



# **Microsoft Exchange Server 2013 on the AWS Cloud: Quick Start Reference Deployment**

**Mike Pfeiffer**

*January 2015*

*Last updated: September 2015 ([revisions](#))*

## Table of Contents

What We'll Cover .....	3
Quick Start Architecture Overview .....	5
The Microsoft Preferred Architecture for Exchange Server .....	6
Designing for Performance .....	8
Processor Sizing.....	8
Memory Sizing.....	11
Storage Sizing.....	11
Additional Storage Considerations .....	13
Validating Storage and Server Performance.....	14
Log Replication and Network Traffic.....	14
Designing for High Availability .....	15
Regions and Availability Zones.....	15
Active Directory Domain Services.....	15
Namespace Design and Planning.....	16
Database Availability Groups .....	18
Load Balancing Client Access .....	20
Sample Deployment Scenarios .....	22
250 Mailboxes.....	22
Quick Start Architecture Deployment Scenario for 250 Mailboxes.....	22
Preferred Architecture Deployment Scenario for 250 Mailboxes .....	23
2,500 Mailboxes.....	24
Additional Considerations.....	25
Network Security .....	25
Security Groups.....	25

Network ACLs.....	25
Edge Transport Servers.....	26
Reverse Proxy Servers.....	26
Remote Administration.....	27
Encryption at Rest.....	28
Transport Limitations for Amazon EC2 Instances.....	28
Backup Options.....	29
Automated Deployment.....	29
Template Customization.....	29
Post-Configuration Tasks.....	30
Further Reading and Resources.....	31
Send Us Your Feedback.....	33
Document Revisions.....	33

## What We'll Cover

This Quick Start reference deployment guide includes architectural considerations and configurations used to build a Microsoft Exchange Server 2013 environment on the Amazon Web Services (AWS) cloud. We discuss how to build and configure the necessary AWS services, such as Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Virtual Private Cloud (Amazon VPC), to deploy a highly available Exchange Server architecture across separate AWS Availability Zones.

We also provide a sample [AWS CloudFormation](#) template to help you deploy a correctly configured infrastructure predictably and repeatedly. This automated template deploys a Microsoft Active Directory Domain Services (AD DS) and Microsoft Exchange Server 2013 infrastructure in multiple Availability Zones in an Amazon VPC. If you've already deployed AD DS, you can use the second template we've provided to launch this Microsoft Exchange Server infrastructure into an existing VPC.

The automated templates build the minimal infrastructure required to run Microsoft Exchange Server 2013 on AWS with high availability for a small deployment that supports 250 mailboxes. We also provide guidance for two additional deployment scenarios (for 250 and 2,500 mailboxes) aligned with the Microsoft preferred architecture for Exchange Server. This architecture provides greater fault-tolerance and eliminates the need for traditional backups. You can use any of these scenarios as a starting point for your own requirements.

[Launch](#) the AWS CloudFormation template for Exchange Server 2013 into the US West (Oregon) region.



This stack takes approximately three hours to create.

**Costs.** You are responsible for all costs incurred by your use of the AWS services used while running this Quick Start Reference Deployment. As of the date of publication, the cost for creating and running the template with default settings is approximately \$5.50 an hour, but prices are subject to change. See the pricing pages of the specific AWS services you will be using for full details.

**License.** You must obtain a license to Microsoft Exchange Server 2013 prior to deploying this Quick Start. Microsoft Exchange Server 2013 can be deployed and licensed via the [Microsoft License Mobility through Software Assurance](#) program. For development and test environments, you can leverage your existing MSDN licenses for Exchange Server using Amazon EC2 Dedicated Instances. For details, see the [MSDN on AWS](#) page.

This guide targets IT infrastructure architects, administrators, and DevOps personnel. After reading it, you should have a good understanding of how to launch the necessary infrastructure to support Exchange Server 2013 on the AWS cloud.

Deploying this Quick Start with the **default input parameters** builds the Exchange Server 2013 environment illustrated in Figure 1 on the AWS cloud.

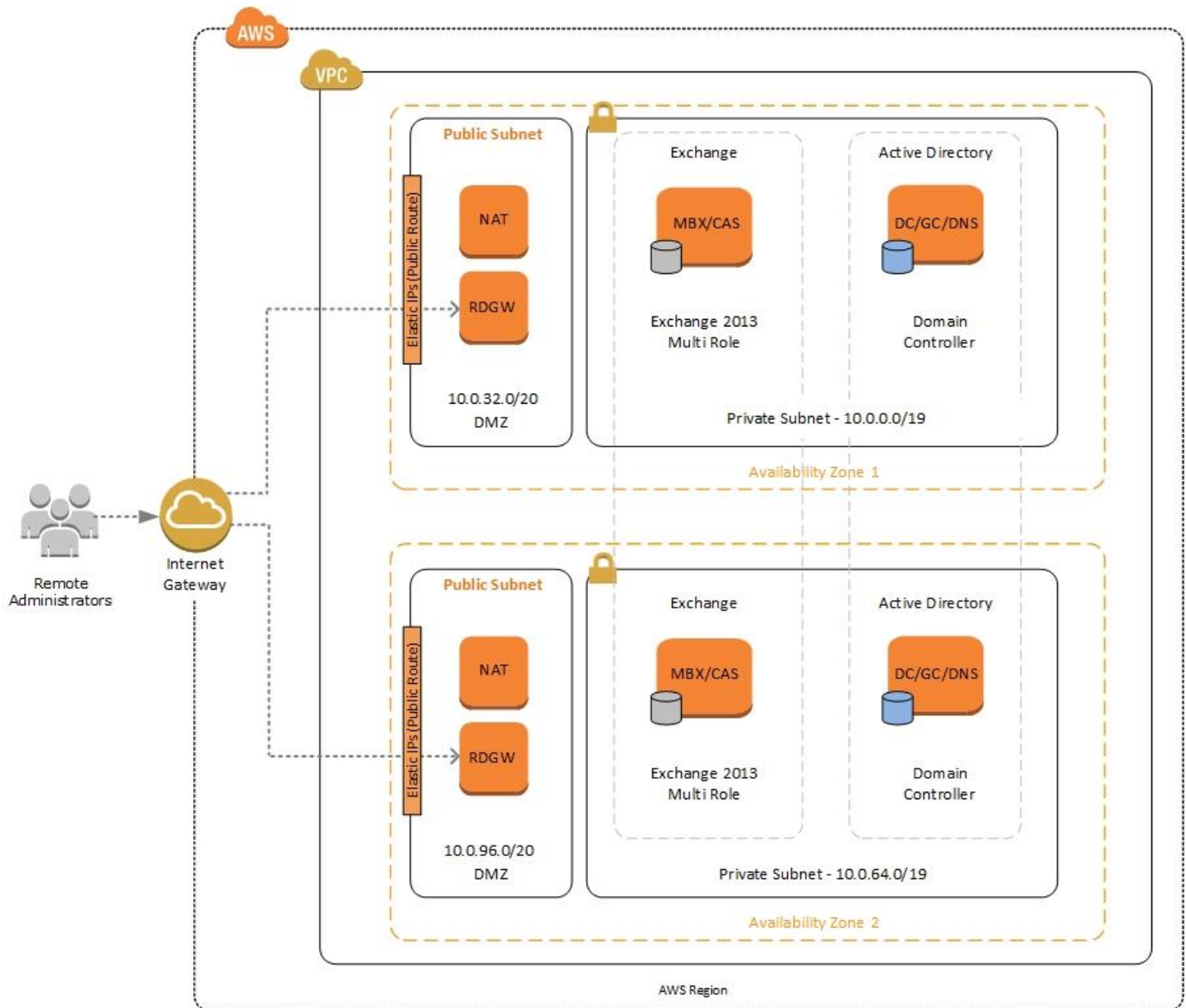


Figure 1: Quick Start Exchange Server 2013 Architecture on AWS

## Quick Start Architecture Overview

This Quick Start gives you the ability to launch an Exchange Server 2013 infrastructure on AWS. The default configuration **deploys the minimal amount of infrastructure to provide Microsoft Exchange Server high availability** for a small deployment that supports 250 mailboxes. The core AWS components used by this Quick Start include the following AWS services:

- [Amazon VPC](#) – The Amazon Virtual Private Cloud (VPC) service lets you provision a private, isolated section of the AWS cloud where you can launch AWS services and other resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

- [Amazon EC2](#) – The Amazon Elastic Compute Cloud (EC2) service allows you to launch virtual machine instances with a variety of operating systems. You can choose from existing Amazon Machine Images (AMIs) or import your own virtual machine images.
- [Amazon EBS](#) – Amazon Elastic Block Store (Amazon EBS) provides persistent block level storage volumes for use with Amazon EC2 instances on the AWS cloud. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability. Amazon EBS volumes provide the consistent and low-latency performance needed to run your workloads.
- [Amazon Route 53 \(optional\)](#) – The Amazon Route 53 service lets you configure Domain Name System (DNS) failover in active-active, active-passive, and mixed configurations to improve the availability of your application. When you have more than one resource—for example, more than one Exchange Server—performing the same function, you can configure Amazon Route 53 to check the health of your resources and respond to DNS queries using only the healthy resources. This Quick Start deploys all the resources shown in Figure 1. You can configure the Amazon Route 53 service manually after you launch the AWS CloudFormation stack.

When deploying a Windows-based environment on the AWS cloud, this Quick Start utilizes an architecture that supports the following best practices:

- Critical workloads are placed in a minimum of two Availability Zones to provide high availability. In this case, the critical workloads are Active Directory domain controllers, Exchange servers, Remote Desktop (RD) gateways for remote administration over the Internet (if needed), Exchange Edge Transport servers, and network address translation (NAT) gateways for outbound Internet access.
- Internal application servers and other non-Internet facing servers are placed in private subnets to prevent direct access to these instances from the Internet. In this Quick Start, domain controllers and multi-role Exchange servers are placed into a private Amazon VPC subnet in each Availability Zone.
- RD gateways are deployed into public subnets in each Availability Zone for remote administration over the Internet. Other components, such as reverse proxy servers, can also be placed into these public subnets if needed. This Quick Start allows you to optionally deploy the Exchange Edge Transport role (an SMTP gateway) into the public subnets for routing Internet email in and out of your environment.

