Transferring Data with AWS DataSync

Introduction

AWS DataSync is an online data movement and discovery service that simplifies data migration and helps you quickly, easily, and securely move your file or object data to, from, and between AWS storage services. DataSync can be particularly useful when migrating large amounts of data between clouds or from on-premises systems to the cloud.

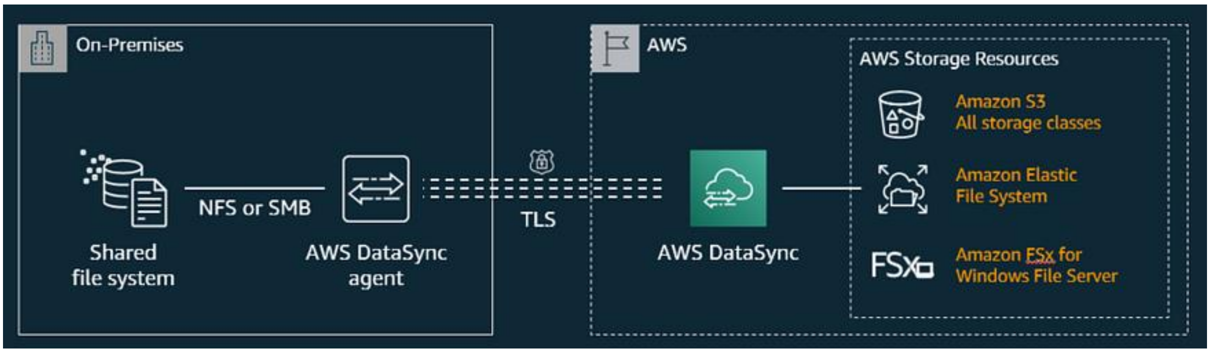
DataSync can copy data to and from:

- Network File System (NFS) file servers
- Server Message Block (SMB) file servers
- Hadoop Distributed File System (HDFS)
- Object storage systems
- Amazon S3 buckets
- Amazon EFS file systems
- Amazon FSx for Windows File Server file systems
- Amazon FSx for Lustre file systems
- Amazon FSx for OpenZFS file systems
- Amazon FSx for NetApp ONTAP file systems
- AWS Snowcone devices
- Google Cloud Storage buckets
- Microsoft Azure Files
- Microsoft Azure Blob Storage

Ajay Kumar, Tech Analyst
Chipsy IT Services Pvt Ltd 17/05/2023 Udupi

Architecture

Below is the high-level architecture of AWS DataSync for copying/migrating data NFS data from on-premise to AWS.



The DataSync agent is deployed on a supported VM in On premise Data centre, connected to storage server and AWS Cloud DataSync service endpoint.

The service endpoint at the AWS end could be 3 types:

1. Private Service Endpoint or VPC End Point
In this type of service endpoint, all communication from on premise DataSync agent to AWS occurs through the VPC endpoint, meaning its uses private network (VPN/Direct connect) established between on-premised data center and AWS.

2. Public Service Endpoint
As the name suggests, all communication from on premise DataSync agent to AWS occurs over the public internet. In this scenario, you do not need to establish direct connect or VPN from your on-premise Datacenter to AWS Cloud.

3. FIPS Service Endpoint
In this type service endpoint, DataSync communicates with the AWS GovCloud (US) or Canada (Central) Region.

Important Consideration

If you have a single VPC shared across multiple AWS accounts using AWS Resource Access Manager (RAM), you can only create the VPC endpoint in the account where VPC is created and you cannot share the VPC endpoint to another account.

Hence the DataSync agent must be setup in the same AWS account where VPC is created. If you have a requirement to transfer the data directly to child account and you are sharing single VPC across all accounts, you can spin up a separate new VPC in child account for the transfer purpose.

Prerequisites

Before we start executing technical setup, we will see the prerequisites you should have or decisions one should make to get started with AWS DataSync.

1. Decide the target AWS account where you want to host your EFS, S3 or FSX.

2. Confirm the transfer method, whether it using private VPC endpoints over direct connect/VPN or public endpoints over the internet.

