Load, Store, and Protect Linux-Based NFS Workloads in AWS

A Cloud Migration Guide with AWS DataSync, Amazon Elastic File System, and AWS Backup

This paper has been archived

June 2021

For the latest technical content, refer to the HTML version:

https://docs.aws.amazon.com/whitepapers/latest/load-store-protect-linux-nfs/load-store-protect-linux.html

aws

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided "as is" without warranties, representations, or conditions of any kind, whether express or implied. 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.

This paper has been archived

For the latest technical content, refer to the AWS Whitepapers & Guides page:
https://aws.amazon.com/whitepapers

Contents

This paper has been archived

For the latest technical content, refer to the AWS
Whitepapers & Guides page:
https://aws.amazon.com/whitepapers

Abstract

This paper outlines a strategy for migrating Linux-based applications that require Network File System (NFS) storage to AWS. It describes how to deploy AWS DataSync to transfer files quickly and effectively from an on-premises file system to Amazon Elastic File System (Amazon EFS). This paper also explains how to use the Amazon Elastic File System Infrequent Access storage classes and file lifecycle management to reduce storage costs by up to 92%, and how to use AWS Backup for automated, centrally managed data protection and recovery.

This paper has been archived

For the latest technical content, refer to the AWS Whitepapers & Guides page:
https://aws.amazon.com/whitepapers

Introduction

The depth and breadth of functionality available through cloud computing continues to grow. Organizations are adopting cloud services to gain additional business agility and lower their IT total cost of ownership. Keeping up with the latest available options and developing a plan for building new applications (or moving existing ones into the cloud with the right architecture), however, can be a challenge.

This whitepaper focuses on best practices for migrating Linux file-based applications to the cloud. The vast majority of data that exists today is file-based.

When considering cloud file storage options, you need to satisfy business objectives. These objectives can include high availability, security and compliance, and backup and recovery. You need to meet them while managing cost and planning for smooth migration.

AWS provides straightforward and cost-effective options for loading, storing, and protecting your file-based data in the cloud. To load data, you can use AWS DataSync, a data transfer service designed to automate moving data between on-premises storage and the cloud.

DataSync automatically handles many of the tasks related to data transfers, reducing the need to write scripts to do replication. You can transfer Data Sync to AWS data to AWS using AWS Direct Connect or the internet, at speeds up to 10 times faster than open-source tools. DataSync also performs data integrity checks and can automatically compare the source and destination files at the end of a transfer.

To store Linux-based file data natively in a file system, you can use Amazon EFS. For Windows-based file data, we offer Amazon FSx for Windows File Server. This paper focuses on Linux-based workloads.

Amazon EFS provides a simple, scalable, fully-managed elastic NFS file system for use with AWS Cloud services and on-premises resources. In addition to the data protection features provided by Amazon EFS, such as high availability, high durability, and encryption, you also can use AWS Backup, a fully-managed backup service that makes it easy to centralize and automate file system backups. With AWS Backup console, you can create backup policies that automate backup schedules and retention management.

This whitepaper discusses these services in detail, highlights features, configurability, and cost, and shows how to use them together as part of an overall migration plan.

Service Overview

AWS DataSync

AWS DataSync is an online data transfer service that simplifies, automates, and accelerates moving data between on-premises storage systems and AWS Storage services, as well as between AWS Storage services. This whitepaper focuses on using DataSync in conjunction with Amazon EFS.

DataSync accelerates and automates data transfer—removing the need to modify applications, develop scripts, or manage infrastructure. DataSync uses a purpose-built protocol and a parallel, multi-threaded architecture to accelerate and secure data transfer at speeds up to 10 times faster than open-source tools.

DataSync has simple, per gigabyte pricing, so that you pay only for what you use. DataSync uses a software agent to handle transferring files and supports both one-time migrations and periodic synchronizations of data between on-premises file systems and Amazon EFS.

DataSync provides end-to-end security including encryption and integrity validation, to ensure your data arrives securely, intact, and ready to use. It also supports VPC endpoints, providing you with the option to transfer data without traversing the public internet, further increasing the security of the data in transit.

