, such as Elastic Load Balancing load balancers or Amazon CloudFront distributions. AWS Certificate Manager then handles the certificate renewals.

AWS Web Application Firewall (AWS WAF)

AWS Web Application Firewall (AWS WAF) is a web application firewall that helps protect web applications from attacks by allowing you to configure rules that allow, block, or monitor (count) web requests based on conditions that you define. These conditions include IP addresses, HTTP headers, HTTP body, Uniform Resource Identifier (URI) strings, SQL injection, and cross-site scripting.

As the underlying service receives requests for your websites, it forwards those requests to AWS WAF for inspection against your rules. Once a request meets a condition defined in your rules, AWS WAF instructs the underlying service either to block or allow the request based on the action you define.

AWS WAF is tightly integrated with Amazon CloudFront and the Application Load Balancer, services that AWS customers commonly use to deliver content for their websites and applications. When you use AWS WAF on Amazon CloudFront, your rules run in all AWS edge locations around the world close to your end users. This means that security doesn't come at the expense of performance. Blocked requests are stopped before they reach your web servers. When you use AWS WAF on Application Load Balancer, your rules run in-region and can be used to protect Internet-facing and internal load balancers.

AWS Trusted Advisor

The *AWS Trusted Advisor* customer support service not only monitors cloud performance and resiliency, but also cloud security. AWS Trusted Advisor inspects your AWS environment and makes recommendations when opportunities may exist to save money, improve

This access can only be modified through the invocation of Amazon VPC APIs. AWS supports the ability to grant granular access to different administrative functions on the instances and the Internet gateway, enabling you to implement additional security through separation of duties.

Dedicated instances Within an Amazon VPC, you can launch Amazon EC2 instances that are physically isolated at the host hardware level (that is, they will run on single-tenant hardware). An Amazon VPC can be created with "dedicated" tenancy so that all instances launched into the Amazon VPC will use this feature. Alternatively, an Amazon VPC may be created with "default" tenancy, but you can specify dedicated tenancy for particular instances launched into it. More information on networking on AWS can be found in Chapter 5.

Dedicated hosts An *Amazon EC2 Dedicated Host* is a physical server with Amazon EC2 instance capacity fully dedicated to your use. Dedicated hosts allow you to use your existing per-socket, per-core, or per-virtual machine software licenses, including Windows Server, Microsoft SQL Server, SUSE, Linux Enterprise Server, and so on. More information on dedicated hosts can be found in Chapter 4.

Amazon CloudFront Security

Amazon CloudFront gives customers an easy way to distribute content to end users with low latency and high data transfer speeds. It delivers dynamic, static, and streaming content using a global network of edge locations. Requests for customers' objects are automatically routed to the nearest edge location, so content is delivered with the best possible performance. Amazon CloudFront is optimized to work with other AWS Cloud services like Amazon S3, Amazon EC2, Elastic Load Balancing, and Amazon Route 53. It also works seamlessly with any non-AWS origin server that stores the original, definitive versions of your files.

Amazon CloudFront requires that every request made to its control API is authenticated so that only authorized users can create, modify, or delete their own Amazon CloudFront distributions. Requests are signed with an HMAC-SHA-1 signature calculated from the request and the user's private key. Additionally, the Amazon CloudFront control API is only accessible via SSL-enabled endpoints.

There is no guarantee of durability of data held in Amazon CloudFront edge locations. The service may sometimes remove objects from edge locations if those objects are not requested frequently. Durability is provided by Amazon S3, which works as the origin server for Amazon CloudFront by holding the original, definitive copies of objects delivered by Amazon CloudFront.

If you want control over who can download content from Amazon CloudFront, you can enable the service's private content feature. This feature has two components. The first controls how content is delivered from the Amazon CloudFront edge location to viewers on the Internet. The second controls how the Amazon CloudFront edge locations access objects in Amazon S3. Amazon CloudFront also supports geo-restriction, which restricts access to your content based on the geographic location of your viewers.

To control access to the original copies of your objects in Amazon S3, Amazon CloudFront allows you to create one or more origin access identities and associate these with your distributions. When an origin access identity is associated with an Amazon CloudFront distribution, the distribution will use that identity to retrieve objects from Amazon S3. You can then use Amazon S3's ACL feature, which limits access to that origin access identity so that the original copy of the object is not publicly readable.

To control who can download objects from Amazon CloudFront edge locations, the service uses a signed-URL verification system. To use this system, you first create a public-private key pair and upload the public key to your account via the AWS Management Console. You then configure your Amazon CloudFront distribution to indicate which accounts you would authorize to sign requests. You can indicate up to five AWS accounts that you trust to sign requests. As you receive requests, you will create policy documents indicating the conditions under which you want Amazon CloudFront to serve your content. These policy documents can specify the name of the object that is requested, the date and time of the request, and the source IP (or CIDR range) of the client making the request. You then calculate the SHA-1 hash of your policy document and sign this using your private key. Finally, you include both the encoded policy document and the signature as query string parameters when you reference your objects. When Amazon CloudFront receives a request, it will decode the signature using your public key. Amazon CloudFront will only serve requests that have a valid policy document and matching signature.

Note that private content is an optional feature that must be enabled when you set up your Amazon CloudFront distribution. Content delivered without this feature enabled will be publicly readable.

Amazon CloudFront provides the option to transfer content over an encrypted connection (HTTPS). By default, Amazon CloudFront will accept requests over both HTTP and HTTPS protocols. You can also configure Amazon CloudFront to require HTTPS for all requests or have Amazon CloudFront redirect HTTP requests to HTTPS. You can even configure Amazon CloudFront distributions to allow HTTP for some objects but require HTTPS for other objects. More information on Amazon CloudFront can be found in Chapters 5 and 6.

Storage

AWS provides low-cost data storage with high durability and availability. AWS offers storage choices for backup, archiving, disaster recovery, and block and object storage.

Amazon Simple Storage Service (Amazon S3) Security

Amazon S3 allows you to upload and retrieve data at any time from anywhere on the web. Amazon S3 stores data as objects in buckets. An object can be any kind of file: a text file, a photo, a video, and more. When you add a file to Amazon S3, you have the option of including metadata with the file and setting permissions to control access to the file. For each bucket, you can control access to the bucket (who can create, delete, and list objects in the bucket), view access logs for the bucket and its objects, and choose the geographical region where Amazon S3 will store the bucket and its contents.

resource. If both record sets are unhealthy, Amazon Route 53 returns the primary resource record set. Health checks are discussed in greater detail in Chapter 10.