## The Microsoft Preferred Architecture for Exchange Server

In addition to providing the minimal amount of infrastructure for high availability, you may want to consider the Microsoft preferred architecture for Exchange Server 2013 (Exchange PA). Although the Exchange PA calls for running Exchange on dedicated physical servers, it also includes many design aspects that can be beneficial in any environment. The Exchange PA includes the following design requirements (source: [Microsoft Exchange team blog](#)):

- Includes both high availability within the data center, and site resilience between data centers
- Supports multiple copies of each database, allowing for quick activation
- Reduces the cost of the messaging infrastructure
- Increases availability by optimizing around failure domains and reducing complexity

You can think of AWS Availability Zones as separate physical data centers. Following the Exchange PA to architect your Exchange Server deployment on AWS provides the added ability to have high availability within a single AWS Availability Zone, as well as across zones in your AWS region.

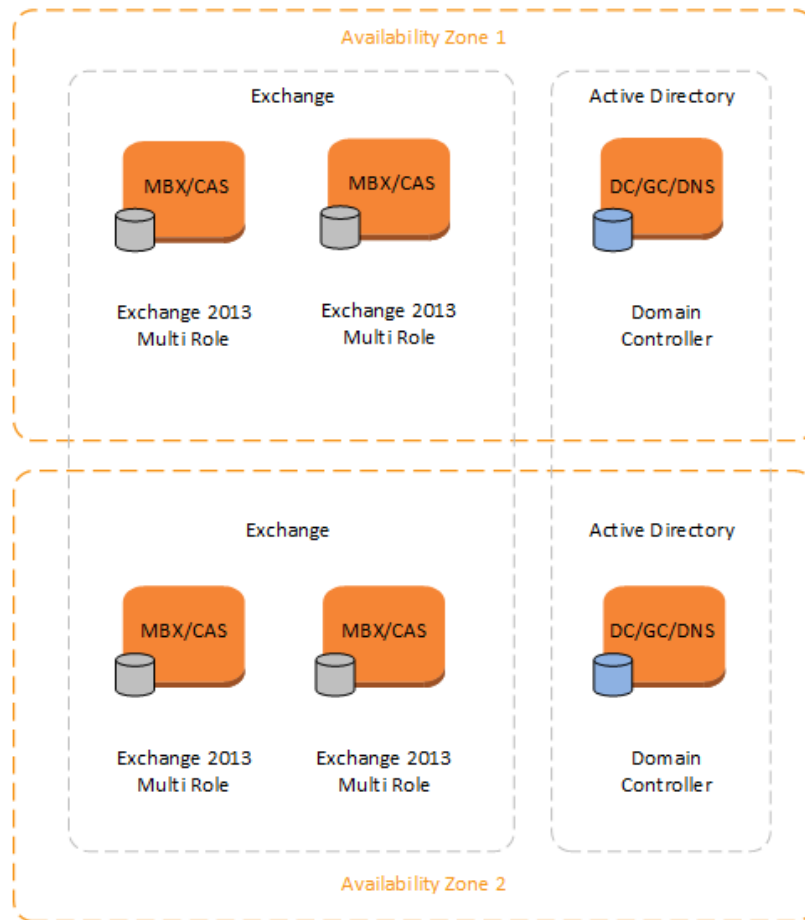


Figure 2: Exchange Server 2013 Architecture on AWS based on the Exchange PA

Following the Exchange PA, each database has four copies, with two copies in each Availability Zone. This means that the Exchange PA requires four servers at a minimum, even for a small deployment of 250 users. Out of these four copies, three servers are configured as highly available (HA) copies, and the fourth server is a lagged database copy configured with the *ReplayLagTime* parameter set to up to 14 days.

Keep in mind that when you design a solution based on the Exchange PA, you end up with an architecture that provides the highest amount of availability, but that also includes a significant amount of infrastructure. As you will see later in this guide, we provide some sample design scenarios so you can get an idea of how much infrastructure is involved. You can use these scenarios to customize your own design.

## Designing for Performance

Much of the capacity planning for a new Microsoft Exchange Server deployment centers on the Exchange Server Mailbox role, and the quantity, activity, and size of the Exchange Server mailboxes required to satisfy the target usage scenario. With that in mind, the first planning tool you should use in any new Exchange Server deployment is the Exchange 2013 Server Role Requirements Calculator. As of this writing, the current version of the calculator is v6.6.

You can download the [Exchange 2013 Server Role Requirements Calculator v6.6](#) from the Microsoft TechNet Gallery.

The following sections address specific areas of the calculator **as they pertain to the AWS cloud infrastructure**. Configure any areas that we do not mention as though your Exchange Server environment were to be deployed using dedicated hardware.

### Processor Sizing

The guest operating system gives you some visibility into the processor specifications of the virtualization host's physical hardware. The actual performance of the virtual CPU type and cores presented to the virtual machine will vary from the dedicated physical equivalent, because the underlying hypervisor controls the scheduling of its tasks. To produce a more quantifiable performance metric, CPU overhead must be accounted for when defining Amazon EC2 instances in the design.

In the Microsoft Exchange 2013 Server Role Requirements Calculator spreadsheet, several fields on the **Input** tab pertain to processor performance. These inputs determine the following:

- Whether you should deduct an arbitrary overhead percentage from the resulting processor performance calculation based on the cost of virtualization
- The number of processor cores that will be made available to the virtual machine
- The optional SPECint2006 performance rating of the target platform onto which the Mailbox role will be deployed

As with any virtualization platform, you must account for hypervisor CPU overhead when planning to run Microsoft Exchange Server 2013. The CPU overhead varies depending on the instance type selected. In addition, AWS is constantly innovating to provide maximum performance for Amazon EC2 instances, and we also add new instance types over time. Therefore, we recommend that you start with a Hypervisor CPU Adjustment Factor of 10% using the Server Role Requirements Calculator. You can perform your own load testing to validate the design and adjust the percentage if needed. Before you start working with the Server Role Requirements Calculator to design CPU requirements, you'll need to determine your planned processor's SPECint2006 Rate Value.

The Microsoft Exchange Server team provides the Exchange Processor Query Tool that will help you determine the SPECint2006 Rate Value. As of this writing, the current version of the tool is v1.1.

You can download v1.1 of the [Exchange Processor Query Tool](#) from the Microsoft TechNet Gallery.

The tool takes a specific processor number as input and searches the SPECint2006 baseline for systems that include that processor number. It outputs an average performance rating for the specified processor.



In step 2 of the Processor Query Tool, we recommend that you use an approximation for your closest matching processor. For example, the R3 instance types use Intel Xeon E5-2670 v2 (formerly Ivy Bridge) processors. If you were planning to use an **r3.2xlarge** instance type, you'd specify **Intel Xeon E5-2670 v2** for the processor model in step 2.

In step 3 of the Processor Query Tool, you query the SPECint2006 baseline for systems that include the specified processor number, and the tool provides an average performance rating for that processor, as shown in Figure 3.

### Exchange 2010 Processor Query Tool

Standard Performance Evaluation Corporation:  
<http://www.spec.org>

---

#### Instructions

**Step 1:** Read and understand the Mailbox Server Processor Capacity Planning article linked above.

**Step 2:** Enter the processor model number to be queried (e.g. X5470).

**Step 3:** Click the button to query the Spec.Org website and obtain the data for the planned processor model.

**Step 4:** Choose the total number of processor cores that will be utilized in your planned mailbox server configuration.

**Step 5:** Examine the data returned by the web query and locate the server model you are planning to use for your Exchange 2010 deployment.  
 Note the SPECint 2006 Rate Value for your planned server model which can be found in the highlighted Result column. If you can't locate the exact server model planned for your deployment, use the average result value listed below.

**Average Result =** **809**

Figure 3: Completing Steps 1-5 in the Processor Query Tool

Next you need to indicate that you'll be deploying virtualized servers. Steps 6 and 7 are used to determine the SPECint2006 Rate Value for the instance type you've selected, as shown in Figure 4.

**Exchange 2010 Processor Query Tool**

**Step 6:** Will you be deploying virtualized mailbox servers? If no, proceed to Step 8. If yes, follow the instructions in Step 7 to calculate the SPECint 2006 rate value of your virtual mailbox role servers.

**Step 7: Virtualized Mailbox Role Servers**  
 Read and understand the System Requirements and support stance for Hardware Virtualization <http://technet.microsoft.com/en-us/library/aa996719.aspx>

- Enter the SPECint 2006 Rate Value of your physical host servers
- Number of physical cores in the host server (from step 4 above)
- Enter the virtual processor ratio that you will use:

Enter 1 if you will be deploying 1:1 virtual processor-to-physical processor on the host  
 Enter 2 if you will be deploying 2:1 virtual processor-to-physical processor on the host

Per Virtual processor SPECint 2006 Rate Value = 40.45

- Enter the number of virtual processors to be allocated to each server

**Virtual Mailbox Server SPECint 2006 Rate Value**

**Step 8:** Go to the latest version of the Mailbox Role Calculator and on the **Input** tab under "Role Requirements Input Factors - Processor Configuration" enter the SPECint2006 Rate Value for your planned mailbox server to determine the adjusted megacycle calculation.

Figure 4: Completing Steps 6-8 in the Processor Query Tool

In this example, the tool provides an average Rate Value result of 809 for the R3 instance processor model. You define a 1:1 virtual-to-physical processor ratio, and enter the number of virtual cores for the selected instance type. In this example, the instance type is r3.2xlarge, which provides 8 vCPUs. The final SPECint2006 Rate Value for the instance is 324.

Next you can move on to the Server Role Requirements Calculator, where you can input the required values. On the **Input** tab, in step 1, indicate that the design will utilize Server Role Virtualization.

Exchange Environment Configuration	Value
Global Catalog Server Architecture	64-bit
Server Multi-Role Configuration (MBX+CAS)	Yes
Server Role Virtualization	Yes
High Availability Deployment	Yes

Figure 5: Server Role Virtualization on the Input Tab

On the **Input** tab, scroll down to the Server Configuration section in step 5, and enter the number of cores and the SPECint2006 Rate Value.

Server Configuration	Processor Cores / Server	SPECint2006 Rate Value
Primary Datacenter Mailbox Server Guest Machines	8	324
Secondary Datacenter Mailbox Server Guest Machines	8	324
Staged Copy Server Guest Machines	16	0

Figure 6: Server Configuration on the Input Tab



Immediately below the **Server Configuration** section, you can define the Hypervisor CPU Adjustment Factor.

