Technical Report

AWS Cloud Volumes ONTAP best practices for EDA workloads

Optimizing NetApp Cloud Volumes ONTAP for EDA and semiconductor workloads on AWS

Michael Johnson, NetApp

November 2022 | TR-4944

Abstract

This technical report examines the setup and performance of NetApp® Cloud Volumes ONTAP® for Amazon Web Services (AWS) for electronic design automation (EDA) workloads. NetApp partners, customers, and employees should use the presented information to make informed decisions about which workloads are appropriate for Cloud Volumes ONTAP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

This document is intended to capture best practices for using NetApp Cloud Volumes ONTAP (CVO) for Electronic Design Automation (EDA) and semiconductor workloads in AWS. The document includes a short background on typical EDA storage use cases and requirements.

EDA and semiconductor workloads are high-performance computing (HPC) workloads that can see up to 100k or more compute cores running parallel jobs on a single project design tree. Example workloads includes frontend verification simulations running Synopsys VCS, Cadence Xcellium, Siemens EDA ModelSim simulators, or chip-timing-analysis jobs like Synopsys Primetime or Fusion Compiler. There are often as many as 25 or more EDA design tools used between design specification, design capture, constrained random simulation, synthesis, Place and Route (P&R), Power Performance Area (PPA) optimization, floor planning, and chip finishing.

Advanced manufacturing technologies in chip design are constantly demanding increased compute and storage for EDA workloads. AWS, with its virtually infinite capacity and wide selection of compute and storage options, is a natural fit for EDA. EDA vendors like Synopsys, Cadence, and Siemens EDA are optimizing their EDA workflows to take advantage of cloud and AI techniques. They use massive parallel computing resources to speed up job completion time, reducing time to results and time to bug identification and fixes.

EDA customers typically use the highest performance NetApp ONTAP All-Flash (SSD) systems in large 16-24 node clusters for their EDA workloads. EDA workloads often cause storage controller bottlenecks due to the high metadata loads and highly parallelized access. On-premises, NetApp recommends using FlexGroup volumes spanning multiple ONTAP nodes with compute servers load balanced across multiple data LIFs.

EDA best practices for Cloud Volumes ONTAP in AWS focus on maximizing and optimizing all available options for maximum scale and throughput. NetApp highly recommends using the maximum settings for throughput and IOPS for active EDA workloads.

Note: This document only covers tier-1 (highest performance) configurations and does not attempt to provide best practice for tier-2 or backup tiering configurations.

Reference architecture for EDA in a hybrid cloud

EDA storage performance requirements

Some of our largest customers use more than 100k parallel compute cores. High parallelism, high file count, high metadata workloads make optimizing for scale-out performance critical. There is no single EDA workload; 50 or more different tools are used in the semiconductor development process from design specification, to design, timing, power and area optimization, and chip finishing.

NetApp has been a data management and storage leader in the semiconductor industry from its earliest days. On-premises, NetApp customers use NetApp A700s and A800 class all-flash storage appliances running the latest ONTAP storage operating system. Over the years, customer workloads have demonstrated that EDA workloads benefit from all-flash SSD-based systems with large CPU and memory configurations like those found in the A700s and A800 systems. Customer workloads have also shown that the scale-out performance of FlexGroup volumes can dramatically improve the number of parallel jobs design teams can run, which in turn leads to a faster TTT of EDA jobs.

The cloud enables EDA design flows to scale compute with almost no limit using the latest compute cores available. As such, the underlying storage and data management system must be designed to support cloud-scale parallel workloads.

The recommendations in this document are focused on configuring Cloud Volumes ONTAP for maximum parallel compute performance while maintaining 1-3ms latency or better. Cost optimization is a secondary requirement after scale-out performance. After you have established your performance requirements, then you can address performance-vs-cost trade-offs.

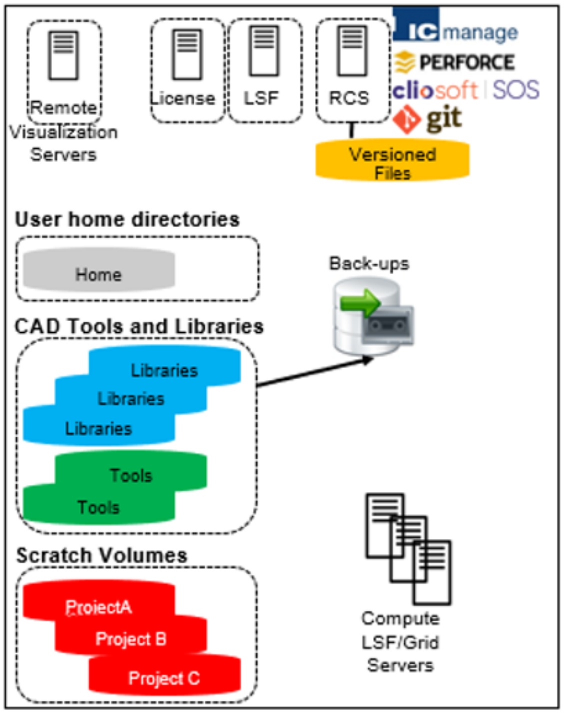
EDA volumes and directories

EDA workflows mount and use data from many different volumes in a single workflow. Not all data volumes have the same characteristics and knowing how each are used in an EDA flow can provide opportunities for optimizing runtime performance and utilization of each of these data infrastructure resources.

For example, `/mnt/tools`, `/mnt/libs`, and `/mnt/libraries` are common storage-volume names representing the location of the EDA tools and design libraries used in a design flow. Within these volume(s), there are often dozens to hundreds of tools and tool versions installed. Tools and libraries are often installed and centrally managed by CAD teams and then shared by multiple teams or projects.

IT and storage teams typically provision storage based on the needs of business units, teams, projects, and specific design flows. Each company's design flows are unique (but not that unique), so we can generalize data and data management for the purposes of classifying data types and requirements.

Figure 1) Typical on-premises design infrastructure.



The on-premises design infrastructure depicted in the previous figure typically includes the following components.

- Compute LSF/grid servers
  - Bare-metal Linux servers with uniform configurations.
  - Optimized for compute performance.
  - Various core counts and memory sizes.
- Load sharing (LSF) – manages job queuing.
  - Provides optimal utilization of resources and fair share.
  - IBM Spectrum LSF, Univa Grid Engine (Sun Grid Engine), SLURM.
- FlexLM license servers
  - Most common EDA license manager.
  - EDA jobs check-out licenses at runtime. After all licenses are in use, additional jobs either wait or fail.
  - After payroll expenses, EDA tool licenses are the next most expensive component of chip design. EDA licenses can run from $5k-$100k per license depending on the tool.

- Revision Control System (RCS)
  - Perforce and ICManage are very popular.
  - Clio SOS, Synchronicity Design Sync, and IBM Clearcase are also commonly used.
  - Git or Artifactory class source and artifact management are less common.
- Data logging, databases, data mining, and AI
  - Tools like Splunk, ElasticSearch, and so on are used for log-file mining and reporting.
  - Databases are used for job and metrics collection.
  - There has been an increased use of AI for design and design flow optimization.
- Remote desktop visualization servers
  - Hummingbird, VNC, terminal services

EDA storage integration employs the following components:

- Read-only tools and library volumes
  - EDA tools (Synopsys, Cadence, Siemens, Ansys, and so on)
  - Technology libraries (TSMC, Global Foundry, and so on)
  - Design IP (ARM, Synopsys, Cadence, and so on)
  - Read-only, except when new tools or libraries are installed.
  - Caching or volume replication with DNS mount load balancing can improve read performance.
- Revision Control System (RCS)
  - Perforce, ICManage, Cliosoft SOS, IBM ClearCase, or Synchronicity.
  - Like Git, but good for very large and small files, text, and binary.
  - Combination of local attached SSD (or SAN) for dbase, logging, and journaling NFS for versioned file store.
  - Very large volumes of both design files and build artifacts. Can grow to PB of data.
  - Long code check-out times can be a day-to-day challenge as is multisite deployments and replication. As build artifacts get larger, so do data transfer times.
- Build artifacts
  - Build artifacts are reusable outputs of the EDA and software build process.
  - Netlists, design databases, compile images, generated file lists, log files, and reports might all be considered build artifacts.
  - Management of build artifacts is handled differently by different teams or companies. Some create versions and redistribute artifacts using tools like Perforce or ICManage. Others store them in a file system.
- Release build volumes
  - Release candidates for the final chip output are generated here. This can be one or many volumes. Typically managed by design leads or central build teams. Often automated by tools like Jenkins.
  - All or many of the design tools are run against a single copy of the source files.
  - Data management is critical: DR, backup, HA, and so on.
- Nightly or CI builds
  - These are regular automated builds typically managed by a design lead or central build teams. These jobs are often automated with CI tools like Jenkins, Bamboo, or CircleCI.
  - These builds use either scratch volumes or normal protected volumes.
- User workspaces
  - User workspaces are almost always provisioned on scratch volumes. This is where developer do most of their work creating and editing designs, writing and running tests, and iterating to improve design quality and functionality.
  - IT often provisions a large volume that multiple users share, often with individual quotas to make sure that one user does not fill up the whole volume.

Note: Users are typically expected to check available storage capacity before launching new jobs. User level and management reporting is key.

The following figure depicts the storage and directory organization of an EDA flow. Many volumes are used in an EDA flow, each with different I/O and data management characteristics.

Figure 2) EDA storage and directory organization.

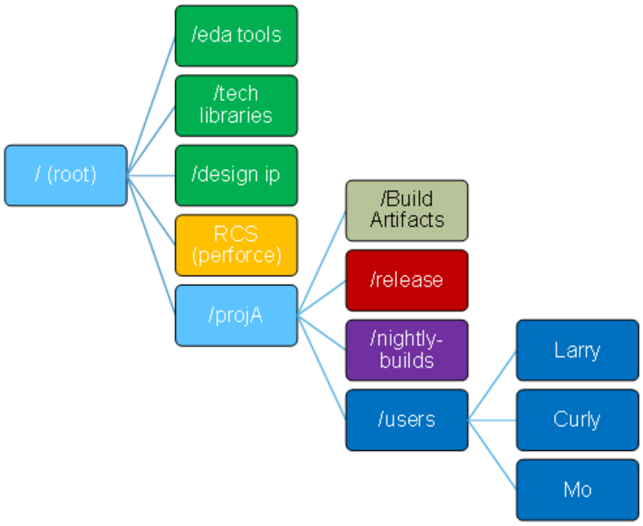
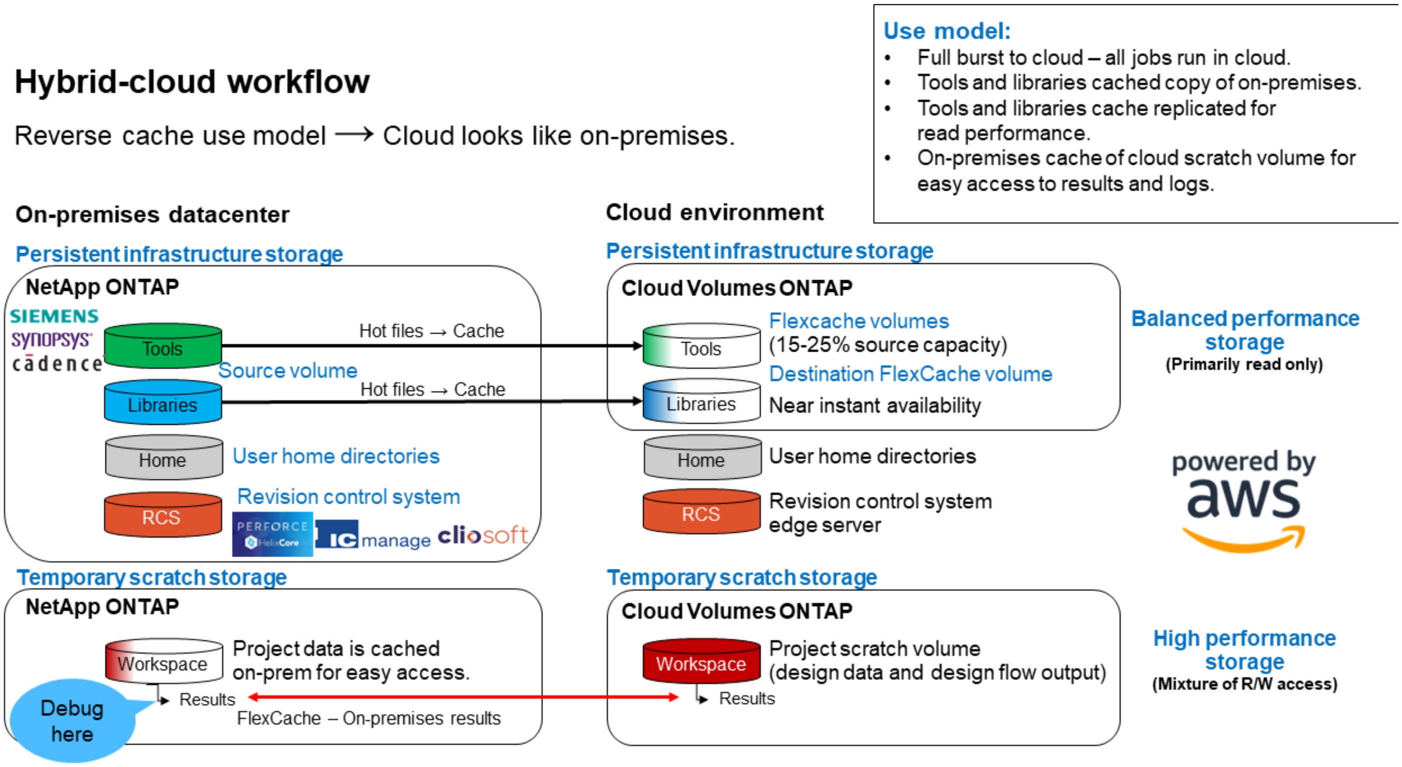


Figure 3) Hybrid cloud and burst-to-cloud architectures.