You can combine routing (for example, have geolocation routing backed up with failover routing) to make sure that you provide the highest level of availability possible. As you can imagine, this can get very complex (and confusing) very quickly, so good documentation is important.

DNS Record Types

Explaining the various DNS records types is out of the scope of this book. However, Table 5.7 shows the supported record types for Amazon Route 53.

TABLE 5.7 Amazon Route 53 Supported DNS Record Types

Record Type	Description
A	Address mapping records
AAAA	IPv6 address records
CNAME	Canonical name records
MX	Mail exchanger record
NAPTR	Name authority pointer record
NS	Name server records
PTR	Reverse-lookup Pointer records
SOA	Start of authority records
SPF	Sender policy framework record
SRV	Service record
TXT	Text records

In addition to the standard DNS record types supported, Amazon Route 53 supports a record type called *Alias*. An *Alias record type*, instead of pointing to an IP address or a domain name, points to one of the following:

- An Amazon CloudFront distribution
- An AWS Elastic Beanstalk environment

- An Elastic Load Balancing Classic or Application Load Balancer
- An Amazon S3 bucket that is configured as a static website
- Another Amazon Route 53 resource record set in the same hosted zone

Health Checks

There are three types of health checks that you can configure with Amazon Route 53. They are as follows:

- The health of a specified resource, such as a web server
- The status of an Amazon CloudWatch alarm
- The status of other health checks

In this section, we explore each type. The level of detail covered may not be tested on the exam. However, as an AWS Certified Systems Operator, the material covered here is a must-know.

The health of a specified resource, such as a web server You can configure a health check that monitors an endpoint that you specify either by IP address or by domain name. At regular intervals that you specify, Amazon Route 53 submits automated requests over the Internet to your application, server, or other resource to verify that it's reachable, available, and functional. Optionally, you can configure the health check to make requests similar to those that your users make, such as requesting a web page from a specific URL.

The status of an Amazon CloudWatch alarm You can create CloudWatch alarms that monitor the status of CloudWatch metrics, such as the number of throttled read events for an Amazon DynamoDB database or the number of Elastic Load Balancing hosts that are considered healthy. After you create an alarm, you can create a health check that monitors the same data stream that CloudWatch monitors for the alarm.

To improve resiliency and availability, Amazon Route 53 doesn't wait for the CloudWatch alarm to go into the ALARM state. The status of a health check changes from healthy to unhealthy based on the data stream and on the criteria in the CloudWatch alarm. The status of a health check can change from healthy to unhealthy even before the state of the corresponding alarm has changed to ALARM in CloudWatch.

The status of other health checks You can create a health check that monitors whether Amazon Route 53 considers other health checks healthy or unhealthy. One situation where this might be useful is when you have multiple resources that perform the same function, such as multiple web servers, and your chief concern is whether some minimum number of your resources is healthy. You can create a health check for each resource without configuring notification for those health checks. Then you can create a health check that monitors the status of the other health checks and that notifies you only when the number of available web resources drops below a specified threshold.

Amazon Route 53 Management

You can access Amazon Route 53 in the following ways:

- AWS Management Console
- AWS SDKs
- Amazon Route 53 API
- AWS CLI
- AWS Tools for Windows PowerShell

The best tool for monitoring the status of your domain is the Amazon Route 53 dashboard. This dashboard will give a status of any new domain registrations, domain transfers, and any domains approaching expiration.

The tools used to monitor your DNS service with Amazon Route 53 are health checks, Amazon CloudWatch, and AWS CloudTrail. Health checks are discussed in the management section. Amazon CloudWatch monitors metrics like the number of health checks listed as healthy, the length of time an SSL handshake took, and the time it took for the health check to receive the first byte, among other metrics. AWS CloudTrail can capture all of the API requests made for Amazon Route 53. You can determine the user who invoked a particular API.

Amazon Route 53 Authentication and Access Control

To perform any operation on Amazon Route 53 resources, such as registering a domain or updating a resource record set, IAM requires you to authenticate to prove that you're an approved AWS user. If you're using the Amazon Route 53 console, you authenticate your identity by providing your AWS user name and a password. If you're accessing Amazon Route 53 programmatically, your application authenticates your identity for you by using access keys or by signing requests.

After you authenticate your identity, IAM controls your access to AWS by verifying that you have permissions to perform operations and to access resources. If you are an account administrator, you can use IAM to control the access of other users to the resources that are associated with your account.

Amazon CloudFront

Amazon CloudFront is a web service that speeds up distribution of your static and dynamic web content—for example, `.html`, `.css`, `.php`, image, and media files—to end users. Amazon CloudFront delivers your content through a worldwide network of edge locations.

When an end user requests content that you're serving with Amazon CloudFront, the user is routed to the edge location that provides the lowest latency, so content is delivered with the

best possible performance. If the content is already in that edge location, Amazon CloudFront delivers it immediately. If the content is not currently in that edge location, Amazon CloudFront retrieves it from an Amazon S3 bucket or an HTTP server (for example, a web server) that you have identified as the source for the definitive version of your content.

Amazon CloudFront can be used to distribute static content, dynamic content, and streaming media.

Amazon CloudFront Implementation

When implementing Amazon CloudFront, the first step is to configure your origin servers, from which Amazon CloudFront gets your files for distribution from Amazon CloudFront edge locations all over the world. An origin server stores the original, definitive version of your objects.

If you're serving content over HTTP, your origin server is either an Amazon S3 bucket or an HTTP server, such as a web server. Your HTTP server can run on an Amazon EC2 instance or on a server that you manage; these servers are also known as custom origins. If you distribute media files on demand using the Adobe Media Server Real-Time Messaging Protocol (RTMP), your origin server is always an Amazon S3 bucket.

The next step is to upload your files to your origin servers. Your files, also known as objects, typically include web pages, images, and media files, but they can be anything that can be served over HTTP or a supported version of Adobe RTMP, the protocol used by Adobe Flash Media Server.

The final step is to create an Amazon CloudFront distribution, which tells Amazon CloudFront which origin servers to get your files from when users request the files through your website or application. In addition, you can configure your origin server to add headers to the files; the headers indicate how long you want the files to stay in the cache in Amazon CloudFront edge locations.