Processor Configuration	Value
Hypervisor CPU Adjustment Factor	10%

Figure 7: Processor Configuration on the Input Tab

After you've defined the remaining input values for your design, you can use the **Role Requirements** tab in the calculator to determine if the predicted server CPU utilization is acceptable. We recommend keeping this value below 80%.

Server Configuration	/ Datacenter 1 Server (Double Failure)
Recommended RAM Configuration	48 GB
Number of Processor Cores Utilized	5
Server CPU Utilization	57%
Server CPU Megacycle Requirements	10911
Server Total Available Adjusted Megacycles	19200

Figure 8: Recommended RAM Configuration on the Role Requirements Tab

You might need to deploy more Exchange Server instances, or choose a more suitable instance type, to reduce the estimated CPU utilization so it's within the recommended threshold.

## Memory Sizing

Microsoft Exchange Server memory requirements start with a minimum of 8 GiB for the recommended multi-role configuration. On the Server Role Requirements Calculator, **Role Requirements** tab, the **Recommended RAM Configuration** output value continues upward from 8 GiB, based on the target quantity and size of mailboxes that you specify on the **Input** tab.

AWS offers a number of instance types with different memory configuration options. The general best practice is to pick the instance type with exact or slightly lower memory configuration based on your requirements. For example, you can pick r3.xlarge as a starting point, which provides 30.5 GiB of memory. You may need to switch to a higher memory configuration instance type (for example, r3.2xlarge) depending on the requirements of your deployment.

Server Configuration	/ Datacenter 1 Server (Double Failure)
Recommended RAM Configuration	48 GB

Figure 9: Recommended RAM Configuration on the Role Requirements Tab

For the Exchange Server Edge Transport role, 4 GiB is the required minimum.

## Storage Sizing

The total storage capacity required to support the target deployment scenario is the result of a variety of parameters, including primary and archive mailbox quantity and size, other mailbox usage profile parameters, database storage overhead, volume free space percentage, the number of copies for each database (database availability groups or DAGs), and others.

On the **Input** tab of the Server Role Requirements Calculator, you might need to specify a custom maximum database size, as shown in Figure 10. This will help avoid having the resulting configuration include volumes that exceed the maximum size of an Amazon EBS volume.

Database Configuration	Value
Maximum Database Size Configuration	Custom
Maximum Database Size (GB)	750
Automatically Calculate Number of Unique Databases / DAG	Yes
Custom Number of Databases / DAG	100
Calculate Number of Unique Databases / DAG for Symmetrical Distributi	Yes

Figure 10: Custom Maximum Database Size on the Input Tab

Each Amazon EBS volume attaches to one of 26 possible mount points on the instance: `/dev/sda1` and `xvdb` through `xvdz`. The operating system root volume is mounted at `/dev/sda1`. You can attach additional Amazon EBS volumes or host-based *instance storage* to the 25 remaining mount points. You might need to add multiple instances so the resulting configuration doesn't exceed the maximum number of Amazon EBS volume mount points per instance.

There are three Amazon EBS volume types to choose from:

- **General Purpose (SSD)** - These volumes are the default for Amazon EC2 instances. They're backed by Solid-State Drives (SSDs) and are suitable for a broad range of workloads. General Purpose (SSD) volumes provide the ability to burst to 3,000 IOPS per volume, independent of volume size. This volume type also delivers a consistent baseline of 3 IOPS/GiB and provides up to 128 MiB/s of throughput per volume.
- **Provisioned IOPS (SSD)** - These volumes offer storage with consistent and low-latency performance, and are designed for applications with I/O-intensive workloads. They're backed by Solid-State Drives (SSDs) and support up to 30 IOPS per GiB, which enables you to provision 4,000 IOPS on a volume as small as 134 GiB. You can also achieve up to 128 MiB/s of throughput per volume with as little as 500 provisioned IOPS. Additionally, you can stripe multiple volumes together to achieve up to 48,000 IOPS or 800 MiB/s when attached to larger Amazon EC2 instances.
- **Magnetic** - These volumes provide the lowest cost per GiB of all Amazon EBS volume types. Magnetic volumes are backed by magnetic drives and are ideal for workloads where data is accessed infrequently, and scenarios where the lowest storage cost is important. Magnetic volumes provide approximately 100 IOPS on average, with an ability to burst to hundreds of IOPS. Magnetic volumes are an option for Exchange databases in a non-production environment. For production use, we do not recommend utilizing Magnetic volumes unless you have a small number of mailboxes with very low I/O demand.

For consistent and predictable bandwidth use cases, use EBS-optimized or 10 gigabit network connectivity instances and General Purpose (SSD) or Provisioned IOPS (SSD) volumes. For additional information, see [Amazon EBS Volume Types](#) in the AWS documentation.

You might need to use the Server Role Requirements Calculator to produce several possible configurations and validate the performance of each using the Exchange Server version of the [Microsoft Jetstress tool](#).

In addition to including Amazon EBS volumes, some Amazon EC2 instances include instance storage. Instance storage is ephemeral; the configured instance storage appears as a new disk each time the instance is stopped and started again. All data located on instance storage is lost after the instance is stopped. Instance storage is included in the runtime cost of the instance.

Instance store volumes are ideal for temporary storage of information that changes frequently, such as operating system paging and temporary file storage.

Keep the following recommendations in mind when you plan the quantity and size of volumes that your Exchange Server instances will require for the volumes that **will not** host Microsoft Exchange Server database files:

- Choose a root volume size that will provide sufficient capacity for patching and diagnostic logging. This Quick Start deploys Exchange multi-role servers with a 300 GiB root volume. You can change the default root volume size when launching the instance either through the AWS Management Console or by using AWS CloudFormation templates.
- Place the operating system paging file on volumes separate from the operating system files for optimal performance. Instance storage is an excellent candidate for this.

When designing for storage using the Server Role Requirements Calculator, on the **Input** tab, step 4, you can use the **Server Disk Configuration** section for each data center, and select **7.2K RPM SATA 3.5"** for the **Disk Type** value. For system volumes, set the **Disk Capacity** to at least 100 GiB. For Database + Log and Restore volumes, set the value to the maximum Amazon EBS volume size, 1,000 GiB. This disk type selection most closely resembles the performance characteristics of a standard Amazon EBS volume.

Datacenter 1 Server Disk Configuration		Disk Capacity	Disk Type
System		300 GB	7.2K RPM SATA 3.5"
Database + Log		1000 GB	7.2K RPM SATA 3.5"
Logs		1000 GB	7.2K RPM SATA 3.5"
Restore Volume		1000 GB	7.2K RPM SATA 3.5"

Figure 11: Disk Configuration on the Input Tab

On the **Input** tab, step 3, the **Backup Methodology** value must be set to either **Software VSS Backup/Restore** or **Exchange Native Data Protection**.

#### Note

Microsoft does not recommend using only Exchange Native Data Protection (no backups/replication only) unless you have configured three or more copies of each database within a database availability group. For architectures based on the Exchange PA, you can select Exchange Native Data Protection.

Backup Configuration	Value
Backup Methodology	Software VSS Backup/Restore

Figure 12: Backup Configuration on the Input Tab

## Additional Storage Considerations

As you provision Amazon EBS volumes to support your Microsoft Exchange Server databases and log files, Microsoft recommends formatting those NTFS volumes with an allocation unit size of 64 KiB. The automated AWS CloudFormation template automatically provisions Amazon EBS volumes for our sample deployment scenario, and we handle formatting the volumes based on recommended best practices.

Additionally, Amazon EBS–optimized instances deliver dedicated throughput to Amazon EBS. When attached to an Amazon EBS–optimized instance, General Purpose (SSD) volumes are designed to deliver within 10% of their baseline and burst performance 99.9% of the time in a given year, and Provisioned IOPS (SSD) volumes are designed to deliver

within 10% of their provisioned performance 99.9% of the time in a given year. For more information, see [Amazon EBS–Optimized Instances](#) in the AWS documentation.

There are a number of other Microsoft best practices for configuring storage in your Microsoft Exchange Server deployment. For details on storage architectures and best practices for storage configuration options, see [Exchange 2013 storage configuration options](#) in the Microsoft TechNet Library.

## Validating Storage and Server Performance

Before you place any Microsoft Exchange Server Mailbox role design into production, you should consider testing the storage subsystem to ensure that the design supports the required IOPS within acceptable thresholds for latency. Storage subsystem performance is critical to an acceptable Exchange Server client experience. You can use two tools to validate storage and server performance: Microsoft Exchange Server Jetstress 2013 and Microsoft Exchange Load Generator 2013.

**Microsoft Exchange Server Jetstress 2013** is a free tool provided by the Microsoft Exchange Server team to simulate realistic Exchange Server I/O patterns against one or more test databases. The tool creates the specified test databases on the target volumes and performs simulated transactions for client access, background maintenance, and transaction log replication. You can customize the behavior of the tool through the GUI and the configuration XML file. Throughout the simulation, the tool collects values for a variety of Exchange Server performance counters, and, at the conclusion of the simulation, compares them to acceptable latency thresholds for database and transaction log operations.

You can download the [Microsoft Exchange Server Jetstress 2013 Tool](#) from the Microsoft Download Center.

**Microsoft Exchange Load Generator 2013** helps validate server performance by simulating the server workload that is generated by user interaction with various messaging client software. It is a useful tool for server administrators or messaging deployment engineers who are sizing servers and validating deployment plans. Exchange Load Generator helps you determine whether each of your servers can handle the load that they are intended to carry.

We recommend that you run Exchange Load Generator against your candidate Exchange Server deployment to validate its ability to handle the anticipated client load. Exchange Load Generator affects the performance of all systems involved, so you should run the tool after you have fully configured Exchange Server for your target deployment and before you introduce production user mailboxes or production data.

You can download the [Exchange Load Generator 2013](#) from the Microsoft Download Center.

## Log Replication and Network Traffic

Amazon Availability Zones within the same region are connected through high-speed links. Start to plan your Exchange Server database availability group (DAG) transaction log replication by choosing **Fast Ethernet** for the **Network Link Type** in the Server Role Requirements Calculator, **Input** tab, step 6. Leave **Network Link Latency** at the default of 50.00, or run your own tests between temporary instances to establish a more precise value.

Network Configuration	Value
Network Link Type	Fast Ethernet
Network Link Latency (ms)	50.00

Figure 13: Network Configuration on the Input Tab

You can use the Exchange Client Network Bandwidth Calculator to predict the network bandwidth requirements for a specific set of clients. The prediction algorithms used in this calculator are entirely new and are derived from significant testing and observation.

