AWS Storage Extras

AWS Snowball

- Highly-secure, portable devices to **collect and process data at the edge**, and **migrate data into and out of AWS**
- Helps migrate up to **Petabytes of data**

Device Types and Specifications

Device	Compute	Memory	Storage (SSD)
Snowball Edge **Storage** Optimized	104 vCPUs	416 GB	**210 TB**
Snowball Edge **Compute** Optimized	104 vCPUs	416 GB	**28 TB**

Data Migrations with Snowball

Challenges

- Limited connectivity
- Limited bandwidth
- High network cost
- Shared bandwidth (can't maximize the line)
- Connection stability

Time to Transfer

Data Size	100 Mbps	1 Gbps	10 Gbps
10 TB	12 days	30 hours	3 hours
100 TB	124 days	12 days	30 hours
1 PB	3 years	124 days	12 days

> ***AWS Snowball**: offline devices to perform data migrations*
> *If it takes more than a week to transfer over the network, use Snowball devices!*

Diagrams

Direct upload to S3

- The client uploads data directly to an Amazon S3 bucket over the internet.
- Assumes a network speed of **10 Gbit/s**.

client ──(www: 10 Gbit/s)──> Amazon S3 bucket

With Snowball

- The client writes data to an **AWS Snowball** device locally.
- The device is **shipped physically** to AWS.
- AWS Snowball then **imports/exports** the data into an Amazon S3 bucket.

client ──> AWS Snowball ──(ship)──> AWS Snowball ──> import/export ──> Amazon S3 bucket

> *This approach avoids long transfer times and network limitations by using physical transport.*

What is Edge Computing?

- Process data while it's being created on **an edge location**
    - Example: a truck on the road, a ship on the sea, a mining station underground
- These locations may have **limited internet** and **no access to computing power**
- We setup a **Snowball Edge** device to do edge computing:
    - **Snowball Edge Compute Optimized** (dedicated for edge computing) and Storage Optimized
    - Run **EC2 Instances** or **Lambda functions** at the edge
- **Use cases**: preprocess data, machine learning, transcoding media

Solution Architecture: Snowball into Glacier

- **Snowball cannot import to Glacier directly**
- You must use **Amazon S3 first**, in combination with an **S3 lifecycle policy**

Data Flow

Snowball ──(import)──> Amazon S3 ──(S3 lifecycle policy)──> Amazon Glacier

> *Use S3 as an intermediate storage tier when moving Snowball data to Glacier.*

Amazon FSx – Overview

- **Launch 3rd party high-performance file systems on AWS**
- Fully managed service

Supported File Systems

- **FSx for Lustre**
- **FSx for Windows File Server**
- **FSx for NetApp ONTAP**
- **FSx for OpenZFS**

Amazon FSx for Windows (File Server)

- **FSx for Windows** is a fully managed **Windows** file system share drive
- Supports **SMB protocol** & **Windows NTFS**
- **Microsoft Active Directory** integration, ACLs, user quotas
- **Can be mounted on Linux EC2 instances**
- Supports **Microsoft's Distributed File System (DFS) Namespaces** (group files across multiple file systems)

Scalability and Storage

- Scale up to **tens of GB/s**, **millions of IOPS**, **hundreds of PB of data**
- **Storage Options**:
    - **SSD** – for latency-sensitive workloads (databases, media processing, data analytics, etc.)
    - **HDD** – for broader use cases (home directories, CMS, etc.)

Connectivity and Availability

- Can be accessed from on-premises (via **VPN** or **Direct Connect**)
- Can be configured to be **Multi-AZ** (for high availability)
- **Data is backed-up daily to S3**

Amazon FSx for Lustre

- **Lustre** is a type of parallel distributed file system, designed for **large-scale computing**
- The name **Lustre** comes from "**Linux**" and "**cluster**"

Use Cases

- Machine Learning, **High Performance Computing (HPC)**
- Video Processing, Financial Modeling, Electronic Design Automation

Performance

- Scales up to **100s of GB/s**, **millions of IOPS**, **sub-millisecond latencies**

Storage Options

- **SSD** – low-latency, IOPS-intensive workloads, small & random file operations
- **HDD** – throughput-intensive workloads, large & sequential file operations