At this point, Amazon CloudFront assigns a domain name to your new distribution and displays it in the Amazon CloudFront console or returns it in the response to a programmatic request. You can also configure your Amazon CloudFront distribution so that you can use your own domain name.

Now that we have discussed (at a very high level) the steps required to Implement an Amazon CloudFront distribution, let's talk about how that distribution is configured. The following are features that relate to Amazon CloudFront web distributions.

Cache Behaviors

A *cache behavior* is the set of rules that you configure for a given URL pattern based on file extensions, file names, or any portion of a URL path on your website (for example, `*.jpg`). You can configure multiple cache behaviors for your web distribution. Amazon CloudFront will match incoming viewer requests with your list of URL patterns, and if there is a match, the service will honor the cache behavior that you configure for that URL pattern. Each cache behavior can include the following Amazon CloudFront configuration values: origin server name, viewer connection protocol, minimum expiration period, query string parameters, and trusted signers for private content.

Regional Edge Caches

You can use Amazon CloudFront to deliver content at improved performance for your viewers while reducing the load on your origin resources. *Regional Edge Caches* sit in between your origin web server and the global edge locations that serve traffic directly to your viewers. As the popularity of your objects is reduced, individual edge locations may evict those objects to make room for more popular content. Regional Edge Caches have larger cache width than any individual edge location, so objects remain in cache longer at these Regional Edge Caches. This helps keep more of your content closer to your viewers, reducing the need for CloudFront to go back to your origin web server and improves overall performance for viewers. For instance, all our edge locations in Europe now go to the Regional Edge Cache in Frankfurt to fetch an object before going back to your origin web server. Regional Edge Cache locations are currently utilized only for requests that need to go back to a custom origin; requests to Amazon S3 origins skip Regional Edge Cache locations.

You do not need to make any changes to your Amazon CloudFront distributions; this feature is enabled by default for all CloudFront distributions. There is no additional cost for using this feature.

Origin Servers

You can configure one or more origin servers for your Amazon CloudFront web distribution. Origin servers can be an AWS resource, such as Amazon S3, Amazon EC2, Elastic Load Balancing, or a custom origin server outside of AWS. Amazon CloudFront will request content from each origin server by matching the URLs requested by the viewer with rules that you configure for your distribution. This feature allows you the flexibility to use each AWS resource for what it's designed for—Amazon S3 for storage, Amazon EC2 for compute, and so forth—without the need to create multiple distributions and manage multiple domain names on your website. You can also continue to use origin servers that you already have set up without the need to move data or re-deploy your application code. Furthermore, Amazon CloudFront allows the directory path as the origin name; that is, when you specify the origin for a CloudFront distribution, you can specify a directory path in addition to a domain name. This makes it easier for you to deliver different types of content via CloudFront without changing your origin infrastructure.

Private Content

You can use Amazon CloudFront's private content feature to control who is able to access your content. This optional feature lets you use Amazon CloudFront to deliver valuable content that you prefer not to make publicly available by requiring your users to use a signed URL or have a signed HTTP cookie when requesting your content.

Device Detection

Amazon CloudFront edge locations can look at the value of the User Agent header to detect the device type of all the incoming requests. Amazon CloudFront can determine whether the end user request came from a desktop, tablet, smart TV, or mobile device and pass that

information in the form of new HTTP headers to your origin server—Amazon EC2, Elastic Load Balancing, or your custom origin server. Your origin server can use the device type information to generate different versions of the content based on the new headers. Amazon CloudFront will also cache the different versions of the content at that edge location.

Geo Targeting

Amazon CloudFront can also detect the country from where the end users are accessing your content. Amazon CloudFront can then pass the information about the country in a new HTTP header to your custom origin server. Your origin server can generate different versions of the content for users in different countries and cache these different versions at the edge location to serve subsequent users visiting your website from the same country.

Cross-Origin Resource Sharing

Amazon CloudFront may be configured to forward the origin header value so that your origin server (Amazon S3 or a custom origin) can support cross-origin access via *Cross-Origin Resource Sharing (CORS)*. CORS defines a way for client web applications that are loaded in one domain to interact with resources in a different domain.

Viewer Connection Protocol

Content can be delivered to viewers using either the HTTP or HTTPS protocol. By default, your web distribution will accept requests on either protocol. However, if you want all of your content or certain URLs delivered only over an HTTPS connection, you can configure your distribution to only accept requests that come over HTTPS for that content. You can configure this feature separately for each URL pattern in your web distribution as part of the cache behavior for that URL pattern.

Protocol Detection

You can configure Amazon CloudFront to include the protocol (HTTP vs. HTTPS) of your end user's request as part of the cache key to identify an object uniquely in cache. This allows you to customize your content based on the protocol that your end users are using to access your content.

Custom SSL

Custom SSL certificate support lets you deliver content over HTTPS using your own domain name and your own SSL certificate. This gives visitors to your website the security benefits of Amazon CloudFront over an SSL connection that uses your own domain name in addition to lower latency and higher reliability. You can also configure CloudFront to use HTTPS connections for origin fetches so that your data is encrypted end-to-end from your origin to your end users. Configuring custom SSL certificate support is easy; you don't need to learn any proprietary code or hire any consultants to configure it for you.

You can provision SSL/TLS certificates and associate them with Amazon CloudFront distributions within minutes. Simply provision a certificate using the new AWS Certificate

Manager (ACM) and deploy it to your CloudFront distribution with a couple of clicks. Then let ACM manage certificate renewals for you. ACM allows you to provision, deploy, and manage the certificate with no additional charges.

Amazon CloudFront still supports using certificates that you obtained from a third-party certificate authority and uploaded to the IAM certificate store.

Geo Restriction

Geo Restriction or *Geoblocking* lets you choose the countries in which you want to restrict access to your content. By configuring either a whitelist or a blacklist of countries, you can control delivery of your content through Amazon CloudFront only to countries where you have the license to distribute. To enable this feature, you can either use the Amazon CloudFront API or the Amazon CloudFront Management Console. When a viewer from a restricted country submits a request to download your content, Amazon CloudFront responds with an HTTP status code 403 (Forbidden). You can also configure Custom Error Pages to customize the response that Amazon CloudFront sends to your viewers.