You can download the [Exchange Client Network Bandwidth Calculator](#) from the Microsoft TechNet Gallery.

## Designing for High Availability

### Regions and Availability Zones

---

You can provision instances in multiple geographic locations called *regions*. You can launch Amazon EC2 instances in these regions so your instances are closer to your customers. For example, you might want to launch instances in Europe to be closer to your European customers or to help meet your legal requirements.

Each region includes *Availability Zones*. Availability Zones are distinct locations that are engineered to be insulated from failures in other zones. They provide inexpensive, low latency network connectivity to other zones in the same region. By launching instances in separate zones, you can protect your applications from any failures that might affect an entire Availability Zone. To help achieve high availability, design your Exchange Server deployment to span two or more Availability Zones.

Based on the needs of your business, you can choose to design your Exchange Server deployment to span multiple regions as well. However, this is more complex and requires additional networking and security, as well as more thorough testing and continuous monitoring.

### Active Directory Domain Services

---

Active Directory Domain Services (AD DS) is a core component of a Microsoft Exchange Server deployment. Exchange Server is tightly coupled with Active Directory. The Active Directory schema must be extended to support additional attributes for objects in which Exchange Server stores configuration settings. Additionally, Active Directory Site Topology is used for routing internal messages between separate physical locations. Designing your deployment based on the Exchange PA assumes that each data center pair (in other words, each Availability Zone) is represented as an individual Active Directory site.

There are three ways to use AD DS in the AWS cloud:

- **Cloud only** – This is the architecture shown earlier in this guide, in Figures 1 and 2. This type of architecture means that your entire Active Directory forest exists only within the AWS cloud. With a cloud-only AD DS architecture, there are no on-premises domain controllers.
- **Traditional hybrid** – The hybrid architecture takes advantage of your existing AD DS environment. You can extend your private, on-premises network to AWS so the resources in the cloud can utilize your existing AD infrastructure. We recommend that hybrid architectures utilize domain controllers from your existing AD forest in the AWS cloud. This is primarily recommended to keep your Exchange servers that are deployed in AWS functional and available in the event of an on-premises outage.
- **AD Connector via AWS Directory Service** – The AD Connector allows you to provision a Directory Service proxy in the AWS cloud. When you have network connectivity from your AWS VPC to the on-premises environment via VPN or AWS Direct Connect, the AD Connector makes it easy to provision Amazon Zocalo sites and Amazon WorkSpaces in your existing AD DS environment. However, the AD Connector should not be used in conjunction

with AWS-based Exchange servers. Exchange servers must have a low latency connection to a writable domain controller and global catalog server in the same Active Directory site in which they reside. We recommend that you use the cloud only or traditional hybrid models for running AD DS, to ensure that writable domain controllers and global catalogs are available in the same Availability Zones as your Exchange servers.

The [Quick Start Reference Deployment for Active Directory Domain Services](#) covers all of our best practices and recommendations for deploying a highly available AD DS environment on AWS. The master AWS CloudFormation template provided in this guide first launches the AD DS Quick Start to provide the foundation for the remaining infrastructure. It's responsible for building the Amazon VPC, public and private subnets, NAT instances and Remote Desktop gateways, and domain controllers in each Availability Zone. We also configure Active Directory Sites and Services as part of this automated deployment. We provision AD sites for each Availability Zone, along with defining objects for each of your Amazon VPC subnets, and mapping them to the appropriate AD site.

Microsoft recommends deploying a ratio of one Active Directory global catalog processor core for every 8 Mailbox role processor cores, assuming domain controllers are running on the x64 (64-bit) Windows platform.

## Namespace Design and Planning

---

Microsoft Exchange Server 2013 includes new functionality that simplifies namespace design. Unlike Exchange Server 2010, Exchange Server 2013 does not require client namespaces to move with the DAG after a failover event. The Client Access role in Exchange Server 2013 proxies requests to the Mailbox server that hosts the active database copy of the user's mailbox, regardless of the Active Directory site in which that server resides. This means that unique namespaces are no longer required for each data center, or in this case, for each Availability Zone.

Based on this new client access architecture, you have two models for namespace design:

- **Unbound namespace** – This model uses a unified namespace that provides access to the Exchange Server infrastructure in each Availability Zone. It allows clients to maintain connectivity without the need to use a different namespace in case one Availability Zone becomes unavailable.
- **Bound namespace** – This model uses a unique namespace for each physical location. Because Availability Zones are connected via high-speed network links, the bound namespace model typically doesn't provide any benefits in a single-region deployment. You might consider this option for a multi-region deployment, to provide a namespace to clients who may be geographically closer to the infrastructure in their region.



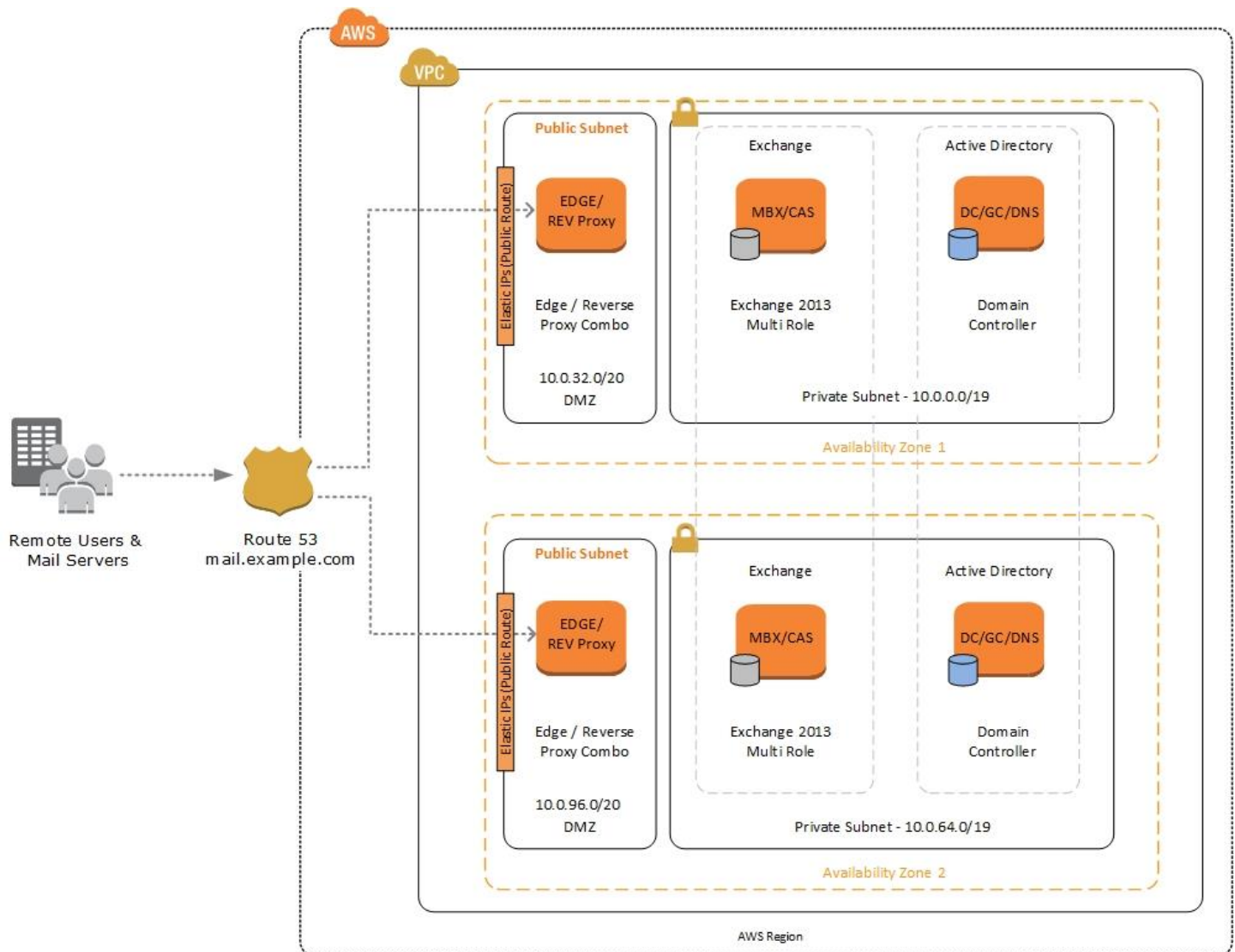


Figure 14: Unbound Namespace Hosted in Amazon Route 53

This Quick Start launches a highly available Microsoft Exchange Server infrastructure in a single region across two Availability Zones. With this architecture we recommend a single, unbound namespace (e.g., mail.example.com). After launching the AWS CloudFormation template and creating the stack, you can proceed to implementing your unified unbound namespace configuration.

Figure 14 includes an Amazon Route 53 hosted zone, along with an active-active failover record set and a reverse proxy solution running on the Edge Transport servers. With this architecture, all client protocols are made highly available through a single unbound namespace. Amazon Route 53, reverse proxy, and Edge Transport configuration options will be explained in greater detail later in this guide.

For more details on namespace design and planning, we recommend reading [Namespace Planning in Exchange 2013](#) on the Microsoft Exchange team blog.

## Database Availability Groups

A database availability group (DAG) is the component for mailbox database high availability and site resilience built into Microsoft Exchange Server 2013. A DAG is a group of up to 16 servers that host a set of databases. DAGs provide automatic database-level recovery from failures that affect entire servers or individual databases. Any server in a DAG can host a copy of a mailbox database from any other server in the DAG. When a server is added to a DAG, it works with the other servers in the DAG to provide automatic recovery from failures that affect mailbox databases, such as a disk, server, or network failures.

### Simple Two-Node DAGs

If you choose not to utilize the Exchange PA for your Exchange Server design, you can implement an architecture similar to the one provided by this Quick Start, which deploys a single, multi-role Exchange Server in each Availability Zone. In this model, you'll have a single DAG, which is stretched across each Availability Zone.

Proper IP addressing needs to be considered when deploying a DAG on AWS. Of course, each DAG member will need a primary IP address for the operating system. Additionally, because Exchange Server DAGs use Windows Server Failover Clustering (WSFC), you must allocate a secondary private IP address to act as the DAG IP address in each Amazon VPC subnet in which the DAG members will reside. DAG IP addresses are used as the Windows Failover Clustering IP Address resource.