Use model:

- Full burst to cloud – all jobs run in cloud.
- Tools and libraries cached copy of on-premises.
- Tools and libraries cache replicated for read performance.
- On-premises cache of cloud scratch volume for easy access to results and logs.

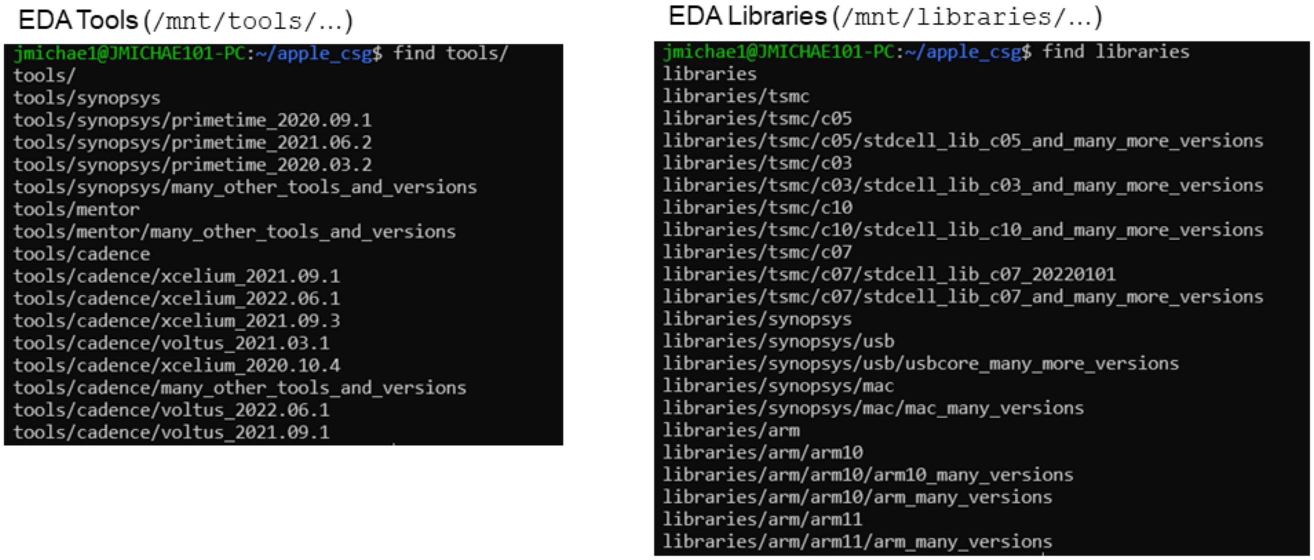
EDA data types

Tool and library volumes – Optimized for read performance

EDA tools like those from Synopsys, Cadence, and Siemens EDA are typically installed in tool directories mounted like `/mnt/tools/synopsys`, `/mnt/tools/cadence`, and so on. Each of these installation directories can contain many different tools and many different tool versions.

Similarly, third-party design-technology libraries are typically installed in directories mounted like `/mnt/libs/tsmc`, `/mnt/libs/arm`, `/mnt/libs/dolphin`, and so on, as can be seen in the following figure. These libraries contain either design IP (like a USB or CPU core) or they can contain tech libraries from device manufactures like TSMC, Intel, Global Foundries, UMC, or Samsung, which describe the physical power, area, and performance data-associated transistor/cell representation of the process node. Tech libraries typically contain thousands of files containing an ever-increasing amount of data.

Figure 4) Typical EDA Tools and Library UNIX directory structure.



EDA Tools (/mnt/tools/...)

```
jmichae1@JMICHAE101-PC:~/apple_csg$ find tools/
tools/
tools/synopsys
tools/synopsys/primetime_2020.09.1
tools/synopsys/primetime_2021.06.2
tools/synopsys/primetime_2020.03.2
tools/synopsys/many_other_tools_and_versions
tools/mentor
tools/mentor/many_other_tools_and_versions
tools/cadence
tools/cadence/xcelium_2021.09.1
tools/cadence/xcelium_2022.06.1
tools/cadence/xcelium_2021.09.3
tools/cadence/voltus_2021.03.1
tools/cadence/xcelium_2020.10.4
tools/cadence/many_other_tools_and_versions
tools/cadence/voltus_2022.06.1
tools/cadence/voltus_2021.09.1
```

EDA Libraries (/mnt/libraries/...)

```
jmichae1@JMICHAE101-PC:~/apple_csg$ find libraries
libraries
libraries/tsmc
libraries/tsmc/c05
libraries/tsmc/c05/stdcell_lib_c05_and_many_more_versions
libraries/tsmc/c03
libraries/tsmc/c03/stdcell_lib_c03_and_many_more_versions
libraries/tsmc/c10
libraries/tsmc/c10/stdcell_lib_c10_and_many_more_versions
libraries/tsmc/c07
libraries/tsmc/c07/stdcell_lib_c07_20220101
libraries/tsmc/c07/stdcell_lib_c07_and_many_more_versions
libraries/synopsys
libraries/synopsys/usb
libraries/synopsys/usb/usbcore_many_more_versions
libraries/synopsys/mac
libraries/synopsys/mac/mac_many_versions
libraries/arm
libraries/arm/arm10
libraries/arm/arm10/arm10_many_versions
libraries/arm/arm10/arm_many_versions
libraries/arm/arm11
libraries/arm/arm11/arm_many_versions
```

Internal libraries or design IP are another class of data used in designs. These might be subsystems or blocks developed by one team within the company and delivered to another team for inclusion in a full chip design. For instance, an ARM core designed by the processor team might be delivered and used by the team that integrates the core into a mobile device.

The key characteristic of tools and libraries is that the data is typically read-only or read-mostly. Writes occur when new versions of tools or libraries are installed, but tools and libraries are seldomly over-written. New tool or library versions are typically installed next to prior versions. As such, we can take advantage of this by optimizing the infrastructure to optimize for read-only performance. This might include NFS mount options to maximize compute server data caching to improve I/O latency.

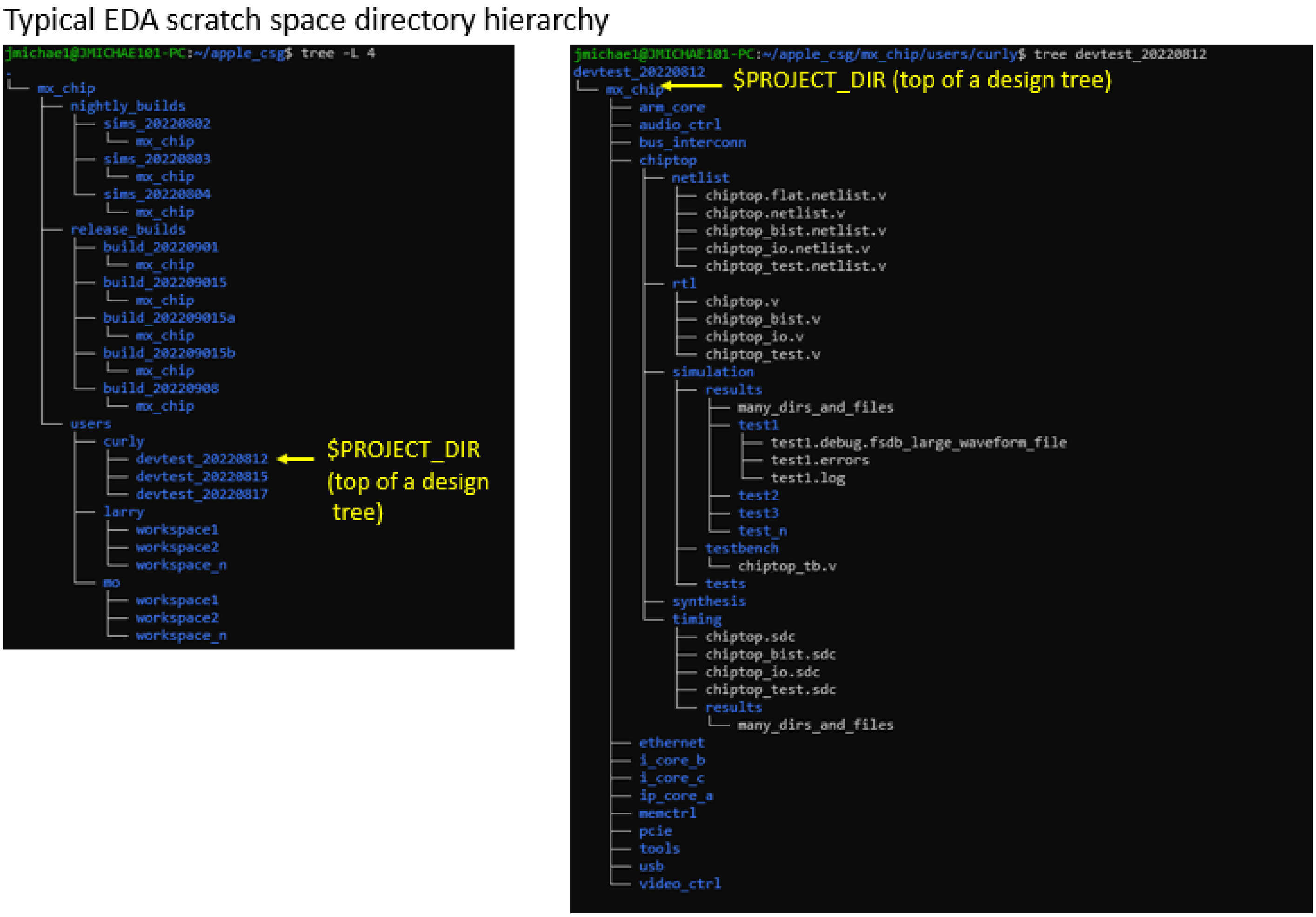
Tools and libraries are accessed by the EDA design flows and might see very high concurrent accesses from the 10k-100k parallel jobs simultaneously running in the compute farm. As such, I/O performance and scale is critical. Tools like Synopsys PrimeTime, a timing analysis tool, load thousands of library files simultaneously from thousands of servers, resulting in high read loads that might cause performance bottlenecks. The reference architecture proposed in this document addresses how FlexCache can improve read performance by load balancing NFS mounts across multiple FlexCached tool and library volumes.

Scratch volumes – Optimized for maximum read/write performance

Scratch volumes are used for user workspace, nightly regressions, CI/CD build flows, and many other development purposes. The term scratch comes from the idea that the data is temporary or has such a high change rate that traditional back-up and archiving is unnecessary.

Scratch is where active development is happening. Engineers launch EDA jobs and wait for them to complete so they can check whether a test passed or failed after a design change or a new functional test. They might be waiting for the results of thousands of jobs to check the power performance and area (PPA) characteristics of a chip after a functional change or optimization. Performance is critical, and both capacity and compute scale are critical attributes.

Figure 5) Sample EDA project UNIX directory structure.



Typical EDA scratch space directory hierarchy

```
jmichae1@JMICHAE101-PC:~/apple_csg$ tree -L 4
.
└── mx_chip
    ├── nightly_builds
    │   ├── sims_20220802
    │   │   └── mx_chip
    │   ├── sims_20220803
    │   │   └── mx_chip
    │   └── sims_20220804
    │       └── mx_chip
    ├── release_builds
    │   ├── build_20220901
    │   │   └── mx_chip
    │   ├── build_202209015
    │   │   └── mx_chip
    │   ├── build_202209015a
    │   │   └── mx_chip
    │   ├── build_202209015b
    │   │   └── mx_chip
    │   └── build_20220908
    │       └── mx_chip
    └── users
        ├── curly
        │   ├── devtest_20220812
        │   ├── devtest_20220815
        │   └── devtest_20220817
        ├── larry
        │   ├── workspace1
        │   ├── workspace2
        │   └── workspace_n
        └── mo
            ├── workspace1
            ├── workspace2
            └── workspace_n
```

$PROJECT_DIR (top of a design tree)

```
jmichae1@JMICHAE101-PC:~/apple_csg/mx_chip/users/curly$ tree devtest_20220812
devtest_20220812
└── mx_chip
    ├── arm_core
    ├── audio_ctrl
    ├── bus_interconn
    ├── chiptop
    │   ├── netlist
    │   │   ├── chiptop.flat.netlist.v
    │   │   ├── chiptop.netlist.v
    │   │   ├── chiptop_bist.netlist.v
    │   │   ├── chiptop_io.netlist.v
    │   │   └── chiptop_test.netlist.v
    │   ├── rtl
    │   │   ├── chiptop.v
    │   │   ├── chiptop_bist.v
    │   │   ├── chiptop_io.v
    │   │   └── chiptop_test.v
    │   ├── simulation
    │   │   ├── results
    │   │   │   ├── many_dirs_and_files
    │   │   │   ├── test1
    │   │   │   │   ├── test1.debug.fsdb_large_waveform_file
    │   │   │   │   ├── test1.errors
    │   │   │   │   └── test1.log
    │   │   │   ├── test2
    │   │   │   ├── test3
    │   │   │   └── test_n
    │   │   ├── testbench
    │   │   │   └── chiptop_tb.v
    │   │   └── tests
    │   ├── synthesis
    │   └── timing
    │       ├── chiptop.sdc
    │       ├── chiptop_bist.sdc
    │       ├── chiptop_io.sdc
    │       ├── chiptop_test.sdc
    │       └── results
    │           └── many_dirs_and_files
    ├── ethernet
    ├── i_core_b
    ├── i_core_c
    ├── ip_core_a
    ├── memctrl
    ├── pcie
    ├── tools
    ├── usb
    └── video_ctrl
```

$PROJECT_DIR (top of a design tree)

Scratch volumes should be optimized for maximum read/write performance and throughput. Modern chip design flows can use as many as one-hundred different tools with a wide range of I/O profiles (read to write ratios, big versus small files, or sequential versus random access) and might vary dramatically. Because these environments are often shared by multiple flows or teams, optimizing for one workflow might not be possible.