TTL Settings: Min, Max, and Default TTL

Amazon CloudFront lets you configure a minimum time-to-live (Min TTL), a maximum TTL (Max TTL), and a default TTL to specify how long CloudFront caches your objects in edge locations. Amazon CloudFront uses the expiration period that your origin sets on your files (through Cache-Control headers) to determine whether CloudFront needs to check the origin for an updated version of the file. If you expect that your files will change frequently, you can configure your origin to set a short expiration period on your files. Amazon CloudFront accepts expiration periods as short as 0 seconds (in which case CloudFront will revalidate each viewer request with the origin). Amazon CloudFront also honors special Cache-Control directives such as private, no-store, and so on. These are often useful when delivering dynamic content that you don't want CloudFront to cache.

If you have not set a Cache-Control header on your files, Amazon CloudFront uses the value of Default TTL to determine how long the file should be cached at the edge before Amazon CloudFront checks the origin for an updated version of the file. If you don't want to rely on the Cache-Control headers set by your origin, you can now easily override the Cache-Control headers by setting the same value for Max TTL, Min TTL, and Default TTL. By setting both a Min TTL and a Max TTL, you can override origin misconfigurations that might cause objects to be cached for longer or shorter periods than you intend. Min TTL, Max TTL, and Default TTL values can be configured uniquely for each of the cache behaviors you define. This allows you to maximize the cache duration for different types of content on your site by setting a lower bound, upper bound, or a default value on the length of time each file can remain in cache.

Query String Parameters

Query string parameters are often used to return customized content generated by a script running on the origin server. By default, Amazon CloudFront does not forward query string parameters (for example, `"?x=1&y=2"`) to the origin. In addition, the query string portion of the URL is ignored when identifying a unique object in the cache. However, you can optionally configure query strings to be forwarded to the origin servers and be included in the unique identity of the cached object. This feature can be enabled separately for each unique cache behavior that you configure. Query string parameters can thus help you customize your web pages for each viewer while still taking advantage of the performance and scale benefits offered by caching content at Amazon CloudFront edge locations.

GZIP

You can configure Amazon CloudFront to apply GZIP compression automatically when browsers and other clients request a compressed object with text and other compressible file formats. This means that if you are already using Amazon S3, CloudFront can transparently compress this type of content. For origins outside S3, doing compression at the edge means that you don't need to use resources at your origin to do compression. The resulting smaller size of compressed objects makes downloads faster and reduces your CloudFront data transfer charges. To use the feature, simply specify within your cache behavior settings that you would like CloudFront to compress objects automatically and ensure that your client adds `Accept-Encoding: gzip` in the request header (most modern web browsers do this by default).

HTTP Cookie Support

Amazon CloudFront supports delivery of dynamic content that is customized or personalized using HTTP cookies. To use this feature, you specify whether you want Amazon CloudFront to forward some or all of your cookies to your custom origin server. You may also specify wildcard characters in the cookie name to forward multiple cookies matching a string format. Amazon CloudFront then considers the forwarded cookie values when identifying a unique object in its cache. This way, your end users get both the benefit of content that is personalized just for them with a cookie and the performance benefits of Amazon CloudFront.

Forward Headers to Origin

You can use Amazon CloudFront to forward all (or a whitelist of) request headers to your origin server. These headers contain information, such as the device used by your visitors or the country from which they accessed your content. You can configure CloudFront to cache your content based on the values in the headers, so that you can deliver customized content to your viewers. For example, if you are hosting multiple websites on the same web server, you can configure Amazon CloudFront to forward the Host header to your origin. When your origin returns different versions of the same object based on the values in the Host header, Amazon CloudFront will cache the objects separately based on those values.

Add or Modify Request Headers Forwarded from Amazon CloudFront to Origin

You can configure Amazon CloudFront to add custom headers or override the value of existing request headers when CloudFront forwards requests to your origin. You can use these headers to help validate that requests made to your origin were sent from CloudFront (shared secret) and configure your origin only to allow requests that contain the custom header values that you specify. This feature also helps with setting up Cross-Origin Request Sharing (CORS) for your CloudFront distribution: You can configure CloudFront always to add custom headers to your origin to accommodate viewers that don't automatically include those headers in requests. It also allows you to disable varying on the origin header, which improves your cache hit ratio, yet forward the appropriate headers so that your origin can respond with the CORS header.

Enforce HTTPS-Only Connection Between Amazon CloudFront and Your Origin Web Server

You can configure Amazon CloudFront to connect to your origin server using HTTPS regardless of whether the viewer made the request by using HTTP or HTTPS.

Support for TLSv1.1 and TLSv1.2 Between Amazon CloudFront and Your Origin Web Server

Amazon CloudFront supports the TLSv1.1 and TLSv1.2 protocols for HTTPS connections between CloudFront and your custom origin web server (along with SSLv3 and TLSv1.0). You can choose the protocols that you want CloudFront to use when communicating with your origin so that you can, for example, choose not to allow CloudFront to communicate with your origin by using SSLv3, which is less secure than TLS.

Default Root Object

You can specify a default file (for example, `index.html`) that will be served for requests made for the root of your distribution without an object name specified, for instance, requests made to `http://abc123.cloudfront.net/` alone, without a file name.

Object Versioning and Cache Invalidation

You have two options to update your files cached at the Amazon CloudFront edge locations. You can use *object versioning* to manage changes to your content. To implement object versioning, you create a unique file name in your origin server for each version of your file and use the file name corresponding to the correct version in your web pages or applications. With this technique, Amazon CloudFront caches the version of the object that you want without needing to wait for an object to expire before you can serve a newer version.

You can also remove copies of an object from all Amazon CloudFront edge locations at any time by calling the *invalidation* API. This feature removes the object from every Amazon CloudFront edge location regardless of the expiration period you set for that object

on your origin server. If you need to remove multiple objects at once, you may send a list of invalidation paths (up to 3,000) in an XML document. Additionally, you can request up to 15 invalidation paths with a wildcard character. The invalidation feature is designed to be used in unexpected circumstances; for example, to correct an encoding error on a video you uploaded or an unanticipated update to your website's CSS file. However, if you know beforehand that your files will change frequently, it is recommended that you use object versioning to manage updates to your files. This technique gives you more control over when your changes take effect, and it also lets you avoid potential charges for invalidating objects.

Access Logs

You can choose to receive more information about the traffic delivered or streamed by your Amazon CloudFront distribution by enabling access logs. *Access logs* are activity records that show you detailed information about each request made for your content. CloudFront access files are automatically delivered multiple times per hour and the logs in those files will typically be available within an hour of your viewers requesting that object.