3. Direct connect/VPN tunnel established to AWS if decided to use VPC endpoint.

4. DataSync agent setup on on-premise as a Virtual Machine (VM).

    a. VMware ESXi Hypervisor (version 6.5, 6.7, or 7.0)

    b. Microsoft Hyper-V Hypervisor (version 2012 R2 or 2016)

    c. Linux Kernel-based Virtual Machine (KVM) *(Not supported for AWS EC2)

    d. AWS EC2

5. NFS feature enabled. Available on Windows 10,11 (Pro and Enterprise Editions only). Available on Ubuntu and other Linux distros as a native feature. However, if the same is not available, it can be installed subsequently.

6. Opening network ports 2049 (NFS), 139/445 (SMB) and 443/80 (Self-managed object storage) between DataSync agent VM and storage server.

7. Opening network ports 1024–1064 for control traffic, 443 (HTTPS) for data transfer if using VPC endpoints scenario.

Agent creation

**AWS CLI should be installed & configured before executing any AWS related commands.

- After installing AWS CLI, execute the command, **‘aws configure’** to setup the AWS account details
- Enter the **AWS Access Key, Secret Access Key, default region i.e ap-south-1** and the **default format** i.e **text**

Use the CLI command below to acquire the most recent DataSync Amazon Machine Image (AMI) ID for the selected AWS Region.

‘aws ssm get-parameter --name /aws/service/datasync/ami --region ap-south-1’

Launch the agent using your AMI from the Amazon EC2 launch wizard from the AWS account where the source file system is stored. To start the AMI, go to the following URL: **https://console.aws.amazon.com/ec2/v2/home?region=source-file-system-region#LaunchInstanceWizard:ami=ami-id**

****Replace the value in region i.e source-file-system-region to 'ap-south-1' and the value in ami i.e ami-id with the ami id obtained after executing the command in the previous step.**

Note: -

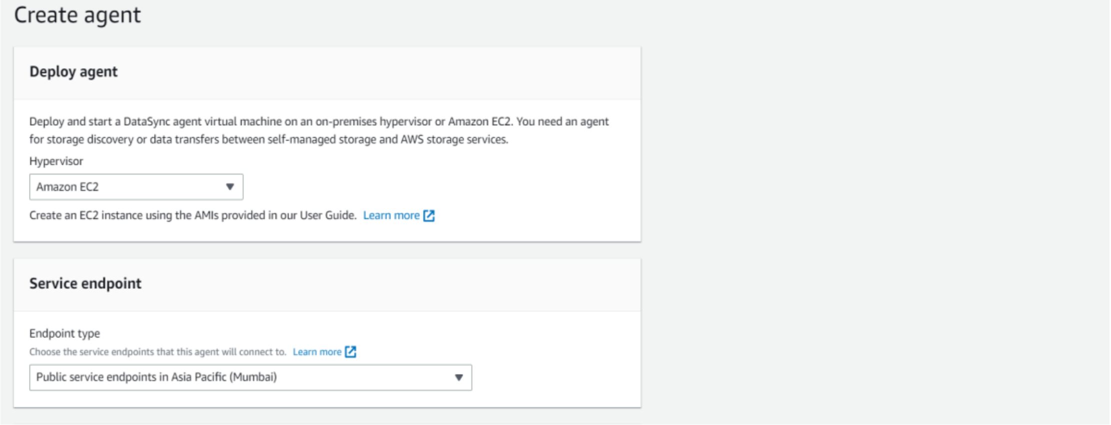
1. **Please make sure that you are launching a machine having RAM of 16 GiB or more i.e an instance with t2.xlarge or m5.xlarge is preferable.**

2. **It is a better option to use Elastic IP address for the Agent instance.**

DataSync Setup

1. Open the DataSync service from the AWS Console and select the below option for the data transfer option. Click on **Discover Storage** or **Transfer Data**. Click on **Agents** from the side-menu and click on **Create Agent**.

2. Select the Hypervisor as AWS EC2, select the Endpoint type as Public service Endpoints,



Create agent

Deploy agent

Deploy and start a DataSync agent virtual machine on an on-premises hypervisor or Amazon EC2. You need an agent for storage discovery or data transfers between self-managed storage and AWS storage services.