Since workloads running on scratch volumes are highly iterative, they are often much more fault tolerant. If a server dies, or if there is high storage latency that causes the job to hang, the job can in many cases just be restarted. Data loss or job restarts are not ideal but are typically not fatal. The worst issue is the time lost due to restarting the job. Some jobs take minutes, others take hours or days to complete. Tolerance to infrastructure issues vary by workflows.

The following EDA best practices provide guidance on how to balance storage performance against tolerance to faults in the infrastructure.

Release or preproduction release volumes – Optimized for maximum durability

Another general class of volumes is release volumes or pre-release volumes. As the chip design approaches design closure and tape-out (also known as release to manufacturing), builds of the chip are performed that are considered release candidates. If the release candidate passes all the quality checks as it passes thru the complete chip design flow, the design is readied for release.

These work areas might be considered more critical and less fault tolerant to infrastructure failures.

Build artifacts

Build artifacts can be any number of generated data from the EDA or software development workflow. A chip or IP netlist generated from Register Transfer Level (RTL) code using synthesis tools is a build artifact. A generated gate-level netlist might be reused by the verification team for gate-level simulations. It might also be reused by the emulation team or by the timing analysis and floor planning team to name a few. Modern semiconductors are made up of both soft, firm, and hard IP that is developed by one team and delivered to another team for integration into a design. The artifacts of an IP or subsystem development team might become reused IP deliverables provided to another team for integration.

Companies manage and share build artifacts differently. Some teams use tools like Perforce/ICManage to check-in and version build artifacts and then use Perforce Helix replication software or edge servers to distribute IP to teams and development sites. Other companies distribute build artifacts by replicating data volumes or directories. Others use symbolic links to share files and directories into development areas. Many NetApp customers have used data replication tools like SnapMirror or FlexCache to distribute artifacts across multiple sites.

EDA performance and cloud-scale

Single job performance measured by wall-clock time

EDA job performance is typically measured in wall-clock time. How long did it take to run a single chip-design simulation? How long did it take to complete a timing analysis or floor plan?

An improvement job runtime is critical since it improves the productivity and utilization of design engineers and EDA licenses. Engineers and EDA licenses make up over 90% of chip development costs. Wall clock improvements of 5-10% are considered huge wins and are worth spending money on. EDA infrastructure cost is only ~10% of the chip development costs and storage is roughly 4% of the IT spend. Therefore, spending more on storage performance pays for itself if it improves the productivity and utilization of the engineers and licenses.

Performance measured by total turn-around time of all jobs

EDA performance is also measured in total turn-around time (TTT). This is the time it takes for a collection of jobs to complete in its entirety. For example, how long does it take for all 1000 simulations to run start to finish? The answer to that question depends on whether the jobs tests were run serially, in parallel, or in some combination. Load sharing tools like LSF, Grid Engine, Slurm, and so on allow engineers to submit large sets of jobs to run in parallel. The more jobs that can run in parallel, typically the faster all the jobs complete.

See the following example of running 20k jobs.

Table 1) Performance measured by total turn-around time of all jobs.

	Total jobs	Max parallel jobs	Ave job runtime (min)	TTT (min)	TTT (hrs)
Example 1	20,000	1,000	10	200	3.33
Example 2	20,000	5,000	10	40	0.67 (5x faster)

As you can see from the table above, running 5000 jobs in parallel results in a 5x improvement in job turn-around time, which in turn means it is 5x faster to find a bug and 5x faster to perform design closure iterations. Therefore, engineers can be 5x more productive.

Optimizing for maximum parallel performance and scale is a key trade-off when considering EDA design architectures. Cloud computing offers unlimited parallel compute scale because the cloud is vast and elastic versus the fixed resources in traditional datacenters. And in turn, Cloud provides significant improvements in TTT and productivity.

The ability to run more jobs in parallel is as important a consideration as optimizing single-job wall-clock-time performance.

Hybrid cloud data management architecture at enterprise scale

As cloud infrastructure scales up to support multiple EDA workflows, projects, and teams, you might need to address the following considerations.

Scaling tools and library

Tools and libraries are typically shared resources on-premises and thus should be in the cloud. Advanced semiconductor manufacturing processes (7nm, 5nm, and 3nm) have increased design complexity and scale, and design technology libraries have become very large as the complexities of representing transistor performance, power, and geometries have caused files to grow significantly. Therefore, library loading times have become a challenge.

The demand on library volumes when tool flows start up on 10k to 100k server cores simultaneously can create a boot storm on the storage filers leading to high read-latency conditions. This is further compounded by sharing tool and library volumes across multiple design flows, projects, and teams. Scaling FlexCache volumes to spread the read load out across the parallel compute cores can dramatically improve library and tool load times.

Because tool and library volumes (Cloud Volumes ONTAP instances) are shared by multiple projects and teams, deploying Cloud Volumes ONTAP in an HA configuration might be desirable for improve reliability, durability, and uptime. Single-node Cloud Volumes ONTAP file system require downtime for upgrades, whereas HA instances can fail-over temporarily to one of the two nodes during upgrades. HA configurations help ensure that tools and library volumes cause minimal downtime during upgrades or fault conditions.

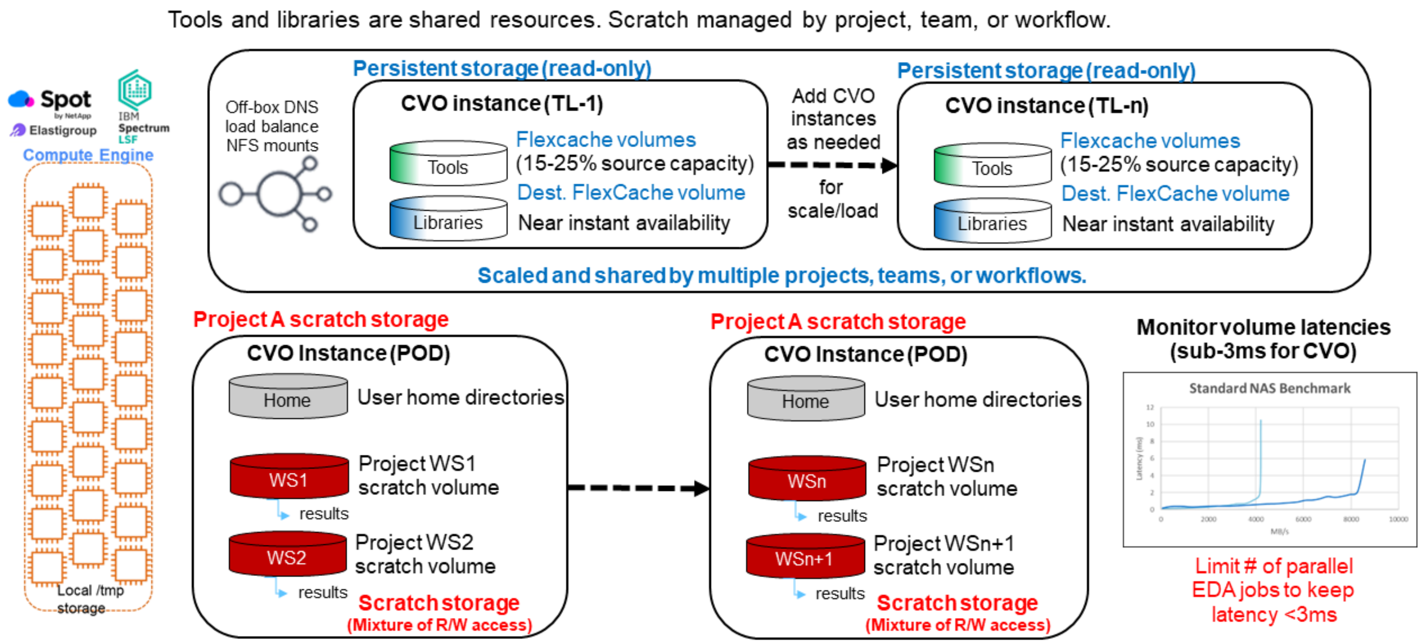
Scratch volumes at scale

Scratch volumes require maximum read/write performance. Single-node Cloud Volumes ONTAP configurations may provide improved write performance as compared to two-node HA configurations. To minimize any potential disruption during ONTAP upgrades or temporary fault conditions, Cloud Volumes ONTAP instances can be deployed on a per flow, team, or project basis. Cloud Volumes ONTAP upgrades can be coordinated with single flow, team, or project owners more easily than trying to coordinate between flows, projects, or teams.

The following architecture diagram shows how shared Cloud Volumes ONTAP file systems can be scaled to meet increasing tool and library read workloads. As more and more EC2 instances are elastically scaled-up to meet increasing design workload requirements, the tool and library FlexCache volumes can be quickly scaled-up and UNIX mount points can be DNS load-balanced to optimize performance. As EC2 instance are scaled back down, FlexCache volumes can be scaled back as well.

The following figure shows the cloud architecture for multi-project and multi-team deployments. Tools and libraries are shared resources, and scratch is managed by project, team, and workflow.

Figure 6) Cloud architecture for multi-project and multi-team deployments.


Tools and libraries are shared resources. Scratch managed by project, team, or workflow.
Spot by NetApp
Elastigroup
IBM Spectrum LSF
Compute Engine
Local /tmp storage
Off-box DNS load balance NFS mounts
Persistent storage (read-only)
CVO instance (TL-1)
Tools
Libraries
Flexcache volumes
(15-25% source capacity)
Dest. FlexCache volume
Near instant availability
Add CVO instances as needed
for scale/load
Persistent storage (read-only)
CVO instance (TL-n)
Tools
Libraries
Flexcache volumes
(15-25% source capacity)
Dest. FlexCache volume
Near instant availability
Scaled and shared by multiple projects, teams, or workflows.
Project A scratch storage
CVO Instance (POD)
Home
User home directories
WS1
Project WS1 scratch volume
results
WS2
Project WS2 scratch volume
results
Scratch storage
(Mixture of R/W access)
Project A scratch storage
CVO Instance (POD)
Home
User home directories
WSn
Project WSn scratch volume
results
WSn+1
Project WSn+1 scratch volume
results
Scratch storage
(Mixture of R/W access)
Monitor volume latencies (sub-3ms for CVO)
Standard NAS Benchmark
Latency (ms)
12
10
8
6
4
2
0
0
2000
4000
6000
8000
10000
MB/s
Limit # of parallel EDA jobs to keep latency <3ms

The cloud can provide a much more elastic compute and storage infrastructure than traditional fixed on-prem resources. As projects scale up and down based on project schedules and workload demands, Cloud Volumes ONTAP file systems can be rapidly created to provide performant storage on a workspace, workflow, or project basis.

Monitoring storage latency is the best way to identify when a file system becomes overloaded. A heavily loaded file system often sees increased latency represented as a knee in the latency curve. In the previous picture, latency remains <3ms until the number of parallel EDA jobs drives a high I/O load, resulting in I/O latency increasing rapidly above 3ms.

Monitoring the file system latency is a good way of calibrating how many EDA jobs can be run while keeping filer performance nominal (<3ms latency). Spreading project workspaces or flows across multiple Cloud Volumes ONTAP file system is a good way of load balancing storage workloads and providing optimal low latency I/O.

Quick start for AWS Cloud Volumes ONTAP setup

The purpose of this section is to document the current best practices for optimizing Cloud Volumes ONTAP for high-performance, high-scale EDA workloads. Details and an explanation of each option and configuration are provided in refence links to the next chapter titled QuickStart Implementation Details.

Cloud Volumes ONTAP creation setup

This step is run at least twice depending on the number of Cloud Volumes ONTAP instances required for tools, libraries, and scratch.

- **Cloud Volumes ONTAP instance for scale-out tool and library FlexCache volumes.** This configuration is optimized for very high read-only I/O performance as well as durability. NetApp recommends using an HA pair file system configuration.
- **Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes.** This configuration is optimized for an absolute max I/O throughput for read/write performance at scale with the trade-off of some durability. NetApp recommends using the single-node file-system configuration.

  OR
- **Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch/release volumes.** This configuration is optimized for maximum I/O throughput for read/write performance at scale while

providing HA durability. This configuration has less write performance than a single-node Cloud Volumes ONTAP configuration. NetApp recommends using an HA-pair file-system configuration.