Amazon CloudFront Usage Charts

Amazon *CloudFront Usage Charts* let you track trends in data transfer and requests (both HTTP and HTTPS) for each of your active CloudFront web distributions. These charts show your usage from each CloudFront region at daily or hourly granularity, going back up to 60 days. They also include totals, average, and peak usage during the time interval selected.

Amazon CloudFront Monitoring and Alarming Using Amazon CloudWatch

You can monitor, alarm, and receive notifications on the operational performance of your Amazon CloudFront distributions using Amazon CloudWatch, giving you more visibility into the overall health of your web application. CloudFront automatically publishes six operational metrics, each at 1-minute granularity, into Amazon CloudWatch. You can then use CloudWatch to set alarms on any abnormal patterns in your CloudFront traffic. These metrics appear in CloudWatch within a few minutes of the viewer's request for each of your Amazon CloudFront web distributions.

Zone Apex Support

You can use Amazon CloudFront to deliver content from the root domain, or "zone apex" of your website. For example, you can configure both `http://www.example.com` and `http://example.com` to point at the same CloudFront distribution, without the performance penalty or availability risk of managing a redirect service.

Using Amazon CloudFront with AWS WAF to Protect Your Web Applications

AWS WAF is a web application firewall that helps detect and block malicious web requests targeted at your web applications. AWS WAF allows you to create rules based on IP addresses, HTTP headers, and custom URIs. Using these rules, AWS WAF can block, allow, or monitor (count) web requests for your web application.

HTTP Streaming of On-Demand Media

Amazon CloudFront can be used to deliver your on-demand adaptive bit-rate media content at scale to a global audience. Whether you want to stream your content using Microsoft Smooth Streaming format to Microsoft Silverlight players or stream to iOS devices using HTTP Live Streaming (HLS) format, you can do so using Amazon CloudFront without the need to set up and manage any third-party media servers. Furthermore, there are no additional charges for using this capability beyond Amazon CloudFront's standard data transfer and request fees. Simply encode your media files for the format you want to use and upload it to the origin they plan to use.

On-demand Smooth Streaming You can specify in the cache behavior of an Amazon CloudFront web distribution to support Microsoft Smooth Streaming format for that origin.

On-demand HLS Streaming Streaming on-demand content using the HLS format is supported in Amazon CloudFront without having to do any additional configurations. You store your content in your origin (for example, Amazon S3). Amazon CloudFront delivers this content at a global scale to a player (such as the iOS player) requesting the HLS segments for playback.

RTMP Distributions for On-Demand Media Delivery

Amazon CloudFront lets you create *RTMP distributions*, which deliver content to end users in real time—the end users watch the bytes as they are delivered. Amazon CloudFront uses Adobe's Flash Media Server 3.5 to power its RTMP distributions. RTMP distributions use the Real-Time Messaging Protocol (RTMP) and several of its variants, instead of the HTTP or HTTPS protocols used by other Amazon CloudFront distributions.

Content Delivery

Now that you have configured Amazon CloudFront to deliver your content, the following will happen when users request your objects:

1. A user accesses your website or application and requests one or more objects.
2. DNS routes the request to the Amazon CloudFront edge location that can best serve the user's request, typically the nearest Amazon CloudFront edge location in terms of latency, and routes the request to that edge location.
3. In the edge location, Amazon CloudFront checks its cache for the requested files. If the files are in the cache, Amazon CloudFront returns them to the user. If the files are not in the cache, CloudFront does the following:
    a. Amazon CloudFront compares the request with the specifications in your distribution and forwards the request for the files to the applicable origin server for the corresponding file type.
    b. The origin servers send the files back to the Amazon CloudFront edge location.
    c. As soon as the first byte arrives from the origin, Amazon CloudFront begins to forward the files to the user.
    d. Amazon CloudFront also adds the files to the cache in the edge location for the next time someone requests those files.

As mentioned earlier, you can configure headers to indicate how long you want the files to stay in the cache in Amazon CloudFront edge locations. By default, each object stays in an edge location for 24 hours before it expires. The minimum expiration time is 0 seconds; there isn't a maximum expiration time limit.

The Amazon CloudFront console includes a variety of reports:

Amazon CloudFront cache statistics reports These reports use the Amazon CloudFront console to display a graphical representation of statistics related to CloudFront edge locations. Data for these statistics are drawn from the same source as CloudFront access logs. You can display charts for a specified date range in the last 60 days, with data points every hour or every day.

Amazon CloudFront popular objects reports These reports are available via the Amazon CloudFront console. You can display a list of the 50 most popular objects for a distribution during a specified date range in the previous 60 days.

Amazon CloudFront top referrers reports These reports provide a list of the 25 domains of the websites that originated the most HTTP and HTTPS requests for objects that CloudFront is distributing for a specified distribution. These top referrers can be search engines, other websites that link directly to your objects, or your own website.

Amazon CloudFront usage reports These are reports which are more detailed than the billing report but less detailed than CloudFront access logs. The usage report provides aggregate usage data by hour, day, or month, and it lists operations by region and usage type.

Amazon CloudFront viewers reports These reports show devices, browsers, and operating systems' versions used to access your content, and also from what locations they are accessing your content.

Amazon CloudFront Management

You can configure Amazon CloudFront using the AWS Management Console, the Amazon CloudFront console, the AWS CLI, or various SDKs available for Amazon CloudFront.

Amazon CloudFront metrics are also accessible in the Amazon CloudWatch Console.

Amazon CloudFront Security

You can control user access to your private content in two ways:

1. Restrict access to objects in the Amazon CloudFront edge cache using either signed URLs or signed cookies.
2. Restrict access to objects in your Amazon S3 bucket so that users can access it through Amazon CloudFront, thus direct access to the S3 bucket. See Amazon S3 bucket policies in Chapter 6, "Storage Systems," for more details.

Understanding how Amazon CloudFront works and how it can deliver content to your end users is important. Knowing how to manage content both at the origin and at the edge is crucial.

Summary

We covered a lot of material in this chapter. While the exam questions may not go into as great a depth in terms of detail, understanding this material will assist you in providing the best answers to the questions on the exam.