Hypervisor: Amazon EC2

Create an EC2 instance using the AMIs provided in our User Guide. Learn more

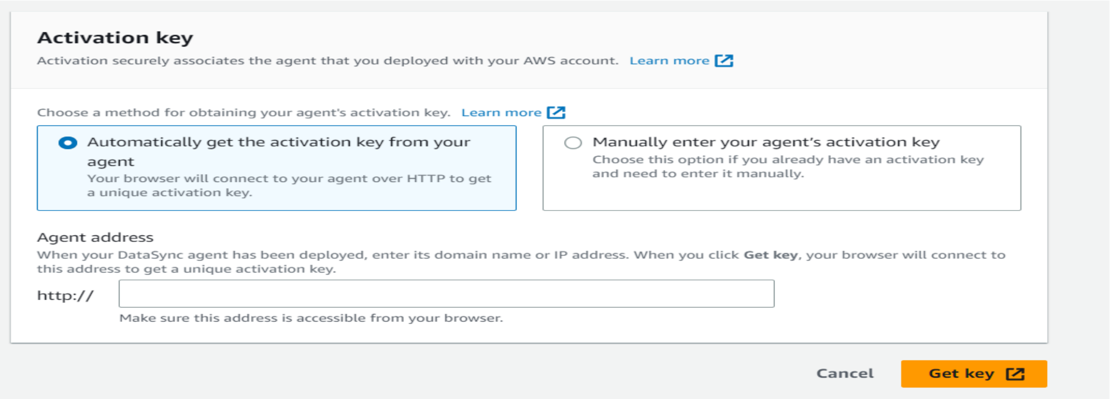
Service endpoint

Endpoint type

Choose the service endpoints that this agent will connect to. Learn more

Public service endpoints in Asia Pacific (Mumbai)

3. Provide the public IP of the agent in the Agent Address as below & click on **Get Key**



Activation key

Activation securely associates the agent that you deployed with your AWS account. Learn more

Choose a method for obtaining your agent's activation key. Learn more

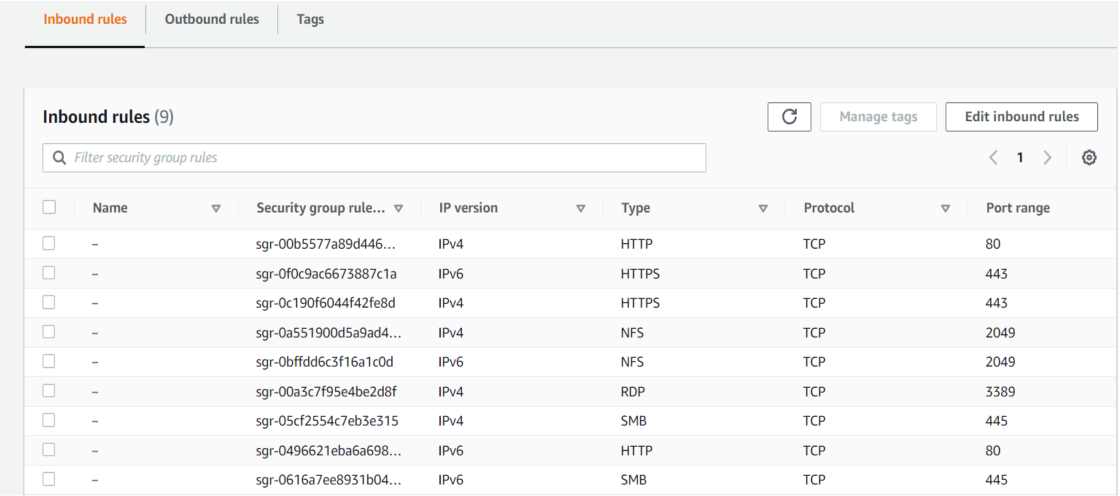
● Automatically get the activation key from your agent
Your browser will connect to your agent over HTTP to get a unique activation key.

○ Manually enter your agent's activation key
Choose this option if you already have an activation key and need to enter it manually.

Agent address
When your DataSync agent has been deployed, enter its domain name or IP address. When you click **Get key**, your browser will connect to this address to get a unique activation key.

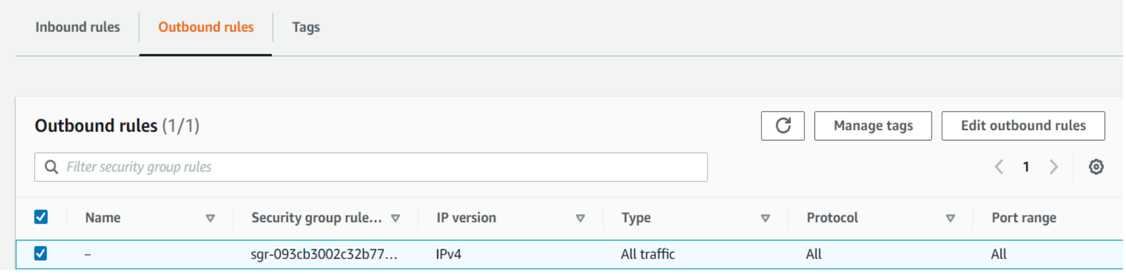
http://
Make sure this address is accessible from your browser.