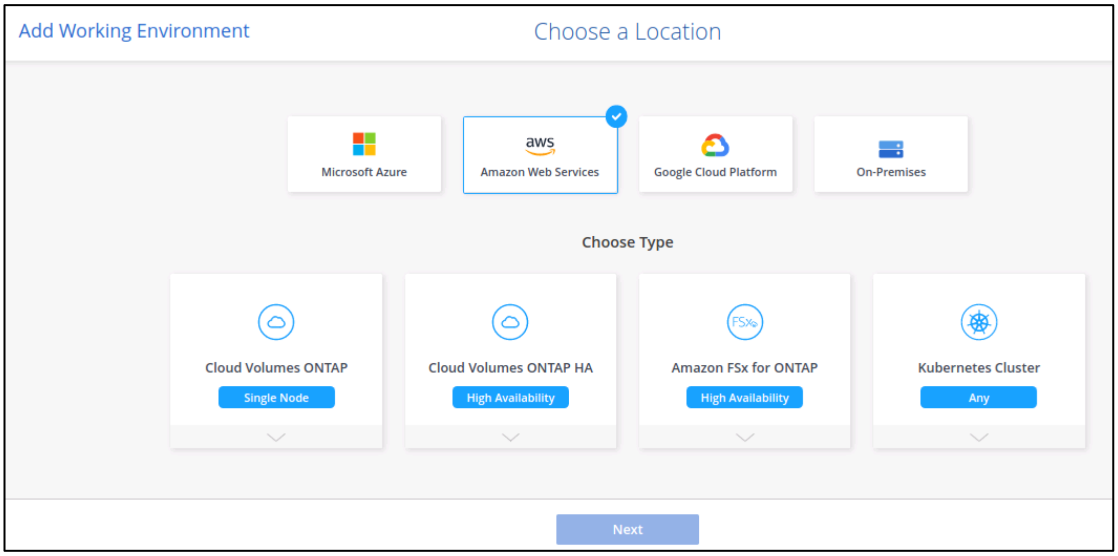
Open NetApp Blue XP

Log into your BlueXP account using your NetApp support site credentials. This is the first step to register your environment with NetApp Active IQ.

Add working environment

Select **AWS** and then select **Choose Type** based on the type of Cloud Volumes ONTAP file system needed.

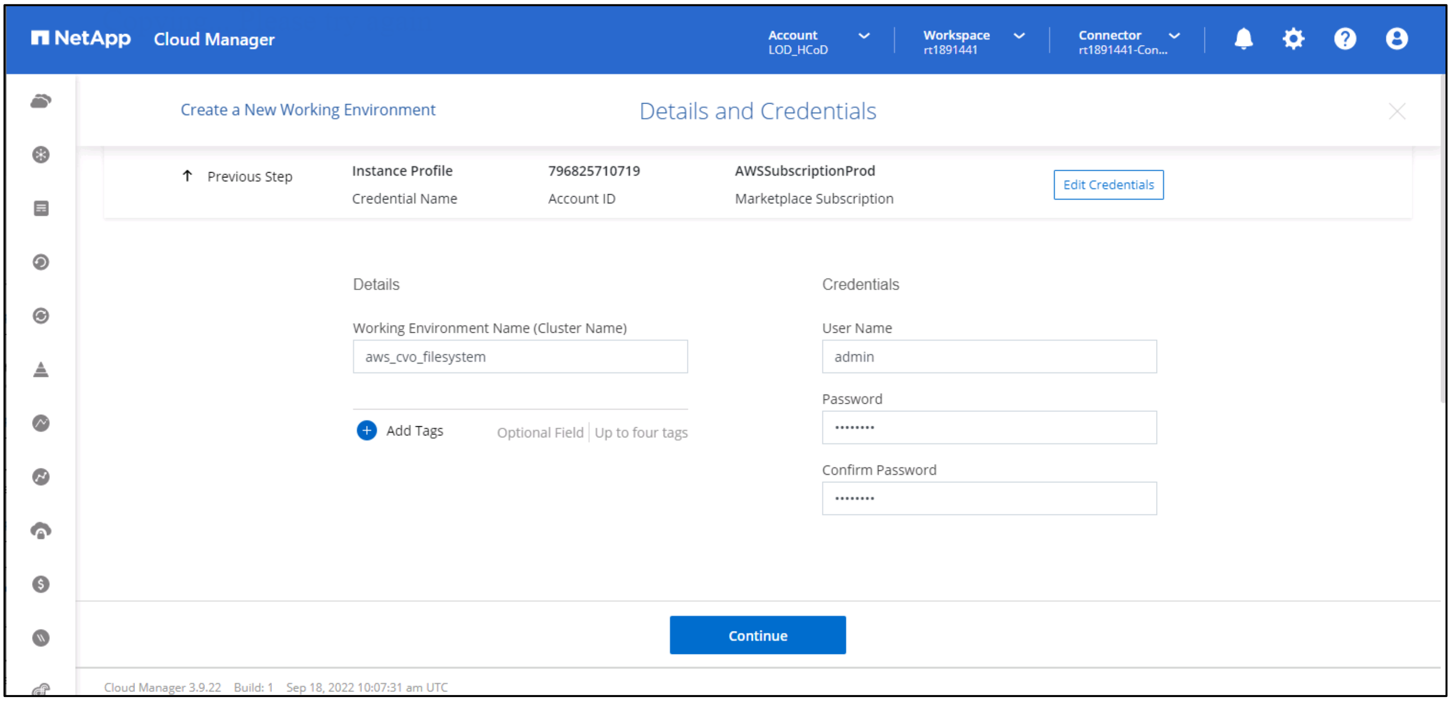
- Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes. Select **Cloud Volumes ONTAP HA**.
- Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes. Select **Cloud Volumes ONTAP Single Node**.
- Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch/release volumes. Select **Cloud Volumes ONTAP HA**.



Details and credentials

Provide a name for the Cloud Volumes ONTAP file system and provide login credentials.

1. Select the **Working Environment Name (Cluster Name)** dialog box and entering a cluster name**.**
2. Select the **User Name** dialog box and entering a user name.
3. Select the **Password** dialog box and enter a strong password. Repeat the password in the **Confirm password** dialog box.



Create a New Working Environment

Details and Credentials

Previous Step

Instance Profile	796825710719	AWSSubscriptionProd
Credential Name	Account ID	Marketplace Subscription

Details

Working Environment Name (Cluster Name): aws_cvo_filesystem

Add Tags — Optional Field | Up to four tags

Credentials

User Name: admin

Password: ••••••••

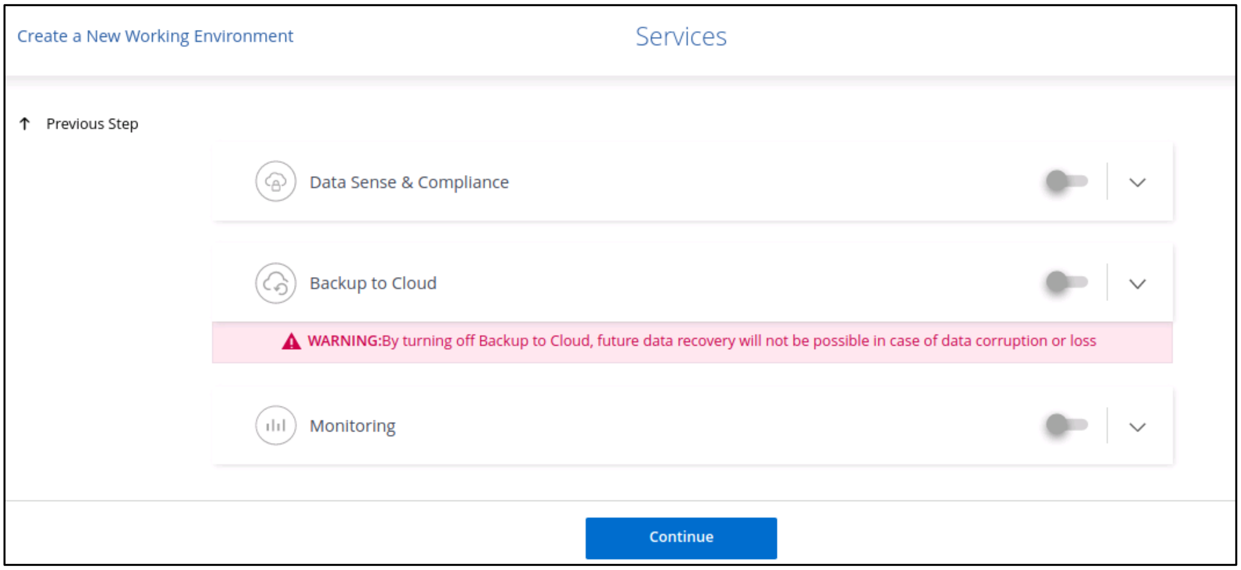
Confirm Password: ••••••••

Cloud Manager 3.9.22 Build: 1 Sep 18, 2022 10:07:31 am UTC

Select added services

For more detailed information about this subject, see the section “Added services.”

- **Data Sense and Compliance.** Optional. Because this feature has a significant effect on performance, NetApp recommends turning it off.
- **Back-up to cloud.** Optional. This is SnapMirror-to-cloud as a service. This feature has a negligible effect on performance, but it might not be needed in most cases.
  - Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes. Select Back-up to Cloud – OFF.
  - Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes. Select **Back-up to Cloud – OFF**.
  - Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch and release volumes. Select **Back-up to Cloud – OFF** (Select **ON** only if backups are required. Scratch volumes are typically not backed up).
- **Monitoring.** Optional. Monitoring has a minimal effect on performance. Select **OFF**. This feature enables Cloud Insights monitoring.


Create a New Working Environment

Services

Previous Step

Data Sense & Compliance

Backup to Cloud

⚠ WARNING:By turning off Backup to Cloud, future data recovery will not be possible in case of data corruption or loss

Monitoring

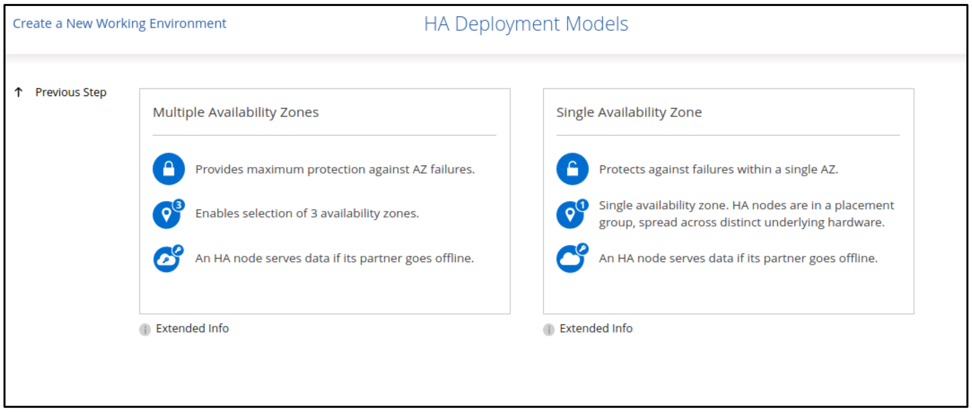
Continue

Select availability zone

For more detailed information about this subject, see the section “Select availability zone (AZ).”

Single availability zone configurations are recommended for all Cloud Volumes ONTAP types.

- Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes (HA config). Select ***Single Availability Zone*** (best for FlexCache optimized volumes).
- Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes (single-node config) N/A – by definition, single-node Cloud Volumes ONTAP file systems exist in a single availability zone.
- Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch/release volumes (HA config). Select **Single Availability Zone HA**.


Create a New Working Environment

HA Deployment Models

Previous Step

Multiple Availability Zones

Provides maximum protection against AZ failures.

Enables selection of 3 availability zones.

An HA node serves data if its partner goes offline.

Extended Info

Single Availability Zone

Protects against failures within a single AZ.

Single availability zone. HA nodes are in a placement group, spread across distinct underlying hardware.

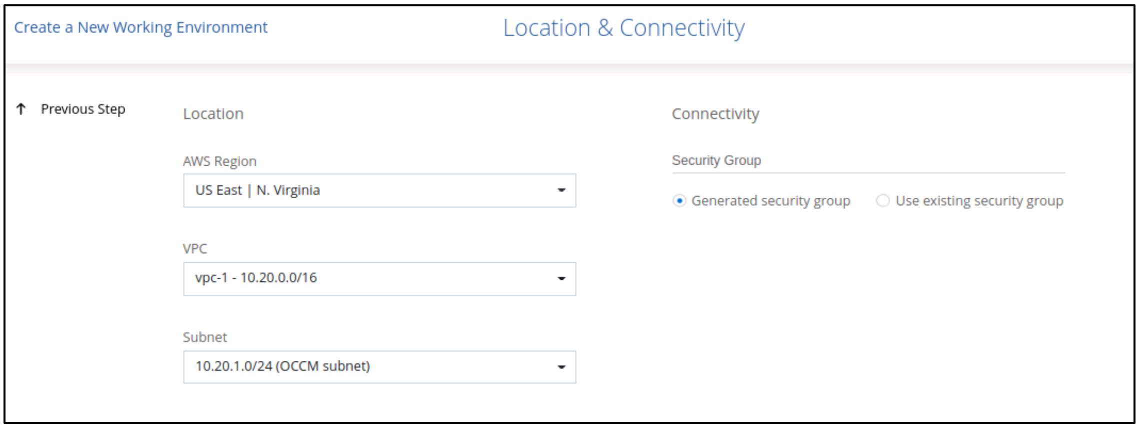
An HA node serves data if its partner goes offline.

Extended Info

Location and connectivity settings

There are no EDA-specific best practice requirements. Options are specific to the AWS Virtual Private Cloud (VPC) design.