In this chapter, we discussed the following:

- Amazon VPC as a logically-isolated section of the AWS Cloud
- AWS Direct Connect and how it allows you to establish a dedicated network connection from your premises to AWS
- The two types of Elastic Load Balancers and their features (Application and Classic Load Balancers)
- How Elastic Load Balancers automatically distribute incoming application traffic across multiple Amazon Elastic Compute Cloud (Amazon EC2) instances within an AWS Region
- Virtual Private Network (VPN) connections and how you can connect Amazon VPC to remote networks to make your AWS infrastructure highly available
- Internet gateways, which are used to connect your public subnets to the Internet
- NAT gateways, which are used to provide connectivity to the Internet from your private instances
- Elastic Network interfaces, which are used to multi-hone an Amazon EC2 instance and can be re-assigned to another Amazon EC2 instance
- Elastic IP addresses (EIP), which are public IPv4 addresses that can be assigned and reassigned to Amazon EC2 instances. IPv6 is not (currently) supported with EIP.
- You learned that Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service. We discussed the various routing types: Simple, Weighted Round Robin (WRR), Latency Based Routing (LBR), geolocation, and Failover routing.
- You learned that Amazon CloudFront is a global Content Delivery Network (CDN) service that accelerates delivery of your websites, Application Programming Interfaces (APIs), video content, or other web assets.

Resources to Review

AWS YouTube channel: `https://www.youtube.com/user/AmazonWebServices`

AWS What's new: `https://aws.amazon.com/new`

Jeff Barr's blog: `https://aws.amazon.com/blogs/aws/`

Amazon VPC documentation: `https://aws.amazon.com/documentation/vpc`

Amazon VPC peering guide:
`http://docs.aws.amazon.com/AmazonVPC/latest/PeeringGuide/Welcome.html`

VPN options:
`http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpn-connections.html`

NAT gateway fundamentals on AWS:
`http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-nat-gateway.html`

Amazon CloudFront documentation:
`https://aws.amazon.com/documentation/cloudfront/`

Amazon Route 53 documentation: `https://aws.amazon.com/documentation/route53`

Elastic Load Balancing documentation:
`https://aws.amazon.com/documentation/elastic-load-balancing/`

AWS Direct Connect and VPN deep dive:
`https://www.youtube.com/watch?v=Qep11X1r1QA`

Amazon CloudFront best practices: `https://www.youtube.com/watch?v=fgbJJ412qRE`

Exam Essentials

Understand what a VPC is. Know how to set up a VPC, and what are the minimum and maximum size of both a VPC and subnets.

Understand the purpose and use of route tables, network ACLs, and security groups. Know how to use each for controlling access and providing security.

Know what are the default values for route tables, network ACLs, and security groups. Know where those default values come from, how to modify them, and why you would modify them.

Understand the difference between a private and public subnet. Public subnets allow traffic to the Internet; private subnets do not. Know how to use Amazon EC2 instances in private subnets to have access the Internet.

Understand the role and function of the various ways to connect the VPC with outside resources. This includes Internet gateway, VPN gateway, Amazon S3 endpoint, VPC peering, NAT instances, and NAT gateways. Understand how to configure these services.

Understand what is an Elastic IP (EIP). Elastic supports public IPv4 addresses (as of this publication). Understand the difference between EIP and an ENI.

Understand what is an Elastic Network Interface (ENI). Elastic Network Interfaces can be assigned and reassigned to an Amazon EC2 instance. Understand why this is important.

Know what services operate within a VPC and what services operate outside a VPC. Amazon EC2 lives within a VPC. Services such as Amazon S3 live outside the VPC. Know the various ways to access these services.

Know what AWS Direct Connect is. Understand why it is used and the basic steps for setting it up. (Remember the seven steps listed in the AWS Direct Connect section of this chapter.)

Understand the concept of VIFs. Understand what a VIF is and the difference between a public and private VIF. Why would you use one versus the other? Understanding these concepts will be very helpful on the exam.

Understand the options for Elastic Load Balancing (Classic Load Balancer vs. Application Load Balancer). Know how each type of load balancer operates, why you would choose one over the other, and how to configure each.

Understand how health checks work in each type of load balancer. Classic Load Balancers and Application Load Balancers have different health check options. Know what they are!

Understand how listeners work. Understand rules, priorities, and conditions and how they interact.

Know how Amazon CloudWatch, AWS CloudTrail, and access logs work. Know what type of information each one provides.

Understand the role of security groups with load balancers. Be able to configure a security group and know how rules are applied.

Understand the various options for establishing an IPsec VPN tunnel from an Amazon VPC to a customer location. Know the operational and security implications of these options.

Know how Amazon Route 53 works as a DNS provider. Understand how it can be used for both public and private hosted zones.

Know what the different routing options are for Amazon Route 53. Understand how to configure the various routing options and how they work.

Know what an Amazon Route 53 routing policy is. Understand how it is applied in Amazon Route 53.

Understand what record types Amazon Route 53 supports and how they work. Know both standard and non-standard record sets.

Know the tools for managing and monitoring Amazon Route 53. Understand how Amazon CloudWatch and AWS CloudTrail work with Amazon Route 53. Go deep in understanding how all of the services in this chapter are monitored.

Know the purpose of Amazon CloudFront and how it works. Know what a distribution is and what an origin is. Know what types of files Amazon CloudFront can handle.

Know the steps to implement Amazon CloudFront. Remember there are three steps to do this.

Know the various methods for securing content in Amazon CloudFront. Know how to secure your content at the edge and at the origin.

Production and Archive

By storing the content (text and media) in Amazon S3 (or Amazon S3-RRS), the CMS is already positioned with the most cost-effective, scalable option to deliver the content.

At some point, the content will be determined to be out of date. In the case of a news website, it may take months before consumers stop searching for or clicking on old stories, but eventually all content will reach a point of diminishing results. When the content reaches that point, automated processes can transition the content to Amazon S3-IA. Why not Amazon Glacier? Because in the case of a news archive, even if consumers don't care about the old story for years, once someone is doing research and wants to open that file (or video), odds are they don't want to wait five hours for the content to be available. By using Amazon S3-IA, the archive is still cost effective but immediately available and searchable.

Additional Storage Solutions

Amazon S3, Amazon EBS, Amazon EFS, and Amazon Glacier represent the core of the AWS storage solutions. When studying to be successful on the exam, you must be very comfortable with these services.

AWS does provide a number of other services that may be considered part of the storage solutions discussion. You should be familiar with them, but they will not likely be a significant portion of the AWS Certified SysOps Administrator – Associate exam.