4. Open the RDP and NFS ports in the On-premise security group



Inbound rules (9)

□	Name	Security group rule...	IP version	Type	Protocol	Port range
□	–	sgr-00b5577a89d446...	IPv4	HTTP	TCP	80
□	–	sgr-0f0c9ac6673887c1a	IPv6	HTTPS	TCP	443
□	–	sgr-0c190f6044f42fe8d	IPv4	HTTPS	TCP	443
□	–	sgr-0a551900d5a9ad4...	IPv4	NFS	TCP	2049
□	–	sgr-0bffdd6c3f16a1c0d	IPv6	NFS	TCP	2049
□	–	sgr-00a3c7f95e4be2d8f	IPv4	RDP	TCP	3389
□	–	sgr-05cf2554c7eb3e315	IPv4	SMB	TCP	445
□	–	sgr-0496621eba6a698...	IPv6	HTTP	TCP	80
□	–	sgr-0616a7ee8931b04...	IPv6	SMB	TCP	445

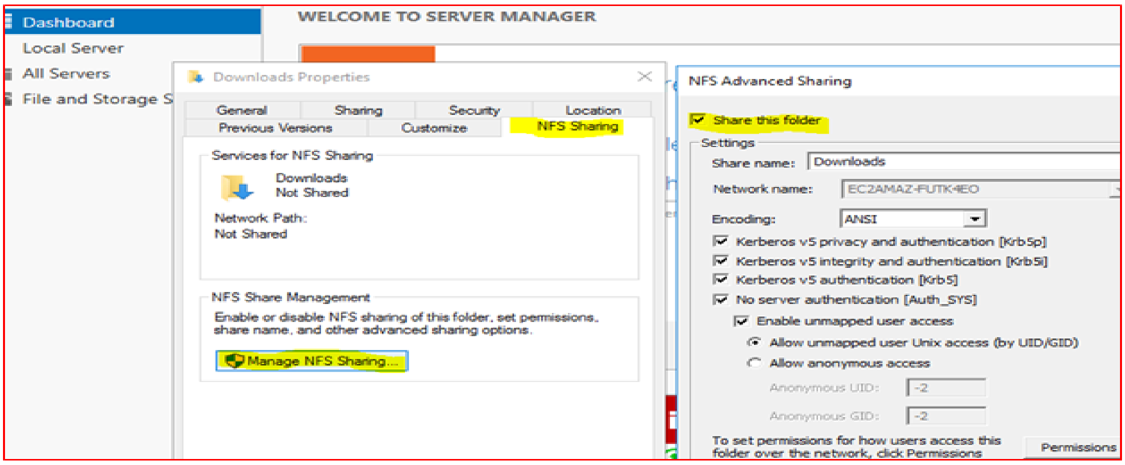


Outbound rules (1/1)

☑	Name	Security group rule...	IP version	Type	Protocol	Port range
☑	–	sgr-093cb3002c32b77...	IPv4	All traffic	All	All

5. Now install the NFS server On-Premise such that it will sync the local files to the cloud through NFS. This can be done through **'Turn Windows features on or off'** menu on Windows 10,11 (Pro and Enterprise Editions Only)

6. Create a folder and text file or any other file inside it. In this particular case, we have created a folder with the name **test** and a text file with the name new file.txt, which we want to sync. Go to the **Properties** of the folder. Select the **NFS Sharing** option and click on **Manage NFS Sharing** and check the **Share this folder** as shown below. Click Apply and then OK to save the configuration over the folder.



Downloads Properties

General | Sharing | Security | Location
Previous Versions | Customize | NFS Sharing

Services for NFS Sharing

Downloads
Not Shared

Network Path:
Not Shared

NFS Share Management

Enable or disable NFS sharing of this folder, set permissions, share name, and other advanced sharing options.

Manage NFS Sharing...