- AWS region
- VPC
- Subnet
- Security group



Location & Connectivity

Location

AWS Region: US East | N. Virginia

VPC: vpc-1 - 10.20.0.0/16

Subnet: 10.20.1.0/24 (OCCM subnet)

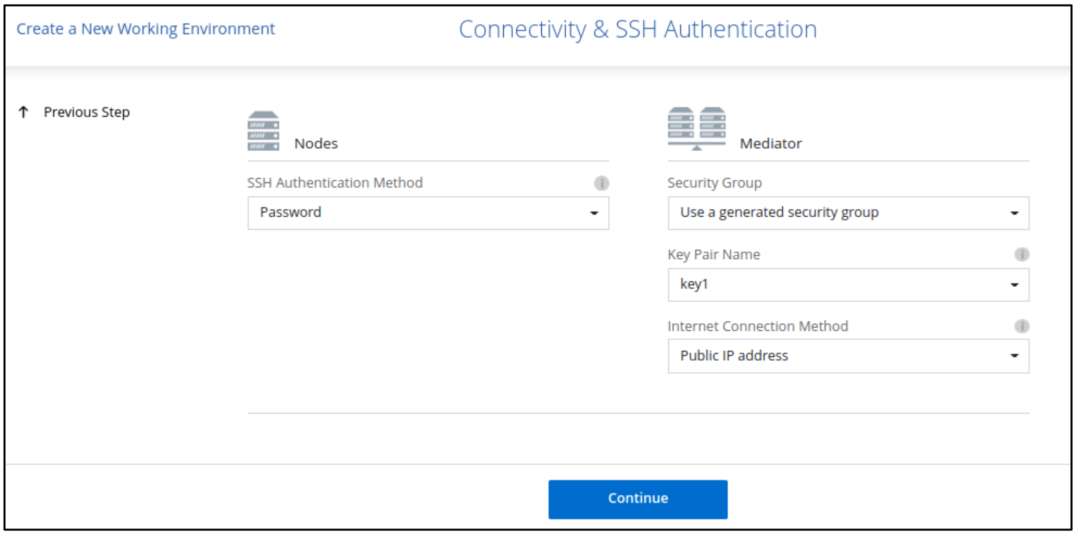
Connectivity

Security Group: ● Generated security group ○ Use existing security group

Nodes and mediator, AWS-managed encryption

There are no EDA-specific best practice requirements.

- **Nodes.** SSH authorization method > Password (only option).
- **Mediator.** Security group/key pair name and internet connection method.



Connectivity & SSH Authentication

Nodes

SSH Authentication Method: Password

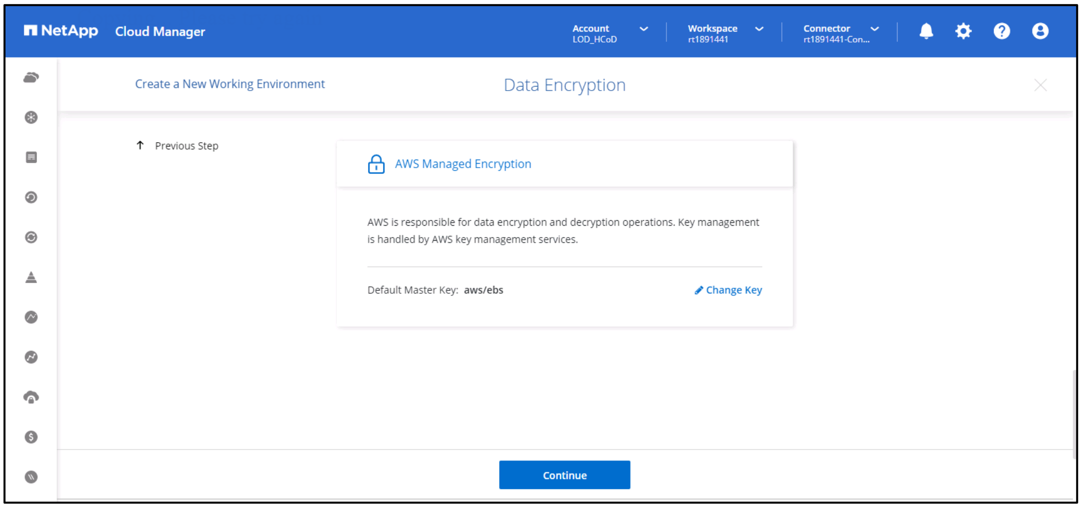
Mediator

Security Group: Use a generated security group

Key Pair Name: key1

Internet Connection Method: Public IP address

- **AWS-managed encryption.** No changes required.



Data Encryption

AWS Managed Encryption

AWS is responsible for data encryption and decryption operations. Key management is handled by AWS key management services.

Default Master Key: aws/ebs

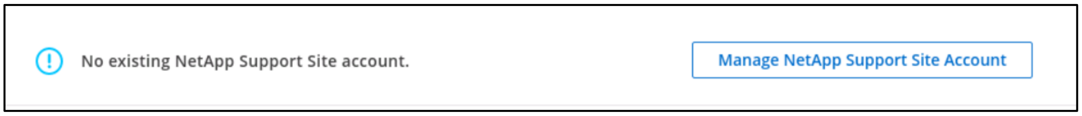
Change Key

Manage NetApp Support Site account

For more detailed information about this subject, see the section “Manage NetApp Support Site account.”

This feature enables Cloud Volumes ONTAP communication with the NetApp Support Site.

Click the **Manage NetApp Support Site Account** button.



No existing NetApp Support Site account.

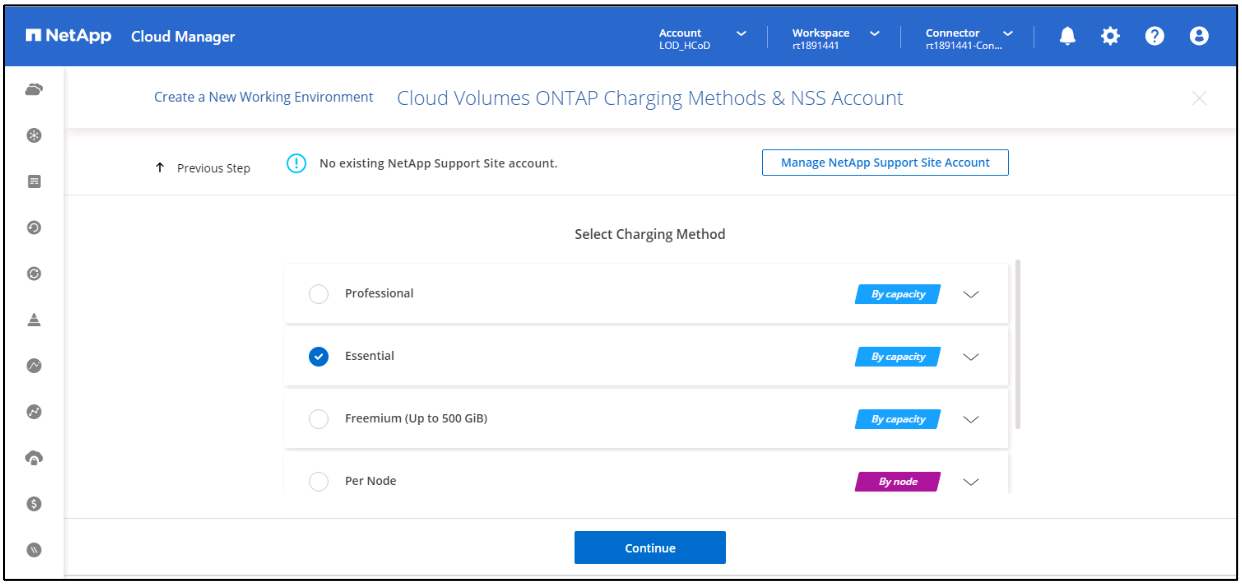
BlueXP must also be registered. Click the link at the top of the BlueXP UI.

Select charging (licensing) method

For more detailed information about this subject, see the section “Select charging method (Licensing).”

Most if not all EDA customers use the Essential License because no back-up to the cloud is required.

- Select **Essential by Capacity**.



NetApp Cloud Manager
Account LOD_HCoD
Workspace rt1891441
Connector rt1891441-Con...
Create a New Working Environment
Cloud Volumes ONTAP Charging Methods & NSS Account
Previous Step
Manage NetApp Support Site Account
Select Charging Method
By capacity
By capacity
By capacity
By node
Continue

No existing NetApp Support Site account.

○ Professional

● Essential

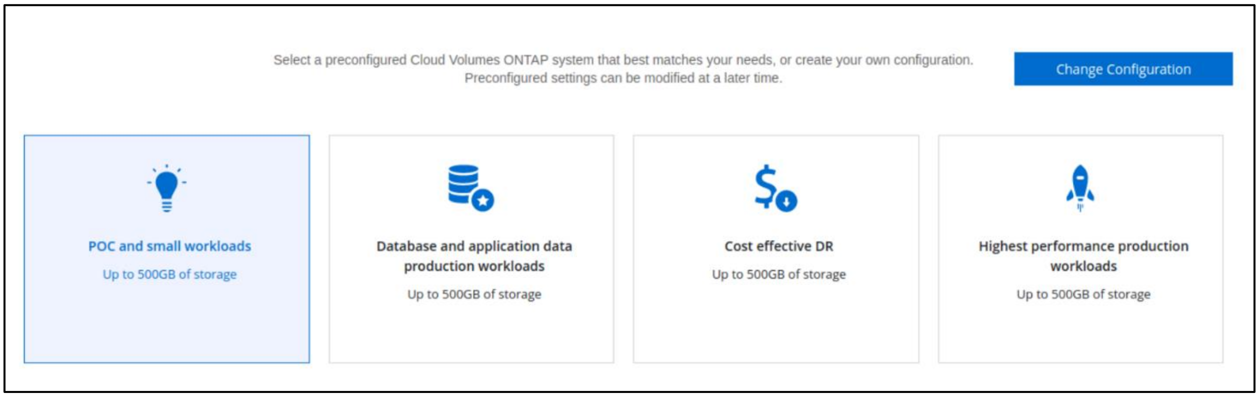
○ Freemium (Up to 500 GiB)

○ Per Node

Configure Cloud Volumes ONTAP instance

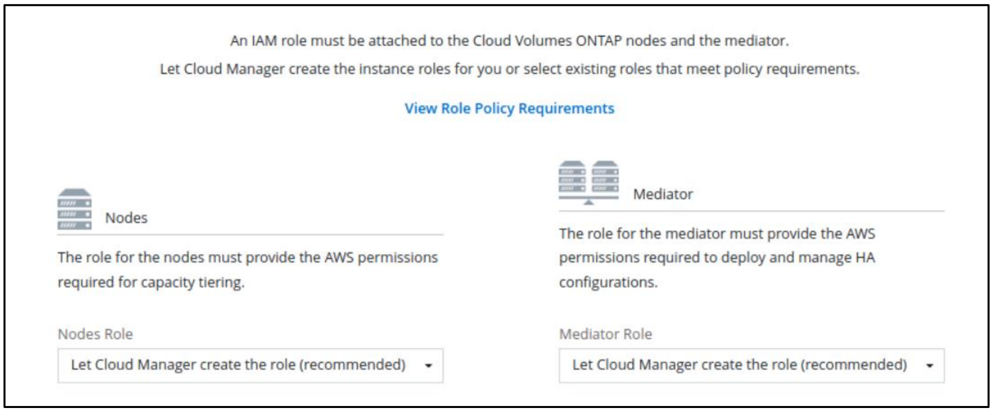
For more detailed information about this subject, see the section “Cloud Volumes ONTAP sizing considerations for EDA workloads.”

To select the Cloud Volumes ONTAP Config Change Configuration, complete the following tasks, click **Change Configuration**. This allows the selection of specific EC2 and EBS instance types.



Select a preconfigured Cloud Volumes ONTAP system that best matches your needs, or create your own configuration. Preconfigured settings can be modified at a later time.

4. **View Role Policy Requirements**. There are no specific EDA requirements.
5. Select the defaults.



An IAM role must be attached to the Cloud Volumes ONTAP nodes and the mediator.
Let Cloud Manager create the instance roles for you or select existing roles that meet policy requirements.

The role for the nodes must provide the AWS permissions required for capacity tiering.

The role for the mediator must provide the AWS permissions required to deploy and manage HA configurations.