Amazon Elastic File System

Amazon Elastic File System (Amazon EFS) provides a simple, serverless, set-and-forget, elastic file system that lets you share file data without provisioning or managing storage. It can be used with AWS Cloud services and on-premises resources, and is built to scale on demand to petabytes without disrupting applications.

With Amazon EFS, you can grow and shrink your file systems automatically as you add and remove files, eliminating the need to provision and manage capacity to accommodate growth. EFS supports sharing storage across up to thousands of Amazon Elastic Compute Cloud (Amazon EC2) instances, containers and AWS Lambda functions. You can access Amazon EFS using a Virtual Private Cloud (VPC), over Virtual Private Networks (VPN), or from an on-premises datacenter using AWS Direct Connect.

Amazon EFS offers you the choice of creating file systems using Standard or One Zone storage classes. Standard storage classes store data within and across multiple availability zones (AZ). One Zone storage classes store data redundantly within a single AZ, at a 47% cost savings compared to file systems using Standard storage classes, for workloads that don't require multi-AZ resilience.

Amazon EFS offers four storage classes: two Standard storage classes (Amazon EFS Standard and Amazon EFS Standard-Infrequent Access (EFS Standard-IA)), and two One Zone storage classes (Amazon EFS One Zone, and Amazon EFS One Zone-Infrequent Access (EFS One Zone-IA)).

Amazon EFS is well suited to support a broad spectrum of use cases from home directories to business-critical applications. Customers can use EFS to lift-and-shift existing enterprise applications to the AWS Cloud.

As more customers are embarking on application modernization using containers and serverless technologies, EFS enables customers to persist data and state from their container and serverless functions. EFS provides fully managed, elastic, highly-available, scalable, and high-performance, cloud-native shared file systems.

These same attributes are shared by Amazon Elastic Container Service (Amazon ECS), Amazon Kubernetes Service (Amazon EKS), AWS Fargate, and AWS Lambda, so developers don't need to design for these traits, the services are simply ready for modern application development with data persistence, using EFS.

Other use cases include big data analytics, web serving and content management, application development and testing, media and entertainment workflows, database backups, and container storage.

AWS Backup

AWS Backup is a fully managed backup service that makes it easy to centralize and automate the back up of data across AWS services in the cloud. AWS Backup provides a policy-based backup solution that simplifies your backup management and enables you to meet your business and regulatory backup compliance requirements. Using AWS Backup, you can backup and restore your Amazon EFS file systems natively.

Migration Guide

To get the most out of this whitepaper, we recommend that you have basic familiarity with the setup and use of each of these services: AWS DataSync, Amazon Elastic File

System, and AWS Backup. The following guide provides considerations for migrating an on-premises file system to AWS, including configuration options.

Prerequisites

To complete the steps outlined in this whitepaper, you need an active AWS account with the permissions required to create new resources for each of the mentioned services. Note that you are responsible for any charges incurred in creating and using these resources.

To transfer files from on-premises to AWS using DataSync, you need a virtualization system that is capable of running the DataSync agent, such as VMware ESXi, Microsoft Hyper-V Hypervisor or Linux Kernel-based Virtual Machine (KVM). You will also need access to a web browser to activate the DataSync agent. For more information, see the DataSync documentation.

Setup and Configuration

Amazon Elastic File System

Create an Amazon EFS file system using the AWS Management Console, Command Line Interface (CLI), or Software Development Kit (SDK). You can use this file system to store the data that you migrate from your on-premises environment. As a fully managed, elastic file system, there is no setup or provisioning needed and no need to plan for or manage storage capacity. EFS file systems automatically grow and shrink as you add and remove files.

General

Amazon EFS is designed to be highly available and highly durable. EFS offers a Service Level Agreement (SLA) for 4 nines of availability for Standard Storage classes, 3 nines for One Zone storage classes and is designed for 11 nines of data durability.

To achieve this level of availability and durability, EFS stores data redundantly - within a single Availability Zone (AZ) for One Zone storage classes, and within and across multiple AZs for Standard storage classes. EFS regional storage classes (EFS Standard and EFS Standard-IA) are designed to tolerate the unavailability of a single AZ without sacrificing performance or features.

Amazon EFS file systems can be accessed by thousands of clients simultaneously.