<b>EXCH1Privatelp</b>	10.0.0.150	Primary private IP for the first Exchange Server
<b>EXCH1Privatelp2</b>	10.0.0.151	Secondary private IP for the first Exchange Server
<b>EXCH2Privatelp</b>	10.0.64.150	Primary private IP for the second Exchange Server
<b>EXCH2Privatelp2</b>	10.0.64.151	Secondary private IP for the second Exchange Server

Figure 15: IP Configuration That Can Be Customized via Template Parameters

You can assign [multiple private IP addresses](#) to an Amazon EC2 Instance using a single elastic network interface (ENI). The AWS CloudFormation template provided by this Quick Start supports this configuration.

```
Machine: EXCH1.example.com
[PS] C:\>Get-DatabaseAvailabilityGroup | fl DatabaseAvailabilityGroupIpAddresses
DatabaseAvailabilityGroupIpAddresses : {10.0.0.151, 10.0.64.151}
```

Figure 16: Viewing the DAG IP Addresses from the Exchange Management Shell

Figure 16 shows the properties of a DAG that was created after successfully launching this Quick Start. The secondary private IP address for each instance has been statically assigned to the DAG.

Note: Exchange 2013 SP1 running on Windows Server 2012 R2 supports DAGs without a cluster Administrative Access Point (AAP). This means that the cluster does not require an IP Address resource. Clusters created without an AAP can be created with the [New-Cluster](#) cmdlet by setting the *AdministrativeAccessPoint* property value to *None*. You can use this cluster setting if you do not want to use a traditional AAP configuration.

### DAG Configuration in the Preferred Architecture

Designing your Microsoft Exchange Server 2013 deployment to run on AWS, while also adhering to the Exchange PA, provides a model where you have a minimum of two Exchange servers in each Availability Zone participating in a single DAG. Here are some of the main benefits of this architecture:

- You get mailbox database high availability within each Availability Zone, as well as across your AWS region.
- Server load is distributed across a larger number of servers in the event of a failure, which reduces resource utilization on remaining servers.
- Because each Availability Zone contains at least two Exchange servers, you end up with a minimum of four copies of each database. In this model, traditional backups are not required, because you can implement three highly available (HA) copies, and one lagged copy. This provides you with enough database durability to implement Exchange Native Data Protection (a solution that doesn't use backups), which reduces the total cost of ownership (TCO) of your deployment.

DAG architecture on AWS requires additional planning and implementation steps when the environment utilizes multiple DAG members in each Availability Zone. As mentioned previously, each DAG member should be assigned a secondary private IP address in Amazon EC2, which will ultimately be dedicated to the DAG and Windows Failover Clustering. Because the Amazon EC2 instances cannot share these secondary IP addresses, you can place DAG members in separate Amazon VPC subnets. This will allow you to statically define a DAG IP address for each member in the DAG.

In order to support the IP addressing requirements, each DAG member is placed within its own Amazon VPC subnet, and the secondary private IP address assigned to each associated instance can be used when defining the DAG. This will ensure that each DAG member can successfully bring the defined IP address online when required.

### Witness Server Placement

In order to provide seamless automatic failover between Availability Zones, we recommend that you place your witness server(s) in a third Availability Zone. This helps ensure that cluster quorum, and therefore automatic failover, can be maintained and achieved in the event of a complete Availability Zone outage, regardless of which Availability Zone becomes unavailable.

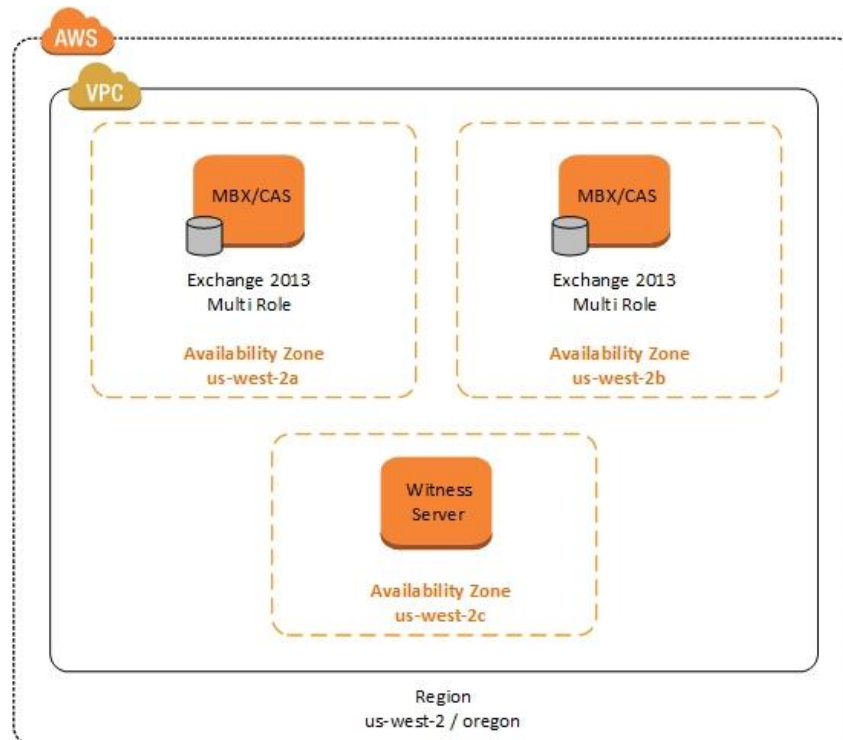


Figure 17: Placing the Witness Server in a Third Availability Zone

Witness server placement in a third Availability Zone is a design aspect that you must implement manually after launching this Quick Start. By default, this Quick Start will launch in Oregon, which includes three Availability Zones at the time of this writing. You can launch the Quick Start using the default settings, and proceed to implement a witness server in a third Availability Zone if desired.

### DAG Network Design

Exchange Server 2010 architectures commonly included multiple network interfaces per mailbox server configured in a DAG. This provided individual network interfaces for client (MAPI) facing networks, along with an isolated and dedicated replication network for database replication.

With fast network throughput becoming commonplace in modern Microsoft Exchange Server designs, and the fact that network interfaces are often only a logical (not physical) separation, Microsoft is moving away from the previous guidance of separating client traffic from replication traffic. This will simplify the management and initial configuration of your Exchange Server deployment on AWS. We recommend that you utilize instances with high network performance or 10 gigabit connectivity in this model.

DAGs are a large topic, and there are many design considerations and operational procedures that you should be familiar with for larger, more complex deployments. We recommend that you consult the [Database availability groups](#) page in the Microsoft TechNet Library for a deeper dive into the subject.

### Load Balancing Client Access

One of the biggest changes in Exchange Server 2013 was the re-architecture of the Client Access role. Exchange Server 2013 no longer requires session affinity (i.e., sticky sessions) at the load balancing layer. In short, the Client Access role can now proxy client connections to the mailbox server that hosts the active copy of the user's mailbox database. This

means you can reliably use layer 4 load balancing, or DNS load balancing, for client connections to your Exchange Server infrastructure.

### Amazon Elastic Load Balancing

Elastic Load Balancing automatically distributes incoming application traffic across multiple Amazon EC2 instances in the cloud. It enables you to achieve greater levels of fault tolerance in your applications, seamlessly providing the required amount of load balancing capacity needed to distribute application traffic. The acceptable port listeners for both HTTPS/SSL and HTTP/TCP connections are 25, 80, 443, 465, 587, and 1024-65535. If you need to load balance other mail protocols such as POP or IMAP, see the remaining solutions in this section.

### DNS Failover with Amazon Route 53

Amazon Route 53 lets you configure DNS failover in active-active, active-passive, and mixed configurations to improve the availability of your application. When you have more than one resource performing the same function, you can configure Amazon Route 53 to check the health of your resources and respond to DNS queries using only the healthy resources.

The following configurations are commonly used to provide DNS load balancing and/or failover:

- **Active-active failover:** Use this configuration when you want all your resources to be available most of the time. When a resource becomes unavailable, Amazon Route 53 can detect that it's unhealthy and stop including it when responding to queries.
- **Active-passive failover:** Use this configuration when you want a primary group of resources to be available most of the time, and you want a secondary group of resources to be on standby in case all the primary resources become unavailable. When responding to queries, Amazon Route 53 includes only the healthy primary resources. If all the primary resources are unhealthy, Amazon Route 53 begins to include only the healthy secondary resources in response to DNS queries.

Amazon Route 53 is an affordable and easy way to provide DNS failover and load balancing over the Internet for all external Exchange Server protocols. Keep in mind that you'll want to use a low Time to Live (TTL) value for your DNS record set. In the unlikely event of an Availability Zone outage, clients will be temporarily disconnected until they start resolving the healthy IP addresses. This also requires that you implement high availability for your Exchange databases using a DAG.

### Other Load Balancing Options

There are a number of third-party load balancing solutions in the AWS Marketplace. Some examples of commonly used solutions are:

- [Citrix NetScaler VPX](#)
- [KEMP Virtual LoadMaster for AWS](#)
- [F5 BIG-IP Virtual Edition for AWS](#)

For details and general guidance, we recommend reading [Load Balancing in Exchange 2013](#) on the Microsoft Exchange team blog.

## Sample Deployment Scenarios

We have developed three sample deployment scenarios for running Microsoft Exchange Server 2013 on AWS that you can use as a starting point for your own requirements. You should use the Microsoft sizing tools as discussed previously in this guide to create your own design, and perform your own validation testing. The scenarios we cover include the following:

- [250 mailboxes](#) – The first scenario supports 250 mailboxes using the Quick Start architecture, which requires one Exchange server per Availability Zone.
- [250 mailboxes](#) – The second scenario supports 250 mailboxes using an architecture based on the Exchange PA, which requires two Exchange servers per Availability Zone.
- [2,500 mailboxes](#) – This scenario supports 2,500 mailboxes and requires four Exchange servers per Availability Zone.

These scenarios focus solely on the Exchange multi-role server infrastructure placed in private subnets. Follow the guidance in this Quick Start for deploying the Edge Transport and reverse proxy infrastructure in your public subnets for a complete, highly available design.

### Note

Each scenario requires a number of Amazon EBS volumes to be attached to each Exchange Server instance. By default, a soft limit of 20 Amazon EBS volumes (per volume type) per AWS account is imposed. You can [request a limit increase](#) if needed.