Amazon CloudFront

Global content distribution can be delivered using *Amazon CloudFront*. This Content Delivery Network (CDN) can help accelerate the delivery of your content by caching copies close to consumers.

Content is stored in origin locations that can be on AWS in an Amazon S3 bucket or served from an Amazon EC2 instance. Amazon CloudFront can even cache content stored on-premises.

Amazon CloudFront is a good choice for distribution of frequently accessed static content that benefits from edge delivery, like popular website images, videos, media files, or software downloads. For on-demand media files, you can also choose to stream your content using Real-Time Messaging Protocol (RTMP) delivery. Amazon CloudFront also supports delivery of live media over HTTP.

Lastly, Amazon CloudFront has a geo-restriction feature that lets you specify a list of countries in which your users can access your content. Alternatively, you can specify the countries in which your users cannot access your content. In both cases, Amazon CloudFront responds to a request from a viewer in a restricted country with an HTTP status code 403 (Forbidden).

AWS Storage Gateway

AWS Storage Gateway allows existing on-premises storage solutions to be extended into AWS. By installing a virtual appliance in the local datacenter, storage can be duplicated or extended into the AWS Region.

For the exam, you will likely see at least a couple of questions that involve AWS Storage Gateway. The important things to understand are the different options available when configuring the appliance.

AWS Storage Gateway Options

The first option is called the *file interface*. This enables you to use Amazon S3 with your on-premises workflows by accessing the Amazon S3 bucket through a Network File System (NFS) mount point. This allows customers to migrate their files into Amazon S3 through object-based workloads while maintaining their on-premises investments.

The second option is the *volume interface*. In this case, you are presented with an iSCSI block storage endpoint. Data is accessed on the local volumes, which is then stored on AWS in the form of Amazon EBS snapshots.

There are two configuration modes using the volume interface. The *cached mode* is designed to extend your local storage. All content is saved in AWS, and only the frequently accessed data is cached locally. This is an excellent operational solution for on-premises storage solutions that are exceeding their local capacity limits. The other mode is the *stored mode*, which is a one-to-one backup of all local content on AWS. This is a primary solution when durable and inexpensive off-site backups are needed. This becomes an excellent start to a hybrid disaster recovery plan.

The third option for AWS Storage Gateway is the *tape interface*. In this configuration, you connect to your existing backup application using an iSCSI Virtual Tape Library (VTL). These virtual tapes are then asynchronously stored in Amazon S3 or Amazon Glacier if less expensive, longer-term storage is needed.

AWS Snowball

For massive data transfers, ground transport can be faster than the Internet. Being able simply to ship large hard drives will get the data into AWS faster (and often cheaper) than public Internet speeds can accommodate. For example, suppose you had 100 TB of data to transfer to AWS. Even over a 100 Mbps data transfer line, it would take 120 days to complete the data transfer. Physically shipping the data is a far faster option.

For those data transfer cases, AWS provides *AWS Snowball*, a physically hardened, secure appliance that ships directly to your on-premises location. Each device weighs around 50 pounds and stores 80 TB (as of this writing). Using multiple AWS Snowball appliances in parallel can provide an easy to implement a migration solution for datasets that are multiple petabytes in size.

AWS Snowball is secured through tamper-resistant seals and a built-in Trusted Platform Module (TPM) that uses a dedicated processor designed to detect any unauthorized modifications to the hardware, firmware, or software.

Data is automatically encrypted when using AWS Snowball with encryption keys managed through AWS KMS. The actual encryption keys are never stored on the appliance.

AWS Snowball with AWS Greengrass

As AWS Snowball has increased in popularity, AWS has added additional functionality, including the ability to access APIs run on the device by leveraging *AWS Greengrass* technology. Those APIs allow the AWS Snowball appliance to act as if it was an Amazon S3 endpoint itself, even when disconnected from the Internet.

AWS Snowmobile

Some companies have dataset transfer needs that range in the exabyte scale. For those unique transfers, AWS can deploy 45-foot-long shipping containers called *AWS Snowmobiles*. Pulled by tractor trailers, each container can transfer up to 100 PB of data.

For security, AWS Snowmobile uses dedicated security personnel, GPS tracking, alarm monitoring, 24/7 video surveillance, AWS KMS encryption, and an optional escort security vehicle while in transit. For more information on AWS KMS, refer to Chapter 3.

Like AWS Snowball, AWS Snowmobile can deliver multiple containers to a single location, providing exabytes of capacity where needed.

Summary

There is no one-size-fits-all solution when it comes to storage. As you prepare for the exam, you need to dive deep into the core storage solutions: the block storage options of Amazon EBS and Amazon EFS and the object storage solutions of Amazon S3 and Amazon Glacier. Make sure that you are comfortable with their use cases and when to deploy each option.

Resources to Review

Amazon EBS documentation: `http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html`

Amazon EFS main page: `https://aws.amazon.com/efs/`

Amazon EFS documentation: `https://aws.amazon.com/documentation/efs/`

Amazon S3 main page: `http://aws.amazon.com/s3/`

Amazon S3 documentation: `http://aws.amazon.com/documentation/s3`

"IAM Policies and Bucket Policies and ACLs! Oh, My! (Controlling Access to Amazon S3 Resources)" blog post: `https://aws.amazon.com/blogs/security/iam-policies-and-bucket-policies-and-acls-oh-my-controlling-access-to-s3-resources/`

Configuring Static Website Hosting section of the Amazon S3 user guide: `http://docs.aws.amazon.com/AmazonS3/latest/user-guide/static-website-hosting.html`

Amazon Glacier main page: `https://aws.amazon.com/glacier/`

Amazon Glacier documentation: `https://aws.amazon.com/documentation/glacier/`

Amazon CloudFront main page: `https://aws.amazon.com/cloudfront/`

Amazon CloudFront documentation: `https://aws.amazon.com/documentation/cloudfront/`

AWS Storage Gateway main page: `https://aws.amazon.com/storagegateway/`

AWS Storage Gateway documentation:

`https://aws.amazon.com/documentation/storage-gateway/`

"File Interface to Storage Gateway" blog post: `https://aws.amazon.com/blogs/aws/category/aws-storage-gateway/`

AWS Snowball main page: `https://aws.amazon.com/snowball/`

AWS Snowball documentation: `https://aws.amazon.com/documentation/snowball/`

What's new at AWS: `https://aws.amazon.com/new`

AWS Simple Monthly Calculator: `http://calculator.s3.amazonaws.com/index.html`

Exam Essentials

Understand block storage vs. object storage. The difference between block storage and object storage is the fundamental unit of storage. With block storage, each file being saved to the drive is broken down into "blocks" of a specific size. With object storage, each file is saved as a single object regardless of size.