To access your file system, you create highly available resources called mount targets, one per AZ, and specify which VPC Security Groups are allowed to access them. The easiest way for your clients to access your file system is by using its DNS name. EFS file system DNS names automatically resolve to the IP address of the mount target located in the same AZ as the client making the request—optimizing performance and removing any potential cross-AZ network charges.

When creating a file system, you can select configuration options for Availability and Durability, Automatic backups, Lifecycle management, Performance mode, Throughput mode and Encryption. For data protection (and to satisfy your organization's compliance requirements), we recommend enabling encryption of data at rest, and mounting your file system using encryption of data in transit.

Enabling encryption of data at rest must be done when you create your file system. You can enable encryption in transit later, when you mount the file system. For more information refer to the whitepaper [Encrypting File Data with Amazon Elastic File System](#)

Performance

You also can choose a performance mode and a throughput mode for your file system. EFS offers two performance modes: General Purpose and Max I/O. Max I/O scales to higher amounts of aggregate throughput than General Purpose mode, but also has slightly higher metadata latencies than General Purpose mode for EFS file systems using Standard storage classes.

As with enabling encryption, you must select your file system performance mode when creating your file system. For the purposes of this whitepaper, we recommend selecting the default performance mode: General Purpose. Additional guidance is provided later in this whitepaper for evaluating which performance mode is optimal for your file system.

EFS offers two throughput modes: Bursting and Provisioned. In the default Bursting throughput mode, EFS file system performance scales linearly with the size of the file system's Standard storage class. When you create an Amazon EFS file system in Bursting Throughput mode, you get an allocation of 2.1 TB of burst credits.

All EFS file systems are able to burst up to 100 MB/s of write-only throughput or up to 300 MB/s read-only throughput (EFS meters read requests at one-third the rate of write requests) when using Bursting Throughput mode, with primary (Standard or One Zone) storage classes file systems larger than 1 TiB able to burst to twice their baseline throughput.

Consider using EFS Provisioned Throughput if you know how much throughput your applications require, and if it's higher than the amount that your storage provides. For more information, refer to the whitepaper [Amazon Elastic File System: Choosing Between the Different Throughput & Performance Modes](#).

In the next section, we will discuss AWS DataSync and how you can use it to quickly transfer data into EFS. If you're planning to migrate more than 2.1 TB of data to a new EFS file system, we recommend provisioning at least 100 MB/s of throughput which will prevent performance slowdowns if you run out of EFS burst credits.

If you must transfer data to an empty EFS file system at speeds greater than 100 MB/s, you will need to use Provisioned Throughput.

Storage Classes and Lifecycle Management

Amazon EFS offers four storage classes that are designed for different use cases. These include primary storage classes: EFS Standard, EFS One Zone and the infrequent access (IA) storage classes: EFS Standard–Infrequent Access (Standard-IA), EFS One Zone–Infrequent Access (EFS One Zone-IA).

The EFS IA storage classes are intended for files that are not accessed every day and stores them on slower storage media at a lower cost. You can start using EFS IA storage classes by enabling Lifecycle Management for your file system.

With Lifecycle Management enabled, EFS automatically moves files that have not been accessed according to the lifecycle policy from the EFS primary storage class to the corresponding IA storage class, which can reduce your storage costs by up to 92%.

Each EFS file system transparently serves files from EFS Standard and EFS Standard-IA or EFS One Zone and EFS One Zone-IA storage classes. We recommend enabling Lifecycle Management to automatically save cost as your file access patterns change and files are less frequently accessed.

AWS DataSync

To perform one-time data migrations or periodic synchronizations from an on-premises file system to Amazon EFS, use AWS DataSync. This whitepaper assumes that you have an AWS Direct Connect connection, but Direct Connect is not required to use DataSync. Using Direct Connect, you can transfer at speeds up to 10 Gbps per Datasync agent. At 10 Gbps sustained, you can transfer 1 terabyte of data from on-premises to the cloud in around 13 minutes, or a petabyte in around 9.5 days.

For data protection, AWS DataSync automatically encrypts data in transit and performs data integrity checks. DataSync uses simple, per-GB pricing. DataSync transfers are always incremental and DataSync has built-in compression and de-duplication to minimize the amount of network activity needed to get the job done.