Nodes Role: Let Cloud Manager create the role (recommended)

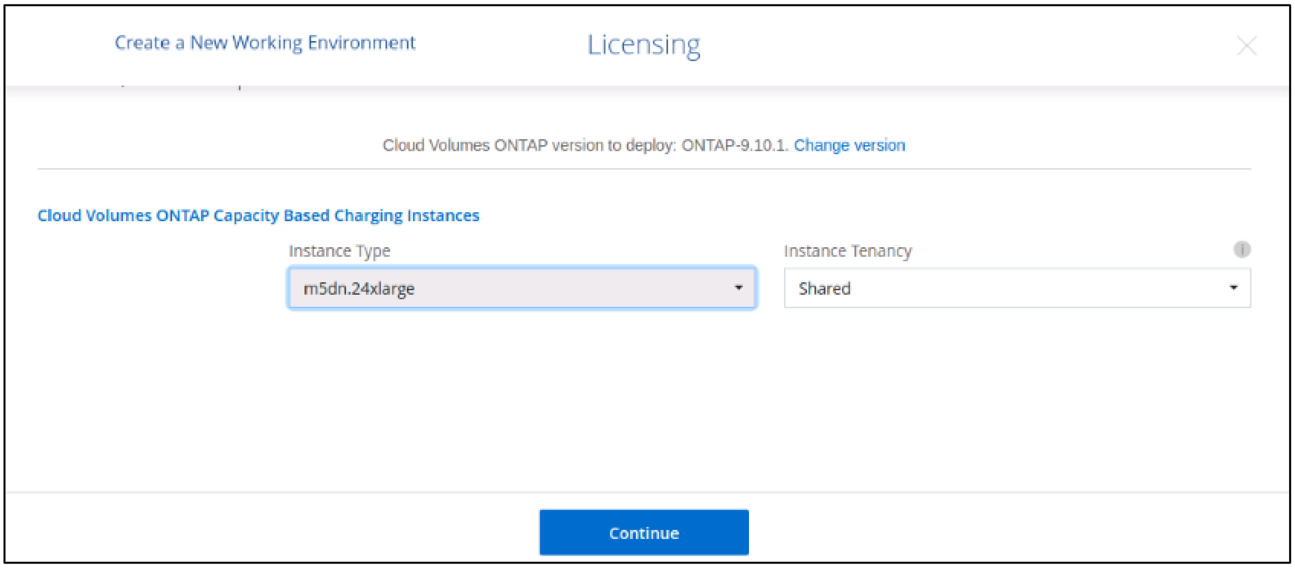
Mediator Role: Let Cloud Manager create the role (recommended)

Cloud Volumes ONTAP version to deploy

To deploy the appropriate Cloud Volumes ONTAP version, select **Change Version** > Select **Latest ONTAP Version**.

Select EC2 Instance Type to Deploy

1. Select the instance type based on the table below.
2. Select **Shared** for the Instance Tenancy.



Create a New Working Environment
Licensing

Cloud Volumes ONTAP version to deploy: ONTAP-9.10.1. Change version

Cloud Volumes ONTAP Capacity Based Charging Instances

Instance Type: m5dn.24xlarge

Instance Tenancy: Shared

Table 2) Table of Cloud Volumes ONTAP configurations (based on GP3 disk type).

Effective (usable) capacity needed (TB)	EC2 instance	AWS disk size	Total aggregates per CVO	EBS volumes per aggregate
40	m5dn.24xlarge	8TB	1 aggr	6
80	m5dn.24xlarge	16TB	1 aggr	6
120	m5dn.24xlarge	16TB*	2 aggr	6
180	m5dn.24xlarge	16TB**	3 aggr	6

* This size requires two data aggregates.

** This size requires three data aggregates.

Note: If additional or more granular configurations are required, consult your NetApp account team.

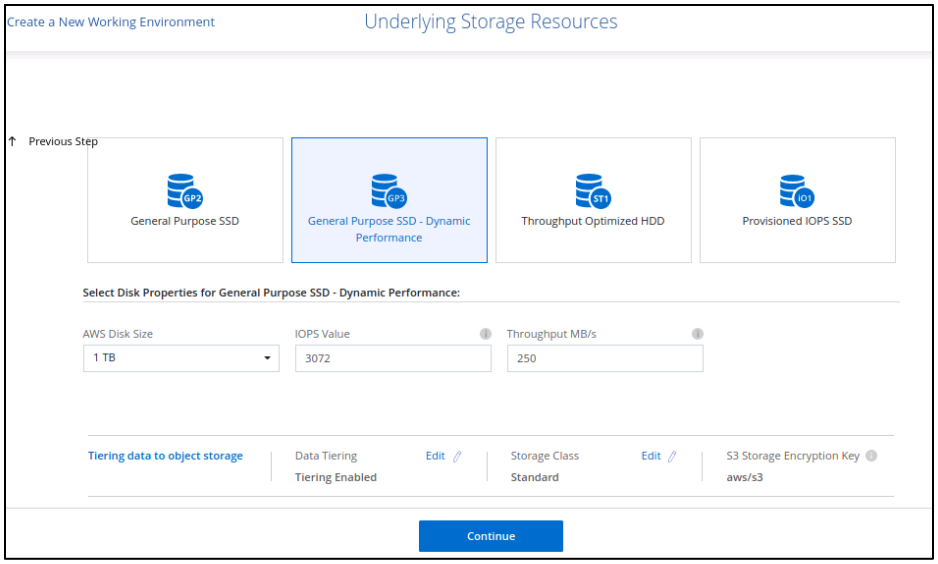
Underlying storage resources

For more detailed information on this subject, see the section “Cloud Volumes ONTAP disk selection and throughput requirements.”

1. Select **Provisioned General Purpose – Dynamic Performance (GP3)**.
2. Select the appropriate disk size value from the previous table.
3. Select the appropriate IOPs value based on the previous table.
4. Select the appropriate throughput value based on the previous table.

This step deploys a single disk. Additional disks are added to the aggregate post Cloud Volumes ONTAP deployment.

See the section “Post Cloud Volumes ONTAP Creation Setup” for more information. Modify aggregates


Create a New Working Environment
Underlying Storage Resources
Previous Step
General Purpose SSD
General Purpose SSD - Dynamic Performance
Throughput Optimized HDD
Provisioned IOPS SSD
Continue

Select Disk Properties for General Purpose SSD - Dynamic Performance:

AWS Disk Size: 1 TB
IOPS Value: 3072
Throughput MB/s: 250

Tiering data to object storage
Data Tiering: Tiering Enabled (Edit)
Storage Class: Standard (Edit)
S3 Storage Encryption Key: aws/s3

Tiering data to object storage

Automated age-based cold data tiering to the object storage tier.

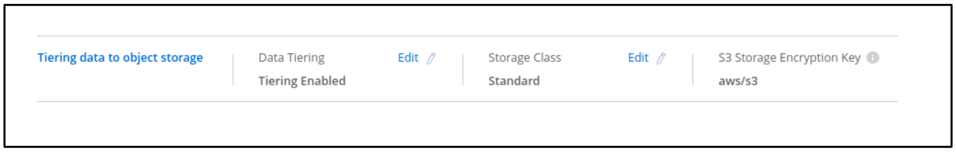
- Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes.

Select **No Tiering Required**. In most cases, tiering FlexCache tool and libraries volumes is unnecessary. Tiering the source volume on-premises does make sense because tool and library volumes often contain greater than 60% cold data.

- Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes.

  Select **No Tiering Required**. Scratch volumes often contain 60-80% cold data. If volumes provisioned for scratch workloads are going to persist in the cloud for long periods of time, tiering is recommended, otherwise select **No Tiering Required**. This can be enabled after Cloud Volumes ONTAP creation.
- Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch/release volumes.

  Select **Tiering Optional** based on data aging data. Scratch volumes often contain 60-80% cold data. If volumes provisioned for scratch workloads are going to persist in the cloud for long periods of time, tiering is recommended.



Tiering data to object storage

Data Tiering: Tiering Enabled

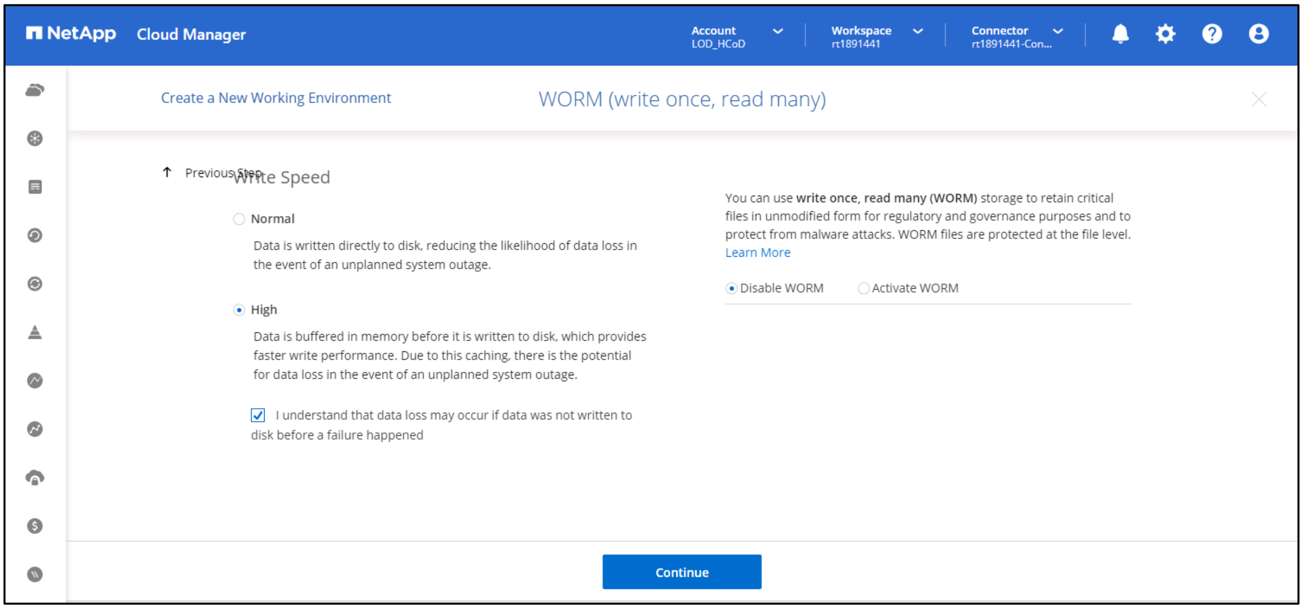
Storage Class: Standard

S3 Storage Encryption Key: aws/s3

Enable High Write Speed mode and WORM mode

- For **Write Speed**, select **High**.
- For **Write-Once Read-Many (WORM)**, select **Disable WORM**.

  WORM is generally not needed for EDA workloads. Rather, it is typically used for data compliance requirements.



WORM (write once, read many)

Write Speed

○ Normal

Data is written directly to disk, reducing the likelihood of data loss in the event of an unplanned system outage.

● High

Data is buffered in memory before it is written to disk, which provides faster write performance. Due to this caching, there is the potential for data loss in the event of an unplanned system outage.

☑ I understand that data loss may occur if data was not written to disk before a failure happened

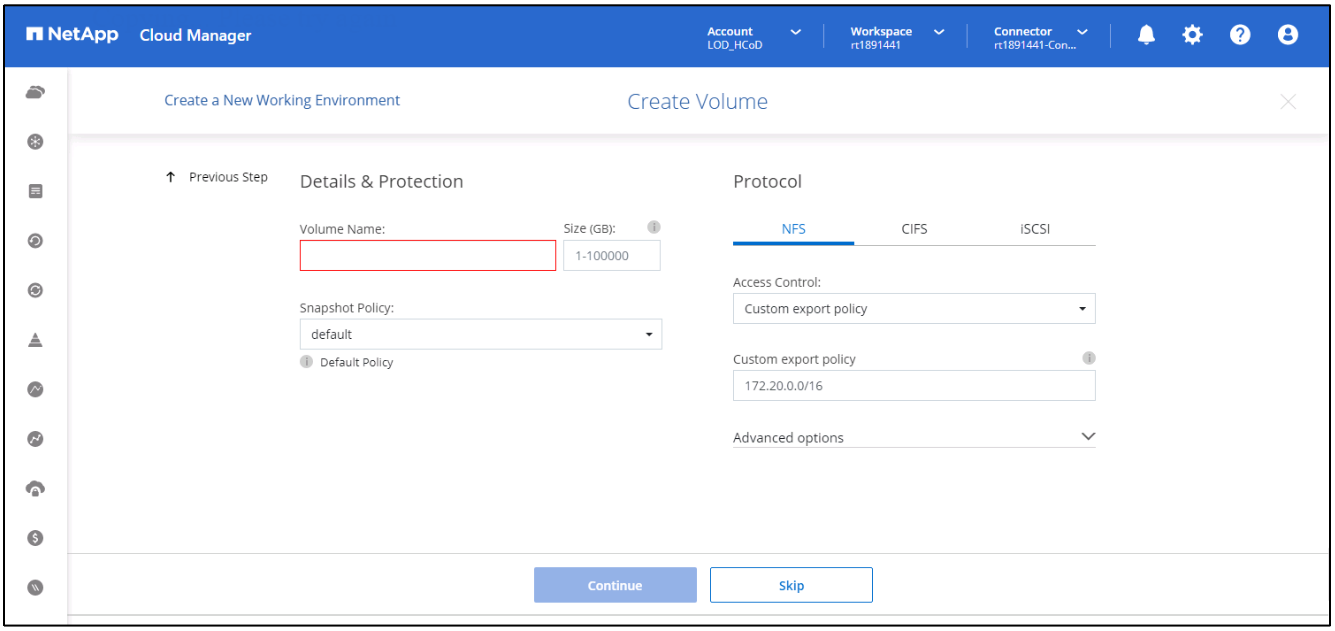
You can use write once, read many (WORM) storage to retain critical files in unmodified form for regulatory and governance purposes and to protect from malware attacks. WORM files are protected at the file level. Learn More

● Disable WORM ○ Activate WORM

Create volume (SKIP)

Volume provisioning is configured later in the section “Post Cloud Volumes ONTAP creation setup.”

Click **Skip**.