S3 Integration

- Can **"read S3"** as a file system (via FSx)
- Can **write outputs** back to S3 (via FSx)
- **Can be used from on-premises servers** (via **VPN or Direct Connect**)

FSx Lustre – File System Deployment Options

Scratch File System

- **Temporary storage**
- Data is **not replicated** (does not persist if file server fails)
- **High burst** performance (6× faster, 200 MBps per TiB)
- **Usage**: short-term processing, cost optimization

Persistent File System

- **Long-term storage**
- Data is **replicated within the same AZ**
- Can **replace failed files within minutes**
- **Usage**: long-term processing, sensitive data

Architecture Diagram (Summary)

Scratch File System

Region └── Availability Zone 1 ├── Compute instances ├── FSx For Lustre (Scratch) └── ENI └── Availability Zone 2 └── Compute instances → Optional S3 bucket as data repository

Persistent File System

Same as above, but FSx is configured as Persistent → Data is replicated → Higher durability for sensitive or long-lived data

Amazon FSx for NetApp ONTAP

- **Managed NetApp ONTAP** on AWS
- **File system compatible** with **NFS**, **SMB**, and **iSCSI** protocols
- Allows you to **move workloads from ONTAP or NAS to AWS**

Works With

- Linux
- Windows
- macOS
- VMware Cloud on AWS
- Amazon Workspaces & AppStream 2.0
- Amazon EC2, ECS, and EKS

Features

- **Elastic storage**: automatically grows or shrinks
- **Snapshots**, **replication**, **low-cost**, **compression**, and **data de-duplication**
- **Point-in-time instantaneous cloning** (useful for testing new workloads)

Integration Protocols

- **NFS**, **SMB**, **iSCSI** supported for connectivity

Amazon FSx for OpenZFS

- **Managed OpenZFS file system** on AWS
- File system compatible with **NFS (v3, v4, v4.1, v4.2)**
- Enables migration of workloads running on **ZFS to AWS**

Works With

- Linux
- Windows
- macOS
- VMware Cloud on AWS
- Amazon Workspaces & AppStream 2.0
- Amazon EC2, ECS, and EKS

Performance and Features

- Up to **1,000,000 IOPS** with **< 0.5ms latency**
- **Snapshots**, **compression**, and **low-cost**
- **Point-in-time instantaneous cloning** (useful for testing new workloads)

Hybrid Cloud for Storage

- AWS is promoting the **"hybrid cloud"** model:
  - Part of your infrastructure is **on the cloud**
  - Part of your infrastructure is **on-premises**

Reasons for Hybrid Cloud

- Long cloud migrations
- Security requirements
- Compliance requirements

- IT strategy

Challenge with S3

- **S3** is a proprietary storage technology (unlike EFS/NFS)
- So, how do you expose S3 data **on-premises**?

> *Use **AWS Storage Gateway**!*

AWS Storage Cloud Native Options

Block Storage

- **Amazon EBS**
- **EC2 Instance Store**

File Storage

- **Amazon EFS**
- **Amazon FSx**

Object Storage

- **Amazon S3**
- **Amazon Glacier**

AWS Storage Gateway

- **Bridge** between **on-premises data** and **cloud data**

Use Cases

- Disaster recovery
- Backup & restore
- Tiered storage
- On-premises cache & low-latency file access

Types of Storage Gateway

- **S3 File Gateway**
- **FSx File Gateway**
- **Volume Gateway**
- **Tape Gateway**

Amazon S3 File Gateway

- Configured **S3 buckets are accessible via NFS and SMB protocols**
- **Most recently used data** is **cached** in the file gateway
- Supports:
  - **S3 Standard**
  - **S3 Standard-IA**
  - **S3 One Zone-IA**
  - **S3 Intelligent-Tiering**
- Can **transition data to S3 Glacier** using a **Lifecycle Policy**
- Access is controlled using **IAM roles** for each File Gateway
- **SMB protocol** supports integration with **Active Directory (AD)** for user authentication

Data Flow Diagram (Summary)