Understand when to use Amazon S3 and when to use Amazon EBS or Amazon EFS. This is an architectural decision based on the content type and the rate of change. Amazon S3 can hold any type of data; however, Amazon S3 would not be a good choice for a database or any rapidly changing data types. Remember the case for eventual consistency.

Understand the lifecycle of Amazon EBS volumes. Amazon EBS volumes can be provisioned at the launch of an Amazon EC2 instance or created and attached at a later time. Amazon EBS volumes can persist through the lifecycle of an Amazon EC2 instance, provided the Deleted on Termination flag is set to false. The AWS best practice for an unused Amazon EBS volume is to snapshot it and delete it. The Amazon EBS volume can be created from the snapshot if needed. Remember that with Amazon EBS, you pay for what you provision, as opposed to paying for what you use with Amazon S3.

Understand different types of Amazon EBS volumes. Understand bursting. Amazon EBS volumes are tied to the Availability Zone.

Understand how Amazon EBS snapshots work. Snapshots are stored in Amazon S3; however, you do not access them via the Amazon S3 Console. They are accessed via the Amazon EC2 console under Snapshots.

Understand instance store storage. Depending on the Amazon EC2 instance family, you have access to the local storage. This storage is called the instance store. Snapshots cannot be taken of the instance store. Instance store storage is ephemeral and doesn't survive a restart or termination of the Amazon EC2 instance. For more information, refer to Chapter 4.

Have a detailed understanding of how Amazon S3 lifecycle polices work. Lifecycle configuration enables you to specify the lifecycle management of objects in a bucket. The configuration is a set of one or more rules, where each rule defines an action for Amazon S3 to apply to a group of objects. Know the two types: Transition actions and expiration actions.

Understand Amazon S3 versioning. When Amazon S3 versioning is turned on, it cannot be turned off, only paused. Understand that when versioning is turned on, items deleted are assigned a delete marker and are unable to be accessed. The deleted objects are still in Amazon S3 and you still pay for them.

Understand how to interface with Amazon Glacier. When objects are moved from Amazon S3 to Amazon Glacier, they can only be accessed from the Amazon S3 APIs (for example, in the case of an object that has been moved to Amazon Glacier as the result of a lifecycle policy).

Understand Amazon Glacier vaults. Amazon Glacier stores objects as archives and stores the archives in vaults. You can have (as of this writing) 1,000 vaults per account per region. Vaults are immutable; when retrieving an object, it is not removed from Amazon Glacier but is copied to Amazon S3.

Know how MFA Delete works with Amazon S3. You can optionally add another layer of security by configuring a bucket to enable MFA Delete, which requires additional authentication for changing the versioning state of your bucket and for permanently deleting an object version. MFA Delete requires two forms of authentication together: your security credentials and the combination of a valid serial number, a space, and the six-digit code displayed on an approved authentication device. MFA Delete thus provides added security in the event, for example, that your security credentials are compromised.

Know how to control access to Amazon S3 resources. IAM policies specify what actions are allowed or denied on what AWS resources. Amazon S3 bucket policies, on the other hand, are attached only to Amazon S3 buckets. Amazon S3 bucket policies specify what actions are allowed or denied for which principles on the bucket to which the bucket policy is attached. Amazon S3 bucket policies are a type of ACL.

Understand the features of Amazon CloudFront. Amazon CloudFront is a CDN that can help accelerate the delivery of content by caching copies close to consumers. Content is stored in origin locations, which can be on AWS in an Amazon S3 bucket or served from an Amazon EC2 instance. Amazon CloudFront can even cache content stored on-premises.

Understand media that can be delivered from Amazon CloudFront. This media includes static website CSS style sheets, HTML pages, images, videos, media files, or software downloads. For on-demand media files, you can also choose to stream your content using RTMP delivery. Amazon CloudFront also supports delivery of live media over HTTP.

Understand Amazon CloudFront's geo-restriction feature. This feature lets you specify a list of countries in which your users can access your content. Alternatively, you can specify the countries in which your users cannot access your content.

Understand AWS Storage Gateway's interfaces and modes of operation. The first option is called the file interface. This enables you to use Amazon S3 with your on-premises workflows by accessing the Amazon S3 bucket through an NFS mount point. The second option is the volume interface. In this case, you are presented with an iSCSI block storage endpoint. The third option for AWS Storage Gateway is the tape interface. In this configuration, you connect to your existing backup application using an iSCSI VTL. The file volume interface operates in one of two modes: cached mode and stored mode. Understand the differences.

Test Taking Tip

Multiple choice questions that ask you to choose two or three true answers require that ALL of your answers be correct. There is no partial credit for getting a fraction correct. Pay extra attention to those questions when doing your review.

Exercises

By now you should have set up an account in AWS. If you haven't, now would be the time to do so. It is important to note that these exercises are in your AWS account and thus are not free.

Use the Free Tier when launching resources. The AWS Free Tier applies to participating services across the following AWS Regions: US East (Northern Virginia), US West (Oregon), US West (Northern California), Canada (Central), EU (London), EU (Ireland), EU (Frankfurt), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Sydney), and South America (Sao Paulo). For more information, see `https://aws.amazon.com/s/dm/optimization/server-side-test/free-tier/free_np/`.

If you have not yet installed the AWS Command Line utilities, refer to Chapter 2, "Working with AWS Cloud Services," Exercise 2.1 (Linux) or Exercise 2.2 (Windows).

The reference for the AWS CLI can be found at `http://docs.aws.amazon.com/cli/latest/reference/`.

Getting to know storage systems can be simple when using your own account. Follow the steps in previous chapters to log in to your account. As always, remember that although AWS is a very cost-effective solution, you will be responsible for all charges incurred. If you don't want to keep the items permanently, remember to delete them after your practice session.

Remember, the goal of this book isn't just to prepare you to pass the AWS Certified SysOps Administrator – Associate exam. It should also serve as a reference companion in your day-to-day duties as an AWS Certified SysOps Administrator.