NFS Advanced Sharing

☑ Share this folder

Settings

Share name: Downloads

Network name: EC2AMAZ-FUTK4EO

Encoding: ANSI

☑ Kerberos v5 privacy and authentication [Krb5p]

☑ Kerberos v5 integrity and authentication [Krb5i]

☑ Kerberos v5 authentication [Krb5]

☑ No server authentication [Auth_SYS]

☑ Enable unmapped user access

● Allow unmapped user Unix access (by UID/GID)

○ Allow anonymous access

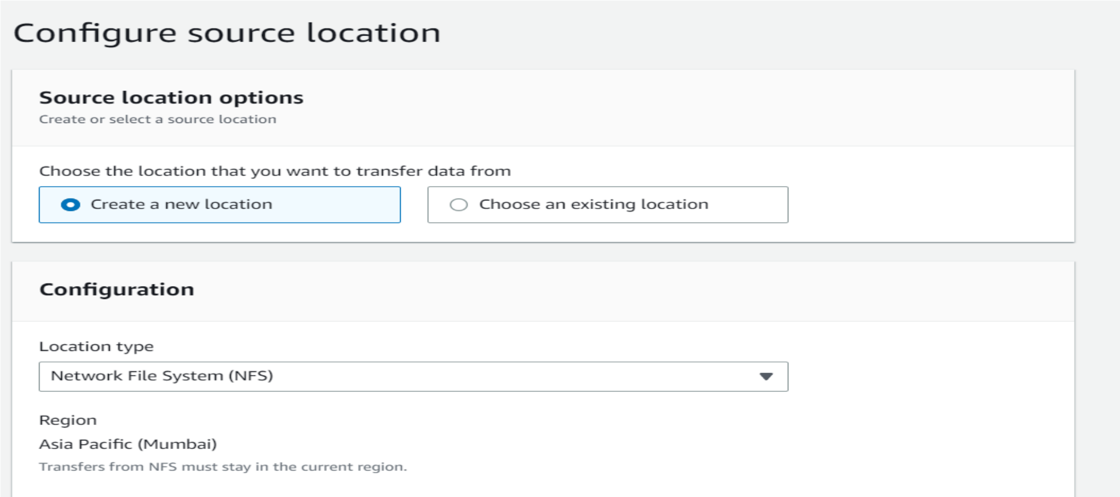
Anonymous UID: -2

Anonymous GID: -2

To set permissions for how users access this folder over the network, click Permissions

Permissions

7. Now go to the DataSync console and create a task and choose the configuration details as below:



Configure source location

Source location options

Create or select a source location

Choose the location that you want to transfer data from

● Create a new location ○ Choose an existing location

Configuration

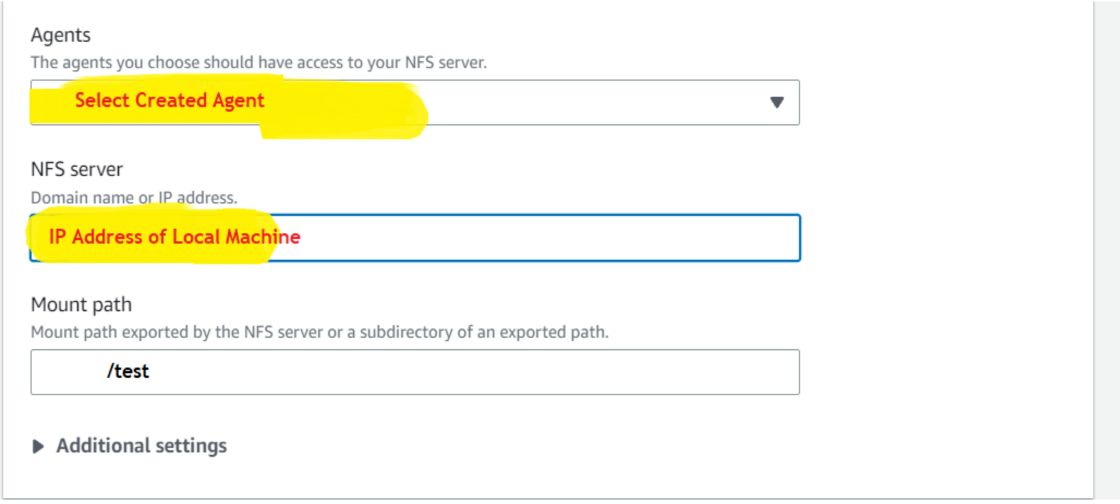
Location type

Network File System (NFS)

Region

Asia Pacific (Mumbai)

Transfers from NFS must stay in the current region.