Application Server (NFS/SMB) ↔ S3 File Gateway (cache) ↔ HTTPS ↔ S3 (Standard, IA, One Zone, Intelligent-Tiering) ↠ [Lifecycle policy] → S3 Glacier

Amazon FSx File Gateway

- **Native access** to Amazon FSx for **Windows File Server**
- **Local cache** for **frequently accessed data**
- Full **Windows native compatibility**:
  - **SMB**, **NTFS**, **Active Directory**, etc.
- Useful for **group file shares** and **home directories**

Architecture (Summary)

SMB Clients ↔ Amazon FSx File Gateway ↔ Amazon FSx for Windows File Server (AWS Cloud)

Volume Gateway

- Provides **block storage** using the **iSCSI protocol**, backed by **Amazon S3**
- Backed by **EBS snapshots**, which allow restoring on-premises volumes

Volume Types

- **Cached volumes**:
  - Low-latency access to the **most recently used data**
  - Full dataset is stored in the cloud
- **Stored volumes**:
  - The **entire dataset** is stored **on-premises**
  - Scheduled **backups** are sent to **S3**

Architecture (Summary)

Application Server ↔ (iSCSI) ↔ Volume Gateway ↔ (HTTPS) ↔ S3 Bucket ↔ EBS Snapshots

Tape Gateway

- Some companies have backup processes using **physical tapes (!)**
- With **Tape Gateway**, companies can use the same processes **in the cloud**
- Provides a **Virtual Tape Library (VTL)** backed by **Amazon S3** and **Amazon Glacier**
- Supports backups using **existing tape-based processes** via **iSCSI interface**
- Compatible with **leading backup software vendors**

Architecture (Explanation)

- A **Backup Server** uses the **iSCSI protocol** to communicate with a **Tape Gateway**
- The Tape Gateway emulates a **Media Changer** and **Tape Drive**
- Data is transferred over **HTTPS** to AWS
- Tapes are stored as **Virtual Tapes** in **Amazon S3**
- Archived tapes can be moved to **Amazon Glacier** for long-term storage

Storage Gateway – Hardware Appliance

- Using Storage Gateway typically requires **on-premises virtualization**
- Alternatively, you can use a **Storage Gateway Hardware Appliance**
- It is **available for purchase on amazon.com**

Key Points

- Works with **File Gateway**, **Volume Gateway**, and **Tape Gateway**
- Comes with required:
  - **CPU**
  - **Memory**
  - **Network**
  - **SSD cache resources**
- Useful for **daily NFS backups** in **small data centers**

Supported Host Platforms

- VMware ESXi
- Microsoft Hyper-V 2012R2/2016
- Linux KVM
- Amazon EC2
- **Hardware Appliance** (prebuilt)

AWS Storage Gateway

On-Premises Integration

- **File Gateway** (local cache)
  - Access via **NFS/SMB**
  - Used for user/group file shares
- **Volume Gateway** (local cache)
  - Access via **iSCSI**
  - Used by application servers
- **Tape Gateway** (local cache)
  - Access via **iSCSI VTL**
  - Used by backup applications

> *All data transfers occur over **encrypted connections** (Internet or Direct Connect)*

Cloud Integration

- **Amazon S3** (excluding Glacier & Glacier Deep Archive)
  - Primary destination for file and volume data
  - Can be transitioned to **any S3 Storage Class**, including Glacier
- **Amazon S3 + AWS EBS**
  - Used for snapshots from **Volume Gateway**
- **Amazon S3 (Tape Library) → Tape Archive**
  - Tapes can be ejected to Glacier & Glacier Deep Archive from backup applications

- **Amazon FSx for Windows File Server**
    - Supports automated backups to **Amazon S3**

Deployment Options

- Can be deployed as:
    - VM (**VMware**, **Hyper-V**, **KVM**)
    - **Hardware Appliance**

AWS Transfer Family

- A **fully-managed service** for file transfers into and out of **Amazon S3** or **Amazon EFS** using the **FTP protocol**

Supported Protocols

- **AWS Transfer for FTP** (File Transfer Protocol)
- **AWS Transfer for FTPS** (File Transfer Protocol over SSL)
- **AWS Transfer for SFTP** (Secure File Transfer Protocol)