To get started, you first download and activate the DataSync agent as a virtual machine in your on-premises environment. The DataSync agent handles the actual data transfer between your on-premises file system and EFS.

After you activate the agent, proceed to the DataSync management console to create a 'task', which is the complete definition of a data transfer that will handle the transfer between two locations: an on-premises Network File System (NFS) server and Amazon EFS.

See the AWS DataSync Getting Started Guide for detailed step-by-step instructions. We recommend using NFS version 4.1 (or later) for the best performance, but versions 4.0 and 3 of the NFS protocol are also supported.

DataSync provides a number of configuration options. First, make sure that you have set the proper network configuration, including firewall and VPC security group rules. You will need to enable traffic on port 2049 in your on-premises firewall and AWS VPC, and you will want to make sure that DataSync is configured to belong to at least one security group that is shared with your EFS mount targets.

If you need to transfer data without traversing the public internet, you can configure the VPC endpoints for DataSync and enable additional ports in your on-premises firewall as described in the network configuration for VPC endpoints. This will keep the DataSync traffic within your network and your AWS VPCs.

DataSync offers options on how to treat your file system metadata. By default, DataSync will preserve ownership information, permissions, and timestamps of your on-premises file system once it has been moved to EFS. We recommend keeping these default settings in place when performing your migration.

DataSync also allows you to turn off the automatic data verification step (the consistency check) it performs after the transfer is complete. Data verification can take a significant amount of time depending on how many files are in your file system and the rate of change to those files. If you are able to stop any file system activity against your on-premises file system before and for the duration of the transfer, you can disable data verification to save time.

DataSync provides a few options for increasing or decreasing file transfer bandwidth and performance. One option is to set a bandwidth limit for your transfer tasks–you can choose to optionally set this limit to enforce a quality of service policy to prevent DataSync from consuming too many network resources.

Alternatively, if you want DataSync to fully saturate the network and load data into EFS as quickly as possible, you can deploy and allocate as many as four DataSync agents to a given task. You may choose this option if you have millions of files in your source file system. You also can use multiple agents when you have network limitations on-premises (such as when not using Direct Connect) and want to get the most out of the limited network resources available.

After your DataSync task is set up, run it on-demand or periodically using the DataSync API to incrementally copy changes from your on-premises NFS share to EFS.

AWS Backup

Once you've copied your data to your EFS file system, you can back up your file system using AWS Backup for additional data protection. When you create a primary (Standard or One Zone) storage class EFS file system using the Amazon EFS Console or creating an One Zone storage class EFS file system using AWS CLI/API, automatic backups are turned on by default.

If using AWS CLI or Amazon EFS API to create a Standard storage class file system, you can turn on automatic backups while or after creating your file system. AWS Backup allows you to create scheduled backups to satisfy compliance requirements and recover from unintended events, such as the accidental deletion of files from your file system. Each backup is incremental—only changes to the file system are recorded.

With AWS Backup, you issue a backup by creating a resource called a backup plan. You can choose from some predefined backup plans that AWS Backup provides, or you can build a new one from scratch. An example of a predefined backup plan is one that executes daily and has a 35-day retention policy.

With your backup plan, you can select a frequency and schedule, or you can run it one time on-demand. Optionally, you can choose when your backup data migrates to a lower-cost 'cold' storage, or you can automatically purge older backups using an expiration policy. When using cold storage for backups, there is an additional retrieval fee if you need to restore it, and it takes a few extra hours to retrieve the backup from cold storage.

After you create your backup plan, you can choose which of your EFS file systems are governed by it. You can do this in one of two ways. One way is by using a tag-based mechanism: you can have your backup plan automatically back up any file system tagged with a user-defined string such as "Production". Tag-based backup plans will cover file systems created in the future that are tagged without requiring any changes to the backup plan. The other way to back up a file system is to specify it by its ID in the backup plan.

With AWS Backup, you create resources called backup vaults to organize and store your backups. When you create a backup vault, you select an AWS Key Management Service (KMS) key to encrypt your backups. When you back up an EFS file system, it is always stored encrypted.