Additionally, these deployment scenarios assume that 1-TiB Amazon EBS volumes are used for Exchange database and log files. To support larger mailboxes, you can deploy larger volumes, but be sure to use the Server Role Requirements Calculator to validate that the required IOPS can be achieved.

## 250 Mailboxes

### Quick Start Architecture Deployment Scenario for 250 Mailboxes

For small or medium-sized deployments, you may want to consider a design that utilizes the minimal amount of infrastructure to provide high availability, meaning one Exchange 2013 multi-role server per Availability Zone. Figure 18 provides a summary of key inputs and outputs from the Server Role Requirements Calculator, along with suggestions for Amazon EC2 instance and Amazon EBS volume types to create a design that supports 250 Exchange mailboxes.

250 Mailbox Deployment Scenario	
Mailbox Size	20 GiB
Total Items Send/Receive per Mailbox Daily	200
Average Item Size	75 KiB
Exchange Instance Type	r3.xlarge (4 vCPU, 30.5 GiB)
Required Memory per Exchange Server	24 GiB
Server CPU Utilization	37%

Number of DAGs	1
Total Number of Exchange Databases	10 (28 mailboxes per database; includes 10% projected growth)
Exchange Databases per EBS Volume	1
EBS Volumes per Exchange Server Instance	11 volumes (1 TiB each; includes restore volume)
Number of HA Database Copies per DAG	2
Number of Lagged Database Copies per DAG	0
Data Overhead Factor	20% (to account for unexpected database growth)
Backup Methodology	Software VSS-based backup
EBS Volume Type	General Purpose (SSD)
Exchange Server Instance Count per AZ	1
Exchange Server Instance Count Total	2

**Figure 18: Sample Deployment Scenario for 250 Mailboxes**

The R3 memory-optimized instance types are good candidates for running Microsoft Exchange Server 2013 on AWS, but you can choose any instance type that makes sense based on your requirements. The r3.xlarge, r3.2xlarge, and r3.8xlarge instance types support Amazon EBS optimization and high network performance, and the r3.8xlarge instance type supports 10 gigabit network connectivity. All of the memory-optimized R3 instance types support enhanced networking.

The r3.xlarge instance type provides 30.5 GiB of memory, which is well over the requirements estimated by the calculator, as shown in Figure 18.

#### **Important**

With only two highly available database copies, Exchange Native Data Protection is not recommended, and you should implement your own backup solution. This design uses a maximum mailbox database size of 750 GiB. You may need to consider a smaller database size limit that meets your Recovery Time Objective (RTO) requirements for database restores. In this design, each Exchange Server has been provisioned with an additional Amazon EBS volume dedicated to data restore procedures.

Microsoft recommends against multiple databases per volume when there are fewer than three highly available database copies in a DAG. Therefore, this design uses a one-database-per-volume model.

#### **Note**

This Quick Start will deploy two Exchange servers across two Availability Zones to support this design scenario. However, we only provision one additional Amazon EBS volume per server for Exchange databases, because the default Amazon EBS volume limits must be increased to support the number of volumes required in this design. This can be handled as a manual post-configuration task, after you've requested a limit increase.

## **Preferred Architecture Deployment Scenario for 250 Mailboxes**

Designing an architecture aligned with the Exchange PA provides greater tolerance to failures and eliminates the requirement for traditional backups. The sample scenario for 250 users in an architecture based on the Exchange PA is

shown in Figure 19, which provides a summary of key inputs and outputs from the Server Role Requirements Calculator along with suggestions for Amazon EC2 instance and Amazon EBS volume types.

<b>250 Mailbox Deployment Scenario</b>	
Mailbox Size	20 GB
Total Items Send/Receive per Mailbox Daily	200
Average Item Size	75 KiB
Exchange Instance Type	r3.xlarge (4 vCPU, 30.5 GiB)
Required Memory per Exchange Server	16 GiB
Server CPU Utilization	23%
Number of DAGs	1
Total Number of Exchange Databases	40 (7 mailboxes per database; includes 10% projected growth)
Exchange Databases per EBS Volume	4
EBS Volumes per Exchange Server Instance	10 volumes (1 TiB each)
Number of HA Database Copies per DAG	3
Number of Lagged Database Copies per DAG	1 (7-day log replay delay)
Data Overhead Factor	20% (to account for unexpected database growth)
Backup Methodology	Exchange Native Data Protection
EBS Volume Type	General Purpose (SSD)
Exchange Server Instance Count per AZ	2
Exchange Server Instance Count Total	4

Figure 19: Sample Deployment Scenario for 250 Mailboxes Based on the Exchange PA

This design provides two Exchange servers per Availability Zone. Therefore, the memory and CPU requirements per server have been reduced compared to the previous one-server-per-zone design.

## 2,500 Mailboxes

As you start to build designs that support thousands of users, the minimal infrastructure (two Exchange 2013 multi-role servers) becomes less practical. Generally this is due to the number of Amazon EBS volumes that you would need to attach to each instance. Designs built for a large number of mailboxes will typically require scaling out to multiple Exchange servers per Availability Zone, so you may want to consider following the Exchange PA for your architecture.

Figure 20 provides a summary of key inputs and outputs from the Server Role Requirements Calculator, along with suggestions for Amazon EC2 instance and Amazon EBS volume types to create a design supporting 2,500 mailboxes in an architecture based on the Exchange PA.

<b>2,500 Mailbox Deployment Scenario</b>	
Mailbox Size	5 GB
Total Items Send/Receive per Mailbox Daily	200
Average Item Size	75 KiB
Exchange Instance Type	r3.2xlarge (8 vCPU, 61 GiB)
Required Memory per Exchange Server	48 GiB



Server CPU Utilization	57%
Number of DAGs	1
Total Number of Exchange Databases	56 (25 mailboxes per database; includes 10% projected growth)
Exchange Databases per EBS Volume	4
EBS Volumes per Exchange Server Instance	14 volumes (1 TiB each)
Number of HA Database Copies per DAG	3
Number of Lagged Database Copies per DAG	1 (7-day log replay delay)
Data Overhead Factor	20% (to account for unexpected database growth)
Backup Methodology	Exchange Native Data Protection
EBS Volume Type	General Purpose (SSD)
Exchange Server Instance Count per AZ	4
Exchange Server Instance Count Total	8

Figure 20: Sample Deployment Scenario for 2,500 Mailboxes Based on the Exchange PA

Now that we've increased the mailbox count to 2,500, the memory requirement per instance is 48 GiB. To accommodate this requirement, we've used the r3.2xlarge instance type, which includes 8 vCPUs and 61 GiB of memory.

## Additional Considerations

### Network Security

As with any enterprise application deployment, an Exchange Server organization in AWS should implement strict security controls. AWS provides a comprehensive set of security features that allow you to control the flow of traffic through your Amazon VPC and associated subnets, and ultimately to each Amazon EC2 instance. These features allow you to reduce the attack surface of your environment while providing end-user access to Exchange Server content and applications, as well as administrator access for securely managing the Microsoft Windows Server infrastructure. The following sections discuss the security features and approaches implemented in this reference deployment.

### Security Groups

When launched, Amazon EC2 instances must be associated with at least one security group, which acts as a stateful firewall. You have complete control over the network traffic entering or leaving your security groups, and you can build granular rules that are scoped by protocol, port number, and source/destination IP address or subnet. By default, all outbound (egress) traffic from a security group is permitted. Ingress traffic, on the other hand, must be configured to allow the appropriate inbound traffic to reach your instances.

This Quick Start creates a number of security groups, and each group has several inbound rules for various TCP and UDP ports for server-to-server and client-to-server communication. You can customize these security groups and their associated rules within the provided AWS CloudFormation templates.

### Network ACLs

A network access control list (ACL) is a set of permissions that can be attached to any network subnet in an Amazon VPC to provide stateless filtering of traffic. Network ACLs can be used for inbound or outbound traffic, and provide an effective way to blacklist a CIDR block or individual IP addresses. These ACLs can contain ordered rules to allow or deny

traffic based upon IP protocol, service port, or source or destination IP address. By default, each Amazon VPC subnet includes a network ACL that permits all traffic from any source or destination.

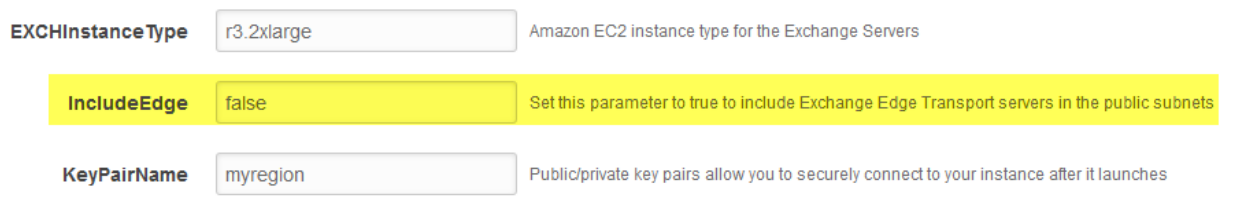
You may choose to either keep the default network ACL configuration or lock it down with more specific rules to restrict traffic between subnets at the network level. One benefit to having multiple layers of network security (security groups and network ACLs) is that they can be managed by separate groups in your organization. If a server administrator inadvertently exposes unnecessary network ports on a security group, a network administrator could override this configuration by using an explicit deny rule, blocking the traffic at the network ACL layer in Amazon VPC.

## Edge Transport Servers

The Exchange Edge Transport server role is deployed in the perimeter network to reduce the attack surface of your messaging system for Internet-facing mail flow. These servers act as an SMTP gateway and run agents that provide additional layers of protection and security.

When you launch this Quick Start, you can optionally include Edge Transport servers in the public subnets in each Availability Zone.

As shown in Figure 22, the AWS CloudFormation template uses a conditional parameter called **IncludeEdge** to control this setting. To include the Edge Transport servers in your deployment, set the parameter to **true**. After the deployment, you'll need to create an Edge Subscription and configure your DNS MX records to resolve to the Elastic IP addresses of your Edge Transport servers. For more information, see [Edge Subscriptions](#) in the Microsoft TechNet Library.