Features

- Managed infrastructure: **scalable**, **reliable**, **highly available** (multi-AZ)
- **Pay per provisioned endpoint** per hour, plus data transfer cost (per GB)
- **Stores and manages users' credentials**
- Can integrate with external authentication systems:
    - Microsoft Active Directory
    - LDAP
    - Okta
    - Amazon Cognito
    - Custom identity systems

Use Cases

- File sharing
- Public dataset hosting
- CRM, ERP system integration

AWS Transfer Family – Architecture Overview

Flow Summary

- **Users (FTP clients)** connect to the service via **Route 53** (optional)
- They access one of the following protocols provided by the **AWS Transfer Family**:
    - **AWS Transfer for SFTP**
    - **AWS Transfer for FTPS**
    - **AWS Transfer for FTP** (only within **VPC**)
- Authentication is handled via:
    - **Microsoft Active Directory**
    - **LDAP**
    - Other identity providers

- Once authenticated, users are granted access to **Amazon S3** or **Amazon EFS**
    - Access is controlled via **IAM Roles**

Key Concepts

- Supports integration with enterprise identity systems
- Can transfer files directly into **Amazon S3** or **Amazon EFS**
- Fully managed and supports high availability

AWS DataSync

- Used to **move large amounts of data** to and from:
    - **On-premises / other cloud to AWS** (via NFS, SMB, HDFS, S3 API, etc.) → **requires agent**
    - **AWS to AWS** (between different storage services) → **no agent needed**

Supported Targets

- **Amazon S3** (supports any storage class, including Glacier)
- **Amazon EFS**
- **Amazon FSx** (Windows, Lustre, NetApp, OpenZFS...)

Features

- **Replication tasks** can be scheduled: **hourly**, **daily**, or **weekly**
- **File permissions and metadata are preserved**
    - Supports NFS POSIX, SMB, etc.
- **One agent task can use up to 10 Gbps** of bandwidth
    - Bandwidth limits can be configured

AWS DataSync – NFS / SMB to AWS (S3, EFS, FSx...)

Architecture Overview

On-Premises

- **NFS or SMB Server** connects to
- **AWS DataSync Agent** (can run on-premises or on a device like **AWS Snowcone**)
- Communication uses **TLS encryption**

AWS Cloud

- The agent connects to the **AWS DataSync service**, which transfers data to:

AWS Storage Resources (Targets)

- **Amazon S3**:
    - S3 Standard
    - S3 Intelligent-Tiering
    - S3 Standard-IA
    - S3 One Zone-IA
    - S3 Glacier
    - S3 Glacier Deep Archive
- **Amazon EFS**
- **Amazon FSx**

AWS DataSync – Transfer Between AWS Storage Services

- AWS DataSync can be used to **copy data and metadata** between AWS storage services.

Supported Sources and Destinations

- **Amazon S3**
- **Amazon EFS**
- **Amazon FSx**
- Transfers can be done **between any combination** of these services:
  - S3 → S3
  - S3 → EFS
  - FSx → EFS
  - EFS → FSx
  - etc.

Notes

- **No agent is needed** for transfers between AWS services.
- Supports **data and metadata preservation**.

Storage Comparison

- **S3**: Object Storage
- **S3 Glacier**: Object Archival
- **EBS volumes**: Network storage for one EC2 instance at a time
- **Instance Storage**: Physical storage for your EC2 instance (high IOPS)
- **EFS**: Network File System for Linux instances, POSIX filesystem
- **FSx for Windows**: Network File System for Windows servers
- **FSx for Lustre**: High Performance Computing Linux file system
- **FSx for NetApp ONTAP**: High OS Compatibility
- **FSx for OpenZFS**: Managed ZFS file system
- **Storage Gateway**: S3 & FSx File Gateway, Volume Gateway (cache & stored), Tape Gateway
- **Transfer Family**: FTP, FTPS, SFTP interface on top of Amazon S3 or Amazon EFS
- **DataSync**: Schedule data sync from on-premises to AWS, or AWS to AWS
- **Snowcone / Snowball / Snowmobile**: To move large amounts of data to the cloud, physically
- **Database**: For specific workloads, usually with indexing and querying