Create Volume

Details & Protection

Volume Name:

Size (GB): 1-100000

Snapshot Policy: default

Default Policy

Protocol

NFS | CIFS | iSCSI

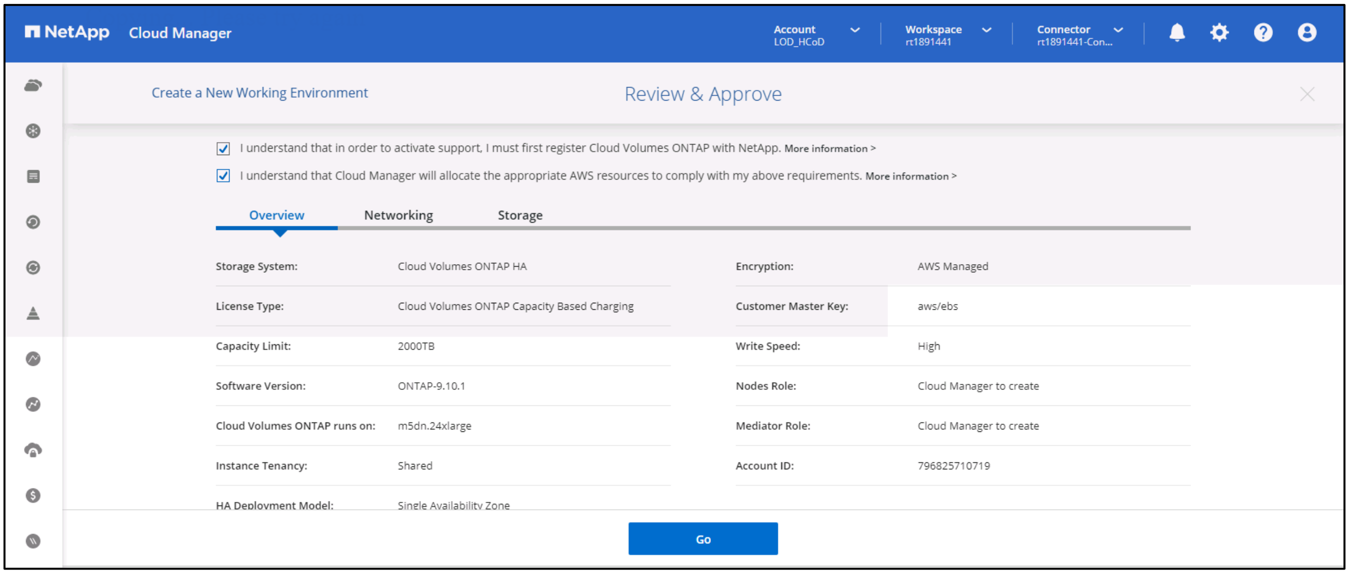
Access Control: Custom export policy

Custom export policy: 172.20.0.0/16

Advanced options

Review and approve

Review the Cloud Volumes ONTAP creation configuration. If the settings look correct, click **Go** to start the automated Cloud Volumes ONTAP creation process.



Review & Approve

☑ I understand that in order to activate support, I must first register Cloud Volumes ONTAP with NetApp. More information >

☑ I understand that Cloud Manager will allocate the appropriate AWS resources to comply with my above requirements. More information >

Storage System:	Cloud Volumes ONTAP HA	Encryption:	AWS Managed
License Type:	Cloud Volumes ONTAP Capacity Based Charging	Customer Master Key:	aws/ebs
Capacity Limit:	2000TB	Write Speed:	High
Software Version:	ONTAP-9.10.1	Nodes Role:	Cloud Manager to create
Cloud Volumes ONTAP runs on:	m5dn.24xlarge	Mediator Role:	Cloud Manager to create
Instance Tenancy:	Shared	Account ID:	796825710719
HA Deployment Model:	Single Availability Zone		

Post Cloud Volumes ONTAP creation setup

For more detailed information on this subject, see the section “Post Cloud Volumes ONTAP creation setup.”

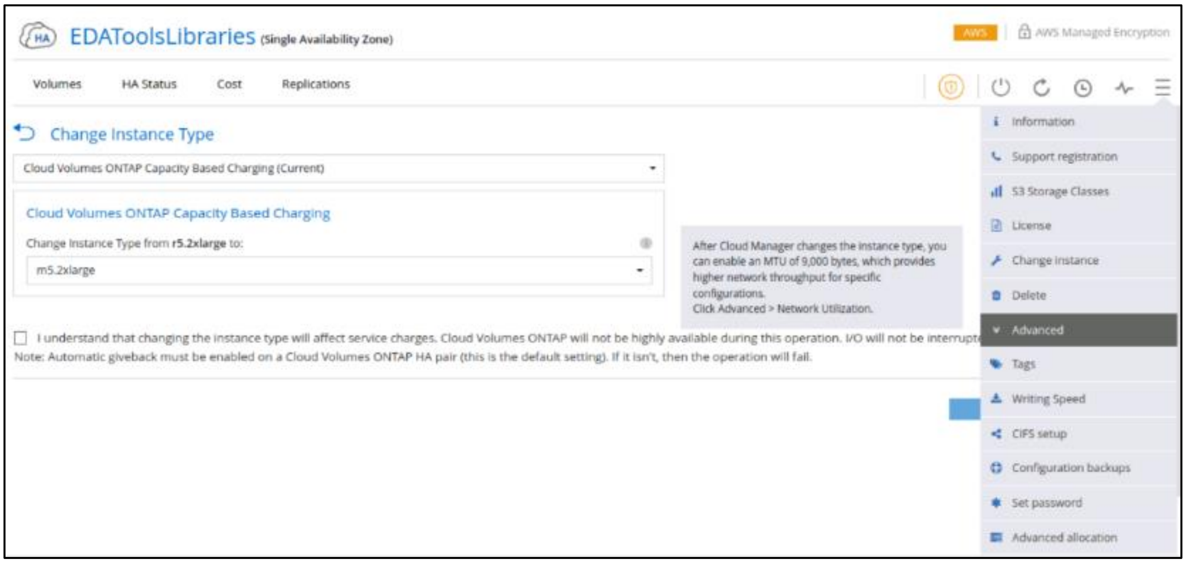
The following steps are run on each Cloud Volumes ONTAP file system after the file system is available and on-line in Cloud Manager.

Modify aggregates

For more detailed information about this subject, see the section “Modify aggregate setup details.”

Change aggregate settings to match the corresponding configuration in Table 2 in the previous section.

From the right drop-down menu, select **Advanced > Advanced Allocation**.



Change Instance Type

Cloud Volumes ONTAP Capacity Based Charging (Current)

Cloud Volumes ONTAP Capacity Based Charging

Change Instance Type from **r5.2xlarge** to:

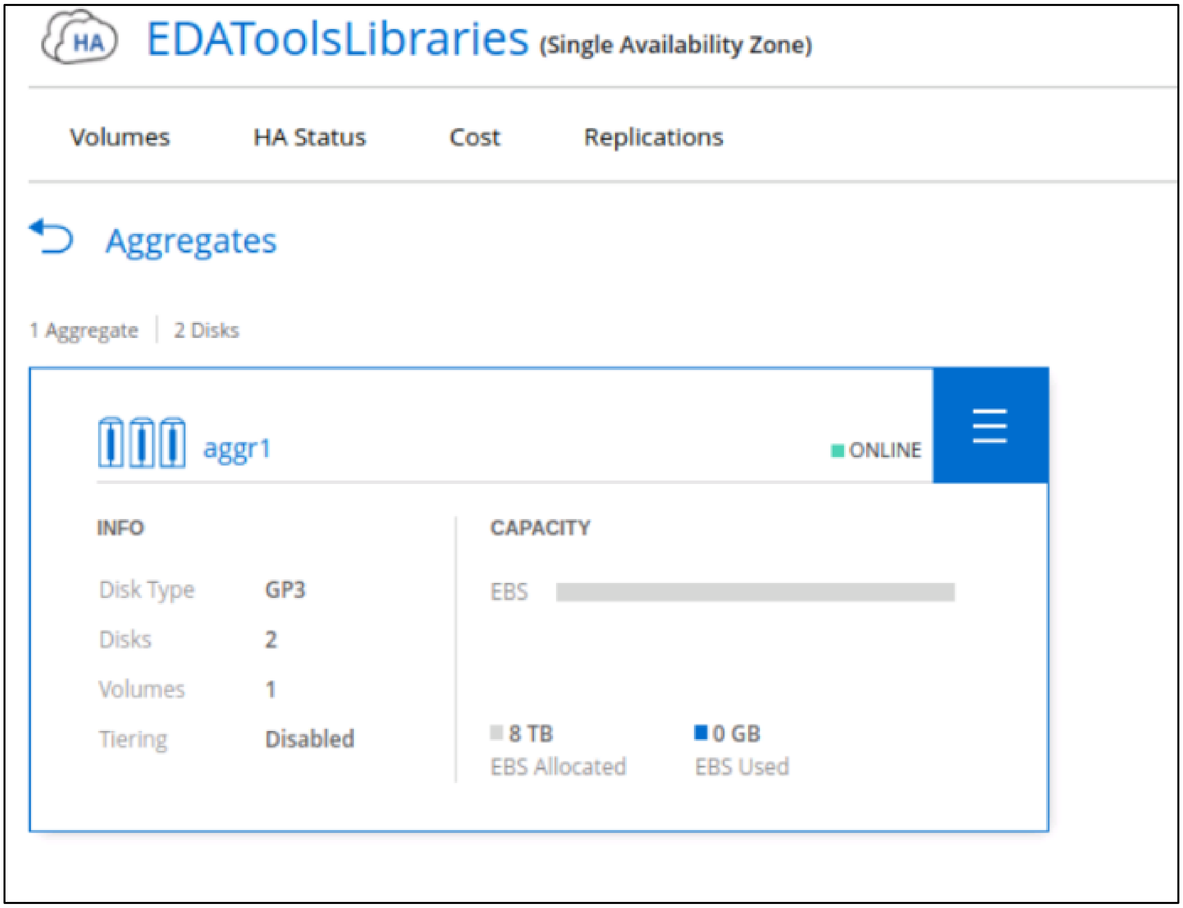
m5.2xlarge

After Cloud Manager changes the instance type, you can enable an MTU of 9,000 bytes, which provides higher network throughput for specific configurations.
Click Advanced > Network Utilization.

☐ I understand that changing the instance type will affect service charges. Cloud Volumes ONTAP will not be highly available during this operation. I/O will not be interrupte

Note: Automatic giveback must be enabled on a Cloud Volumes ONTAP HA pair (this is the default setting). If it isn't, then the operation will fail.

1. Select **Aggregates Menu > Add AWS disk**.
2. Select **AWS Disks** to **Add Per Aggregate** and change the number of disks to **5**, so the total number of disks match the disks per aggregate listed in Table 2, which in this example is 6 total disks per aggregate.

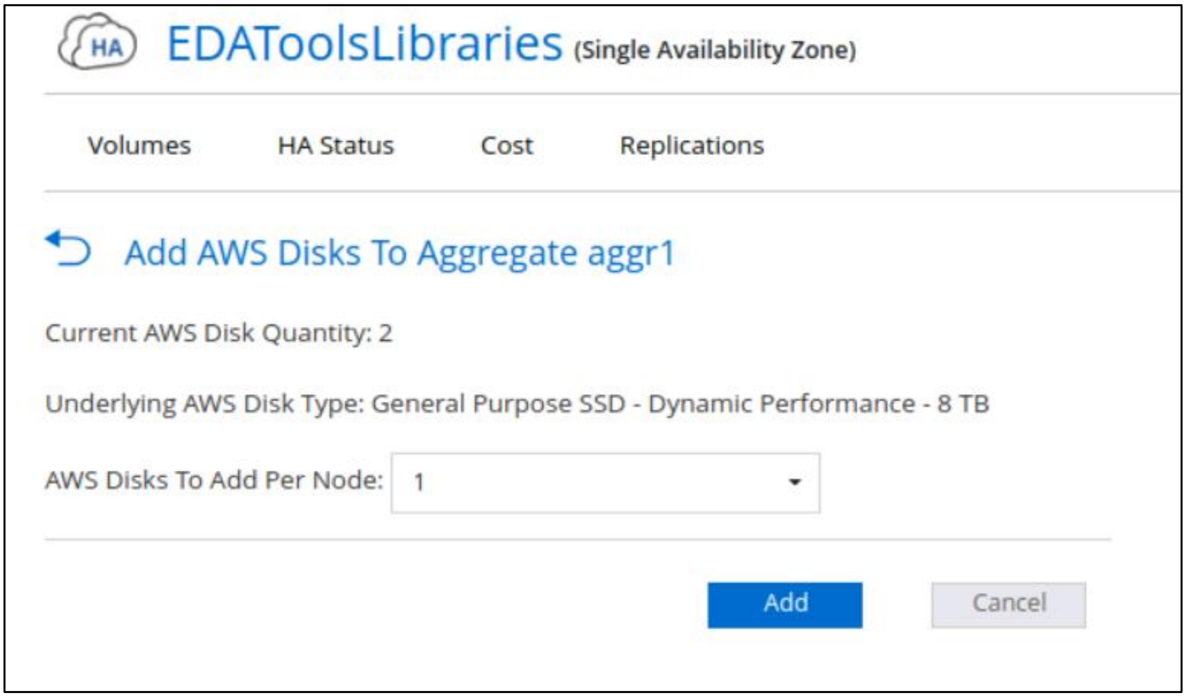


Disk Type	GP3
Disks	2
Volumes	1
Tiering	Disabled

8 TB EBS Allocated

0 GB EBS Used

This step results in one aggregate with six disks per node in total.