<b>EXCHInstanceType</b>	<input type="text" value="r3.2xlarge"/>	Amazon EC2 instance type for the Exchange Servers
<b>IncludeEdge</b>	<input type="text" value="false"/>	Set this parameter to true to include Exchange Edge Transport servers in the public subnets
<b>KeyPairName</b>	<input type="text" value="myregion"/>	Public/private key pairs allow you to securely connect to your instance after it launches

Figure 22: Specifying Edge Server Options in Template Parameters

The edge servers deployed by this Quick Start will not be domain-joined. To authenticate to these servers for remote administration, you'll need to authenticate locally on each instance. The edge server in the first Availability Zone is named *edge1*, and the edge server in the second Availability Zone is named *edge2*. To log in, use *edge1\administrator* or *edge2\administrator* for the user name, and use the password specified via the **DomainAdminPassword** parameter when launching the AWS CloudFormation template. You can authenticate to the remaining domain-joined servers by using the domain administrator user name and password provided when launching the template.

## Reverse Proxy Servers

When you allow client access to your Exchange infrastructure over the Internet for HTTP based workloads such as Microsoft Outlook Web App, Exchange ActiveSync, or Outlook Anywhere, you can add an additional layer of security by placing reverse proxy/threat management servers into your public Amazon VPC subnets. The public subnets in this reference architecture can be considered a perimeter network (DMZ) that you would typically use in a traditional physical network environment.

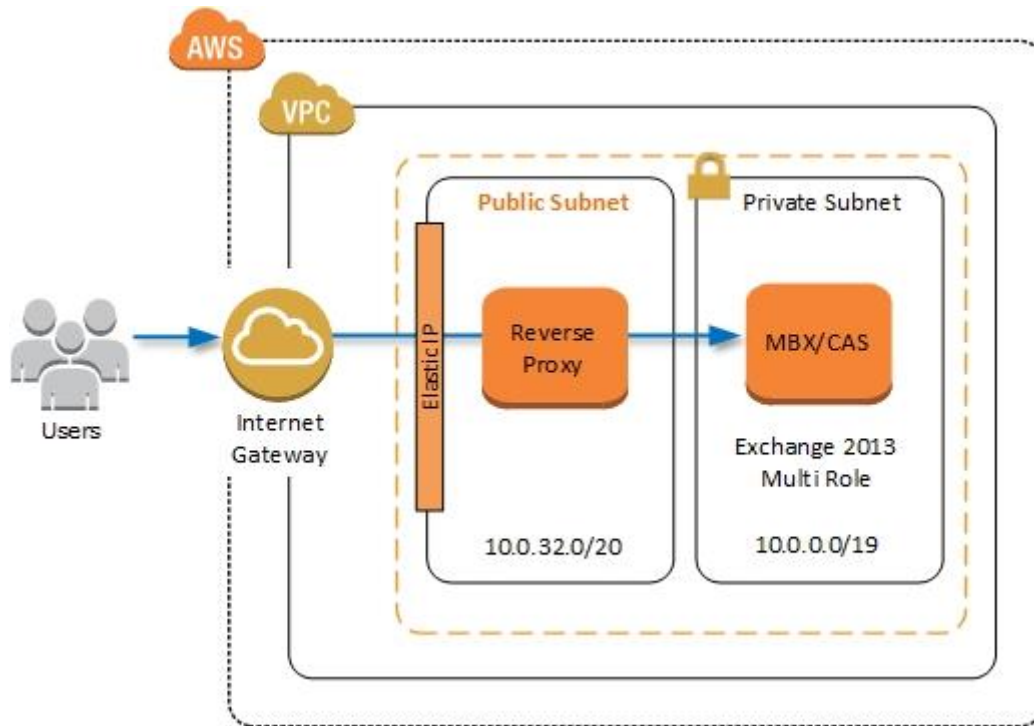


Figure 23: Reverse Proxy Server in a Public VPC Subnet

One of the benefits of this architecture is the ability to pre-authenticate users at the perimeter of your network, while shielding your internal Exchange servers from the public Internet. Several third-party appliances and applications can be used for this task. The Web Application Proxy role in Microsoft Windows Server 2012 R2 also provides support for publishing your Exchange servers to the Internet. Another option is to use Microsoft IIS Application Request Routing (ARR).

Keep in mind that the Exchange Edge Transport handles only the SMTP protocol. If you choose to deploy the optional Edge Transport servers into the public subnets upon launching this Quick Start, you may be able to install reverse proxy software on those servers to handle HTTP requests and minimize the number of instances required in your architecture.

## Remote Administration

When architecting workloads on AWS, you should always eliminate single points of failure. This also applies to any infrastructure you will use for remote administration over the Internet (if required). You can do this in a Windows environment by deploying Remote Desktop (RD) Gateway in each Availability Zone. In case of an Availability Zone outage, this architecture allows access to the resources that may have failed over to the other Availability Zone.

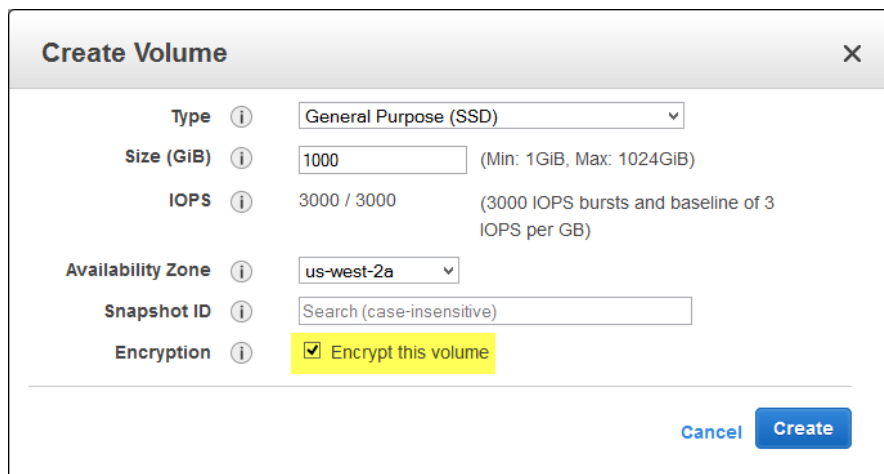
The RD Gateway uses the Remote Desktop Protocol (RDP) over HTTPS to establish a secure, encrypted connection between remote administrators on the Internet and Windows-based, Amazon EC2 instances, without needing to configure a virtual private network (VPN) connection. This allows you to reduce the attack surface on your Windows-based instances while providing a remote administration solution for administrators.

The AWS CloudFormation templates provided in this Quick Start automatically deploy the architecture and configuration steps outlined in the [Quick Start Reference Deployment for the Microsoft Remote Desktop Gateway](#).

After you've launched your Exchange Server infrastructure using the AWS CloudFormation template in this guide, you will initially connect to your instances using a standard RDP TCP port 3389 connection. You can then follow the steps in the [RD Gateway Quick Start deployment guide](#) to secure future connections via HTTPS.

## Encryption at Rest

Amazon EBS encryption offers you a simple encryption solution for your Amazon EBS volumes without requiring you to build, maintain, and secure your own key management infrastructure. When you create an encrypted Amazon EBS volume and attach it to a supported instance type, data stored at rest on the volume, disk I/O, and snapshots created from the volume are all encrypted. The encryption occurs on the servers that host Amazon EC2 instances, providing encryption of data in transit from Amazon EC2 instances to Amazon EBS storage.



The screenshot shows the 'Create Volume' dialog box with the following configuration:

- Type: General Purpose (SSD)
- Size (GiB): 1000 (Min: 1GiB, Max: 1024GiB)
- IOPS: 3000 / 3000 (3000 IOPS bursts and baseline of 3 IOPS per GB)
- Availability Zone: us-west-2a
- Snapshot ID: Search (case-insensitive)
- Encryption:  Encrypt this volume

Buttons: Cancel, Create

Figure 24: Provisioning an Amazon EBS Volume with Encryption

Amazon EBS encryption uses AWS Key Management Service (AWS KMS) Customer Master Keys (CMKs) when creating encrypted volumes, and creating any snapshots from your encrypted volumes. The first time you create an encrypted Amazon EBS volume in a region, a default CMK is created for you automatically. This key is used for Amazon EBS encryption unless you select a CMK that you created separately using AWS KMS. Creating your own CMK gives you more flexibility, including the ability to create, rotate, disable, and define access controls, and audit the encryption keys used to protect your data.

## Transport Limitations for Amazon EC2 Instances

In order to maintain the quality of Amazon EC2 addresses for sending email, we enforce default limits on the amount of email sent from Amazon EC2 accounts. If you want to send larger amounts of email from Amazon EC2, you can apply to have these limits removed from your account. To do so, submit a request using the [AWS Request to Remove Email Sending Limitations](#) form.

If you intend to send email to third parties from Amazon EC2 instances, we also suggest that you provide the Elastic IP address (EIP) for each Exchange Edge Transport server to AWS through the same form. AWS works with ISPs and Internet anti-spam organizations such as Spamhaus to help reduce the chance that email sent from your addresses will be flagged as spam.

You can also help avoid having your email flagged as spam by assigning a static reverse DNS record to the EIP used to send email. You have the option to provide AWS with a reverse DNS record (such as mail.example.com) to associate with

your EIP(s). Note that a corresponding forward DNS record (an **A** record) pointing to your EIP must exist before AWS can create your reverse DNS record. It may take up to a week before the anti-spam organization approves your EIP(s).

## Backup Options

The Microsoft PA for Exchange Server 2013 provides Exchange Native Data Protection, which relies on built-in Exchange Server features to protect your mailbox data without the use of backups. By combining Exchange Native Data Protection with Legal Hold, lagged database copies, and other built-in features, you can reduce or eliminate your reliance on conventional point-in-time backups, and lower your costs.

If you decide to deploy only the minimal amount of infrastructure to provide high availability (i.e., two Exchange multi-role servers), you should implement a traditional backup solution. There are a number of third-party, VSS-based backup solutions that are compatible with Exchange Server.

For details on Exchange Native Data Protection and traditional backup configurations, see the [Backup, restore, and disaster recovery](#) guidance for Exchange Server 2013 in the Microsoft TechNet Library.

## Automated Deployment

This automated AWS CloudFormation template deploys a highly available architecture including Active Directory domain controllers and Exchange 2013 servers in multiple Availability Zones into an Amazon VPC.

**Launch** the AWS CloudFormation template in the US-West (Oregon) region.

### Note

You are responsible for all costs incurred by your use of the AWS services used while running this Quick Start Reference Deployment. As of the date of publication, the cost for creating and running the template with default settings is approximately \$5.50 an hour, but prices are subject to change. See the pricing pages of the specific AWS services you will be using for full details.