Agents

The agents you choose should have access to your NFS server.

Select Created Agent ▼

NFS server

Domain name or IP address.

IP Address of Local Machine

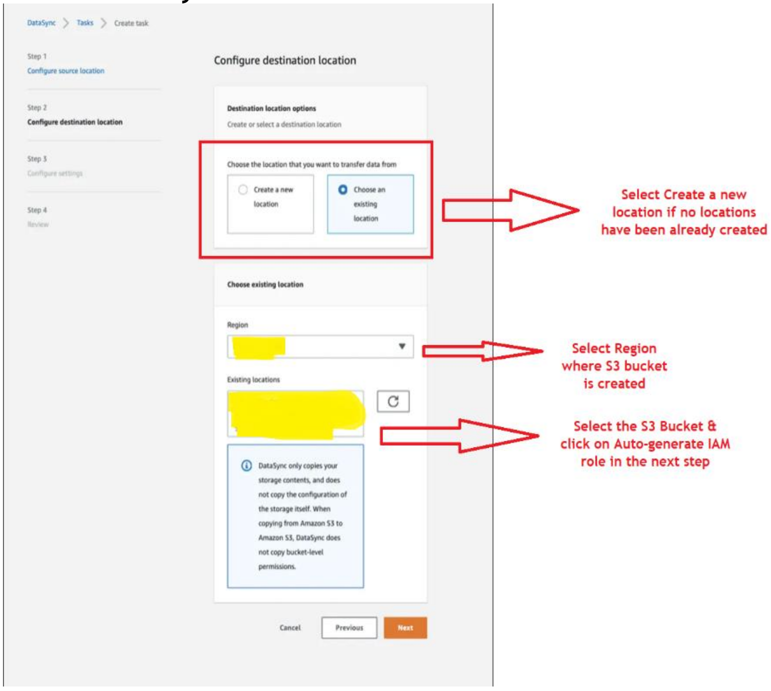
Mount path

Mount path exported by the NFS server or a subdirectory of an exported path.

/test

▶ Additional settings

8. The destination location is S3, and select the S3 bucket to where these files need to be sync



DataSync > Tasks > Create task

Step 1
Configure source location

Step 2
Configure destination location

Step 3
Configure settings

Step 4
Review

Configure destination location

Destination location options

Create or select a destination location

Choose the location that you want to transfer data from

○ Create a new location

● Choose an existing location

Choose existing location

Region ▼

Existing locations

DataSync only copies your storage contents, and does not copy the configuration of the storage itself. When copying from Amazon S3 to Amazon S3, DataSync does not copy bucket-level permissions.

Cancel | Previous | Next

Select Create a new location if no locations have been already created

Select Region where S3 bucket is created

Select the S3 Bucket & click on Auto-generate IAM role in the next step

9. Leave the remaining fields as default and click on create.
*Suppose, you want to delete the object from S3 bucket if it is not available in the source, then uncheck **'Keep deleted files'.**

10. Select the autogenerate for creating the IAM in the provided fields.

11. Once the previous step is completed, wait until your task's status becomes available.

12. Now start the task and wait for a while; this starts syncing your on-premise data to AWS S3 bucket.

Observations

We investigated and highlighted the relevance of data synchronization. When just the changed data is transferred, data synchronization will work smoothly. As a result, each synchronization procedure uses a marker to determine the most up-to-date information.

Conclusion

AWS DataSync is an online data transfer service, which automates and simplifies transferring large amounts of data from on-premise to AWS cloud storage services.

It is more beneficial to use DataSync than a custom script for transferring as it saves up script development time. DataSync is easy to setup, uses parallel multithreaded architecture, encrypts and validates the integration of transferred files. Also does have an option to use VPC endpoints over direct connect/VPN tunnels, so that your data never leaves your corporate network.

It also has an option to limit the use of bandwidth on direct connect/VPN, this is useful if you have other workloads in cloud using the same connection.

Consider the latency, if you are only migrating shared storage to EFS and presenting them to on-premise servers, test it thoroughly in all possible scenarios. It’s always a good idea to combine the migration of storage and servers together to achieve low latency to shared storage.