You can easily restore your file system from backup using the AWS Backup console. The backup can be restored to a specific folder in an existing file system, or to a brand new file system.

Operations and Monitoring

Now that you've set up your EFS file system, your DataSync agent and task, and your backup plan, you are ready to move files from on-premises to the cloud and begin to run applications against EFS. AWS provides tools and metrics to help you evaluate and optimize performance on an ongoing basis, both from within each respective AWS service as well as through dedicated operational services such as Amazon CloudWatch.

AWS DataSync

DataSync provides status information about each task, task execution, and whether the locations that it is configured for (such as on-premises NFS servers and EFS file systems) are reachable. After you create a task, you can easily tell whether it's available to run, whether it's unavailable to run because one or more of the file systems are unreachable, or whether it's currently running. If the task is running, you can inspect the task execution status to see what phase of the workflow it is in, and ultimately if it succeeded or failed.

If you want to integrate DataSync operational health information into a custom monitoring solution, you can use the DataSync API to obtain a task's status or a task's execution status. You also can enable DataSync to send task logging information to an Amazon CloudWatch Log Group to monitor and debug your tasks.

DataSync also sends event and health metrics to Amazon CloudWatch. Using CloudWatch Events you can monitor and configure alarms for state changes in an agent, location, task, or task execution. Using CloudWatch metrics, as a function of time, you can see the number of files written, the number of bytes written, and the number of bytes transferred over the network as a function of time. For more information, see Monitoring Your Task.

Amazon EFS

As discussed previously, EFS has two performance modes: General Purpose and Max I/O. General Purpose mode is designed to be the right fit for most applications, but if you require the highest amounts of aggregate throughput possible, you might be better served by Max I/O mode.

To determine whether your workload is a good fit for Max I/O, we recommend that you first test using General Purpose mode and monitor the EFS PercentIOLimit CloudWatch metric. If you see this value consistently at or near 100%, then a Max I/O file system might be a better option for you. Another option would be to split your workload across multiple General Purpose mode file systems.

If you're not reaching the I/O limit of General Purpose mode, but still want to drive additional throughput, there are two techniques that you can pursue. First, to take maximum advantage of the distributed design of EFS, where storage is replicated within and across multiple Availability Zones, your application should be multi-threaded and leveraging parallel I/O as much as possible.

EFS provides a tutorial hosted on GitHub with several examples showing how to put this technique into practice, including an AWS CloudFormation stack to help you get up and running.

A second option is to determine if your throughput is being rate-limited based on your storage. You can use the Throughput utilization (%) graph in the Amazon EFS console Monitoring section to monitor your throughput utilization.

You can also monitor the throughput using MeteredIOBytes CloudWatch metric which represents the calculation of the overall metered throughput that you are driving on your file system and comparing it with the PermittedThroughput CloudWatch metric. You can do this using a CloudWatch metric math expression. if the values for MeteredIOBytes and PermittedThroughput are equal, your file system is consuming all available throughput.

This might be the case if you're running in EFS Bursting Throughput mode, have a relatively small file system, and have run out of burst credits. Each file system begins with 2.1 TB of burst credits, which accrue when I/O is idle but are consumed as I/O is performed. If this is the case and you have deterministic application throughput requirements, consider EFS Provisioned Throughput.

EFS provides CloudWatch metrics that allow you to monitor your throughput, burst credits, and client connections.

Pricing and Cost Considerations

AWS DataSync

AWS DataSync has simple, per-GB pricing. You pay $0.0125/GB or $12.80/TB transferred in all AWS Regions where DataSync is available.

Amazon EFS

Amazon EFS has six price dimensions: Standard Storage, Standard-Infrequent Access (IA) Storage, One Zone Storage, One Zone-Infrequent Access (IA) storage, Infrequent Access IO requests, and Provisioned Throughput. There are no minimum fees and you pay only for what you use.

Using the US-EAST-1 Region, EFS Standard storage costs $0.30/GB-month. EFS Standard IA storage costs $0.025/GB-month. EFS One Zone Storage costs $0.16/GB-month and EFS One Zone IA storage costs $0.0133/GB-month.