The servers in this stack are bootstrapped from scratch using the base Amazon Machine Image (AMI) for Microsoft Windows Server 2012 R2, which allows you to customize the environment based on a number of input parameters in the AWS CloudFormation template. It takes approximately three hours to create.

You can download the nested template (which deploys AD DS and Exchange Server) directly from [https://s3.amazonaws.com/quickstart-reference/microsoft/exchange/latest/templates/Exchange\\_2013\\_Master.template](https://s3.amazonaws.com/quickstart-reference/microsoft/exchange/latest/templates/Exchange_2013_Master.template).

After the stack has been created, you will have two Exchange Server 2013 instances deployed across two Availability Zones. You can navigate to the Exchange Administrative Center (EAC) at <https://exch1/ecp> or <https://exch2/ecp> to configure your Exchange organization. You'll need to sign in to the EAC with the administrative user name and password used when launching the stack.

## Template Customization

This automation allows for rich customization of several template parameters. You can modify these parameters, change the default values, or, if you choose to edit the code of the template itself, you can create an entirely new set of parameters based on your specific deployment scenario. The parameters include the following default values:



Parameter	Default	Description
KeyPairName	<user-provided>	Public/private key pairs, which allow you to connect securely to your instance after it launches
ADInstanceType	m4.xlarge	Amazon EC2 instance type for the first Active Directory instance
AD2InstanceType	m4.xlarge	Amazon EC2 instance type for the second Active Directory instance
NATInstanceType	t2.small	Amazon EC2 instance type for the NAT instances
RDGWInstanceType	m4.xlarge	Amazon EC2 instance type for the Remote Desktop Gateway instance
EXCHInstanceType	r3.xlarge	Amazon EC2 instance type for the Exchange 2013 multi-role servers
EdgeInstanceType	m3.large	Amazon EC2 instance type for the Exchange 2013 Edge Transport servers
DomainDNSName	example.com	Fully qualified domain name (FQDN) of the forest root domain
DomainNetBIOSName	example	NetBIOS name of the domain, for users of earlier versions of Windows (maximum 15 characters)
ADServerNetBIOSName1	DC1	NetBIOS name of the first AD server (maximum 15 characters)
ADServerNetBIOSName2	DC2	NetBIOS name of the second AD server (maximum 15 characters)
RestoreModePassword	<user-provided>	Password for a separate administrator account when the domain controller is in restore mode. This must be a <a href="#">complex password</a> that's at least 8 characters long.
DomainAdminUser	stackadmin	User name for the account that will be added as domain administrator (separate from the default "Administrator" account)
DomainAdminPassword	<user-provided>	Password for the domain administrator user. This must be a <a href="#">complex password</a> that's at least 8 characters long.
UserCount	25	Total number of test user accounts to create in Active Directory
DMZ1CIDR	10.0.32.0/20	CIDR block for the public DMZ subnet located in Availability Zone 1
DMZ2CIDR	10.0.96.0/20	CIDR block for the public DMZ subnet located in Availability Zone 2
PrivSub1CIDR	10.0.0.0/19	CIDR block for the AD server tier located in Availability Zone 1
PrivSub2CIDR	10.0.64.0/19	CIDR block for the AD server tier located in Availability Zone 2
VPCCIDR	10.0.0.0/16	CIDR block for the Amazon VPC
AD1PrivateIp	10.0.0.10	Fixed private IP for the first Active Directory server located in Availability Zone 1
AD2PrivateIp	10.0.64.10	Fixed private IP for the second Active Directory server located in Availability Zone 2
EXCH1PrivateIp	10.0.0.150	Primary private IP for the first Exchange server located in Availability Zone 1
EXCH1PrivateIp2	10.0.0.151	Secondary private IP for the first Exchange server in Availability Zone 1
EXCH2PrivateIp	10.0.64.150	Primary private IP for the second Exchange server located in Availability Zone 2
EXCH2PrivateIp2	10.0.64.151	Secondary private IP for the second Exchange server located in Availability Zone 2
IncludeEdge	false	A value that controls the inclusion of Exchange Edge Transport servers. Set this parameter to <b>true</b> to include Exchange Edge Transport servers in the public subnets

Figure 25: Input Parameters for the AWS CloudFormation Template

If you have already deployed Active Directory Domain Services in AWS, you can launch this Microsoft Exchange Server infrastructure into an existing VPC by using the standalone AWS CloudFormation template for Exchange Server. You can download the standalone template directly from [https://s3.amazonaws.com/quickstart-reference/microsoft/exchange/latest/templates/Template\\_1\\_EXCH\\_2013.template](https://s3.amazonaws.com/quickstart-reference/microsoft/exchange/latest/templates/Template_1_EXCH_2013.template).

## Post-Configuration Tasks

This Quick Start provisions Active Directory Domain Services and Microsoft Exchange Server 2013 across two Availability Zones. There are a number of post-configuration tasks specific to your environment that you should perform to complete the deployment. See the [Microsoft TechNet Library](#) for extensive documentation on how to configure each of these components. We've provided a few links for your convenience.

- [Configure accepted domains and email address policies](#)
- [Configure mail flow and client access](#)
- [Configure certificates](#)
- [Configure the Edge Transport servers](#)
- [Create databases](#)
- [Create a database availability group \(DAG\) for mailbox database high availability](#)
- Implement [load balancing](#) or [Amazon Route 53 DNS failover](#) for client access

## Further Reading and Resources

### AWS services

- AWS CloudFormation  
<http://aws.amazon.com/documentation/cloudformation/>
- Amazon EBS
  - User guide: <http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html>
  - Volume types: <http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSVolumeTypes.html>
  - Optimized instances: <http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSOptimized.html>
- Amazon EC2
  - User guide for Microsoft Windows: <http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/>
  - AWS request to remove email sending limitations: <https://portal.aws.amazon.com/gp/aws/html-forms-controller/contactus/ec2-email-limit-rdns-request>
- Amazon Route 53  
<http://aws.amazon.com/documentation/route53/>
- Amazon VPC  
<http://aws.amazon.com/documentation/vpc/>

### Microsoft Exchange Server 2013

- Microsoft preferred architecture  
<http://blogs.technet.com/b/exchange/archive/2014/04/21/the-preferred-architecture.aspx>
- Storage configuration options  
[http://technet.microsoft.com/en-us/library/ee832792\(v=exchg.150\).aspx](http://technet.microsoft.com/en-us/library/ee832792(v=exchg.150).aspx)
- Namespace planning  
<http://blogs.technet.com/b/exchange/archive/2014/02/28/namespace-planning-in-exchange-2013.aspx>
- Database availability groups  
[http://technet.microsoft.com/en-us/library/dd979799\(v=exchg.150\).aspx](http://technet.microsoft.com/en-us/library/dd979799(v=exchg.150).aspx)

- Load balancing  
<http://blogs.technet.com/b/exchange/archive/2014/03/05/load-balancing-in-exchange-2013.aspx>
- Edge subscriptions  
[http://technet.microsoft.com/en-us/library/aa997438\(v=exchg.150\).aspx](http://technet.microsoft.com/en-us/library/aa997438(v=exchg.150).aspx)
- Backup, restore, and disaster recovery  
[http://technet.microsoft.com/en-us/library/dd876874\(v=exchg.150\).aspx](http://technet.microsoft.com/en-us/library/dd876874(v=exchg.150).aspx)

### Deploying Microsoft software on AWS

- Microsoft on AWS  
<http://aws.amazon.com/microsoft/>
- Secure Microsoft applications on AWS  
[http://media.amazonwebservices.com/AWS\\_Microsoft\\_Platform\\_Security.pdf](http://media.amazonwebservices.com/AWS_Microsoft_Platform_Security.pdf)
- Microsoft Licensing Mobility  
<http://aws.amazon.com/windows/mslicensibility/>
- MSDN on AWS  
<http://aws.amazon.com/windows/msdn/>
- AWS Windows and .NET Developer Center  
<http://aws.amazon.com/net/>

### Tools

- Exchange 2013 Server Role Requirements Calculator  
<https://gallery.technet.microsoft.com/office/Exchange-2013-Server-Role-f8a61780>
- Exchange Processor Query Tool  
<http://gallery.technet.microsoft.com/Exchange-Processor-Query-b06748a5>
- Exchange Server Jetstress 2013  
<http://www.microsoft.com/en-us/download/details.aspx?id=36849>
- Exchange Load Generator 2013  
<http://www.microsoft.com/en-us/download/details.aspx?id=40726>
- Exchange Client Network Bandwidth Calculator  
<https://gallery.technet.microsoft.com/office/Exchange-Client-Network-8af1bf00>
- Load-balancing solutions in the AWS Marketplace  
<https://aws.amazon.com/marketplace/>



## Associated Quick Start Reference Deployments

- Microsoft Active Directory on AWS  
[https://s3.amazonaws.com/quickstart-reference/microsoft/activedirectory/latest/doc/Microsoft\\_Active\\_Directory\\_Quick\\_Start.pdf](https://s3.amazonaws.com/quickstart-reference/microsoft/activedirectory/latest/doc/Microsoft_Active_Directory_Quick_Start.pdf)
- Microsoft Remote Desktop Gateway on AWS  
[https://s3.amazonaws.com/quickstart-reference/microsoft/rdgateway/latest/doc/Microsoft\\_Remote\\_Desktop\\_Gateway\\_Quick\\_Start.pdf](https://s3.amazonaws.com/quickstart-reference/microsoft/rdgateway/latest/doc/Microsoft_Remote_Desktop_Gateway_Quick_Start.pdf)
- Additional reference deployments  
<https://aws.amazon.com/quickstart/>

## Send Us Your Feedback

Please post your feedback or questions on the [AWS Quick Start Discussion Forum](#).

## Document Revisions

Date	Change	In sections
September 2015	In the sample templates, changed the default type for Active Directory and RD Gateway instances from <b>m3.xlarge</b> to <b>m4.xlarge</b> for better performance and price.	<a href="#">Template customization</a>
August 2015	Updated DAG guidance and deployment scenarios.	<a href="#">Database Availability Groups</a> , <a href="#">Sample Deployment Scenarios</a>
March 2015	Optimized the underlying Amazon VPC design to support expansion and to reduce complexity.	Architecture diagram and template updates
January 2015	Initial publication	—

© 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

### Notices

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.