Using EFS Standard IA or One Zone IA, you also pay a fee each time you read a file from the IA storage class, and each time lifecycle management moves a file to the IA storage class. This fee is $0.01/GB transferred (US-EAST-1).

You also have the option of provisioning throughput for EFS at the rate of $6/MB/s (US-EAST-1). You only pay for the throughput in excess of what your file system is permitted to drive, based on primary (Standard or One Zone) storage classes, when in bursting throughput mode. As an example, if you have a file system with 10 TiB in the primary storage class, it is permitted to drive 500 MB/s of throughput continuously. If you provision 1 GB/s of throughput, you pay only for the additional 500 MB/s relative to what's permitted from your storage.

AWS Backup

Backing up an EFS file system costs $0.05/GB-month for warm storage and $0.01/GB-month for cold storage (US-EAST-1). Restores cost $0.02/GB from warm storage and $0.03/GB from cold storage.

Migration Cost Example

Here is an example for calculating migration costs. Ultimately, your workload and access patterns will drive your actual cost, so we've made simplifying assumptions based on generally accepted industry estimates. To test your own assumptions and analysis, refer to the AWS Pricing Calculator.

Let's say that you are migrating 100 TB of storage from an on-premises NFS server to AWS and storing it for 1 year, and 80% of this storage is infrequently accessed, aging off and never read again. You also want to keep a daily backup of your file system. Your yearly cost will be:

Component	Cost
AWS DataSync	$0.0125/GB * 1000 GB/TB * 100 TB = **$1,250**
Amazon EFS Standard Storage	$0.30/GB-month * 1000 GB/TB * 100 TB * 20% frequently accessed * 12 months/year = **$72,000**
Amazon EFS Standard IA I/O	$.01/GB * 1000 GB / TB * 100 TB * 80% infrequently accessed = **$800**
Amazon EFS Standard IA Storage	$0.025/GB-month * 1000 GB/TB * 100 TB * 80% infrequently accessed * 12 months/year = **$24,000**
AWS Backup	$0.05/GB-month * 1000 GB/TB * 100 TB * 12 months/year = **$60,000**

Total cost: **$157,250** or **$157.25/TB/year** to migrate to a highly available, highly durable, regional, elastic, fully managed file system–and keep an extra backup to protect against accidental data deletion.

For workloads that don't require multi-AZ resilience, One Zone storage classes store data redundantly within a single AZ, at a lower price compared to file systems using Standard storage classes. So, for same migration scenario described above, your yearly cost will be **$112,418** or **$112.42/TB/year**:

Component	Cost
AWS DataSync	$0.0125/GB * 1000 GB/TB * 100 TB = **$1,250**
Amazon EFS One Zone Storage	$0.16/GB-month * 1000 GB/TB * 100 TB * 20% frequently accessed * 12 months/year = **$38,400**
Amazon EFS One Zone IA I/O	$.01/GB * 1000 GB / TB * 100 TB * 80% infrequently accessed = **$800**
Amazon EFS One Zone IA Storage	$0.0133/GB-month * 1000 GB/TB * 100 TB * 80% infrequently accessed * 12 months/year = **$12,768**
AWS Backup	$0.05/GB-month * 1000 GB/TB * 100 TB * 12 months/year = **$60,000**

Conclusion

In summary, this whitepaper provides guidance on how to setup, tune, and monitor migrating applications that use file system storage to AWS. Using AWS DataSync to quickly and efficiently transfer files from an on-premises NFS file system to Amazon EFS and AWS Backup for automated data protection and recovery, you can take advantage of all the benefits of a public cloud service for around $157.25/TB/year on Standard Storage class or $112.42/TB/year on One Zone Storage class.

Contributors

Contributors to this document include:

- Joe Travaglini, Principal Product Manager, Amazon Elastic File System

- Peter Buonora, Principal Solutions Architect

- Siva Rajamani, Senior Solutions Architect

Document Revisions

Date	Description
May 2019	First publication
September 2019	Updated costs and reference to additional lifecycle policies
June 2021	Updated Amazon EFS Service features and cost

This paper has been archived

For the latest technical content, refer to the AWS Whitepapers & Guides page:
https://aws.amazon.com/whitepapers