Add AWS Disks To Aggregate aggr1

Current AWS Disk Quantity: 2

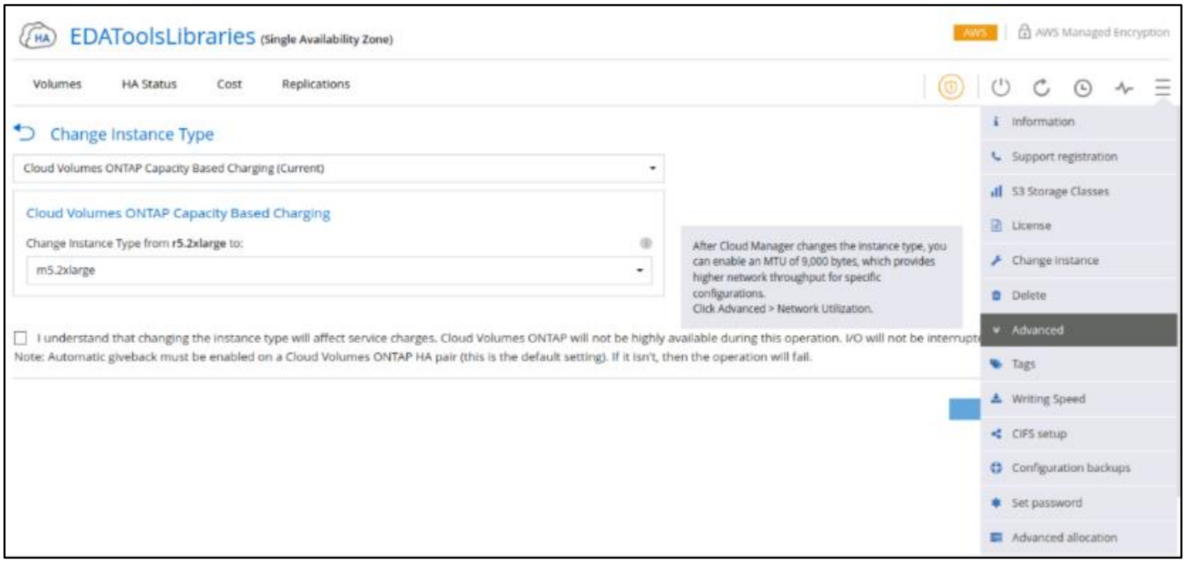
Underlying AWS Disk Type: General Purpose SSD - Dynamic Performance - 8 TB

AWS Disks To Add Per Node: 1

Add aggregates

Add additional aggregates based on the configuration requirements in Table 2.

1. From the right drop-down menu, select **Advanced > Advanced Allocation**.



Change Instance Type

Cloud Volumes ONTAP Capacity Based Charging (Current)

Cloud Volumes ONTAP Capacity Based Charging

Change Instance Type from **r5.2xlarge** to:

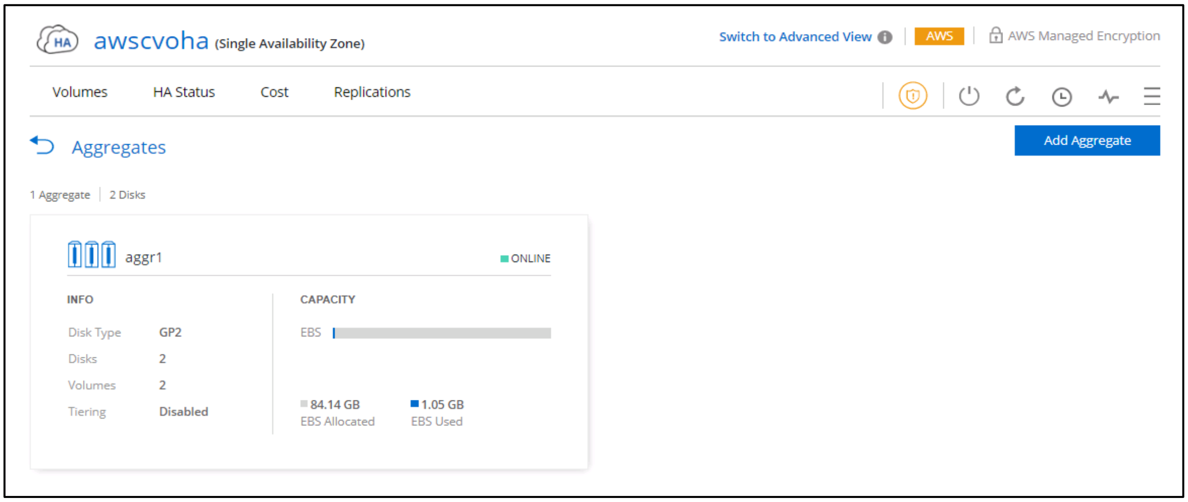
m5.2xlarge

After Cloud Manager changes the instance type, you can enable an MTU of 9,000 bytes, which provides higher network throughput for specific configurations.
Click Advanced > Network Utilization.

□ I understand that changing the instance type will affect service charges. Cloud Volumes ONTAP will not be highly available during this operation. I/O will not be interrupte

Note: Automatic giveback must be enabled on a Cloud Volumes ONTAP HA pair (this is the default setting). If it isn't, then the operation will fail.

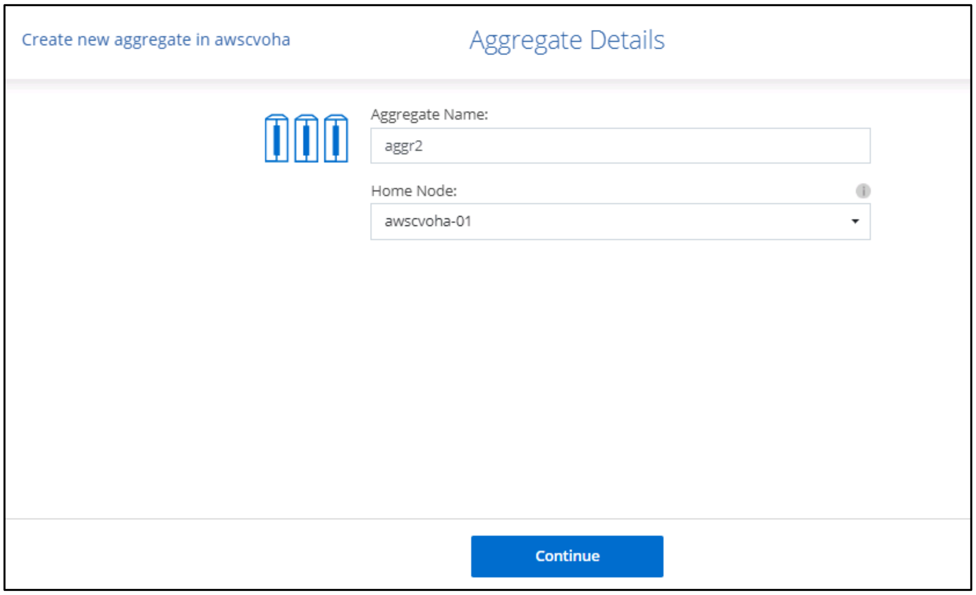
2. Select **Add Aggregates**.



Disk Type: GP2
Disks: 2
Volumes: 2
Tiering: Disabled

84.14 GB EBS Allocated
1.05 GB EBS Used

3. Select the **Aggregate Name** dialog box and type an aggregate name like aggr2. Press **Continue.**



Aggregate Name: aggr2
Home Node: awscvoha-01

4. Repeat these steps to add additional aggregates and disks per aggregate as required by Table 2.

Active-passive or active-active

For more detailed information about this subject, see the section “Active-active vs active-passive.”

Active-passive is the default for Cloud Volumes ONTAP and is recommended for all Cloud Volumes ONTAP file system types, tools, libraries, and scratch. Because it is the default, there is no need to make any changes.

Recommended EDA-specific options (pre-volume provision)

NetApp recommends enabling 64-bit identifiers as follows:

```
ONTAP cli> set -privilege advance
ONTAP cli> vserver nfs modify -vserver [VSERVER_NAME] -v3-64bit-identifiers enabled
```

Create volumes

For more detailed information about this subject, see the section “Volumes provision (creating a new volume).”

EDA workloads significantly benefit from FlexGroup volumes for maximum scale and performance. BlueXP does not support the creation of FlexGroup or FlexCache volume types from the web UI at time of publishing (this may change with newer BlueXP releases).

FlexGroup and FlexCache volumes must be provisioned with the Cloud Volumes ONTAP CLI or with the ONTAP REST API.

ONTAP CLI commands are shown below.

Create FlexCache volume for tool and library volumes

Tool and library volumes are automatically replicated (cached) to AWS from on-premises ONTAP volume(s) by FlexCache. More than one FlexCache volume might be required depending on the number of tool and library volumes that need to be replicated to the cloud.

This NetApp ONTAP documentation link provides the instructions for setting up an ONTAP cluster and SVM peering relationship between two ONTAP instances.

After the cluster and SVM peering relationships have been set up and verified, FlexCache volumes can be created using the following command.

```
ONTAP cli> volume flexcache create -origin-vserver [ORIGIN_SVM_NAME] -origin-volume [ORIGIN_VOLUME_NAME] \
                                   -vserver [DESTINATION_SVM_NAME] -volume [DESTINATION_VOLUME_NAME] \
                                   -size [VOLUME_SIZE] -junction-path [JUNCTION_PATH] \
                                   -aggr-list aggr -aggr-list-multiplier 8 \
                                   -policy [EXPORT_POLICY]
```

FlexCache volumes should be constrained to a single node, single aggregate, and single LIF configuration. FlexCache volumes should NOT span multiple aggregates. This ensures that cross node latency and throughput performance does not limit overall volume performance.

Size volume capacity to address large-file issues (see below). See the section “Large file considerations” for more information.

Set atime-update on origin/source to False. (ONTAP 9.11 or earlier)

Set `atime-update` on the origin (or source volume) to `False`. To avoid invalidations on files that are cached when there is only a read at the origin, turn off the last accessed time updates on the origin volume. If the origin or source volume is running ONTAP 9.6 or later, the creation of a FlexCache automatically sets `atime-update` to `False`.

From the origin or source volume ONTAP CLI (ONTAP version earlier than 9.6), run the following command:

```
ONTAP cli> volume modify -vserver origin-svm -volume vol_fc_origin -atime-update false
```

In releases earlier than NetApp ONTAP 9.11.1, it is a best practice is to disable `atime-update` on the FlexCache origin volumes to reduce the amount of data being updated in the cache because of read access. Failing to disable `atime-update` on the source volume might lead to overly aggressive cache eviction. For example, simple reads at the source volume could evict the data from the cache.

NetApp ONTAP 9.11.1 and later provides an improved workflow that can allow for `atime-update` to be propagated across the environment and not affect evictions at the cache. Even reads at the cache can update atimes on the origin. The cache is never the authority on atime, only the origin is.

After upgrading the origin or source volume to ONTAP 9.11.1 or later, if there is no need for atime, then leave `atime-update` off. If there are workflow decisions made based on atime, then use the knobs created in 9.11.1 appropriately for your needs.

Create FlexGroup volumes for read/write scratch volumes

Create one or more FlexGroup volumes for the scratch workload.

Based on SPECstorage 2020 EDA benchmark testing, the best FlexGroup volume performance is achieved when FlexGroup volumes use the following recommendations:

- FlexGroup volumes should be constrained to a single node, single aggregate, and single LIF configuration. This provides cross-node latency, and throughput performance does not limit overall volume performance.
- For best scale-out performance, a single FlexGroup volume with qtrees is recommended, vs provisioning multiple smaller FlexGroup volumes. .
- Size volume capacity to address large-file issues (see below). Provisioning a single large FlexGroup volume managed with qtrees is recommended over provisioning multiple smaller volumes.

FlexGroup CLI command

```
ONTAP cli> volume create -vserver [SVM_NAME] -volume [VOLUME_NAME] \
                         -size [VOLUME_SIZE] -junction-path [JUNCTION_PATH] \
                         -aggr-list [AGGR_NAME] -aggr-list-multiplier 8 \
                         -policy [EXPORT_POLICY] \
                         -files-set-maximum true [-supporttiering true]
```

Large file considerations

The `aggr-list-multiplier` specifies the number of constituent volumes per node or aggregate pair. Eight-constituent volumes are recommended for EDA workloads. If an 80TB volume is provisioned, each constituent will be 10TB in size. A large file is any file that is 1% to 5% of the member constituent volume. In this example, a file large is any file larger than (80TB/8 x 5%) or 0.5TB (500GB). In general, it is better to provision larger thin-provisioned FlexGroup volumes than smaller volumes.

Mount command for read optimized tool and library volumes

The following NFS v3 mount command is recommended to maximize server-side caching using the nocto and actimeo options. Aggressive use of nocto and actimeo values can reduce EDA's intensive metadata I/O on the filer by as much as 80-90% in EDA workflows.

Review and apply this best practice according to the EDA flow requirements. Overly aggressive server-side caching can lead to slow cache updating issues.

```
%> mount -t nfs \
         -o "nocto,actimeo=600,hard,rsize=262144,wsize=262144,vers=3,tcp,mountproto=tcp" \
         <ontap data lif ip>:/<volname> /mnt/<volname>
```

Mount command for optimized mixed read/write performance on scratch volumes

The following NFS v3 mount command is recommended for EDA scratch volume workloads.

```
%> mount -t nfs \
         -o "hard,rsize=262144,wsize=262144,vers=3,tcp,mountproto=tcp" \
         <ontap data lif ip>:/<volname> /mnt/<volname>
```

Improving tool and library scale-out performance

Tool and library volumes are often shared by many different workflows, projects, or even business units, resulting in a high number of connected Linux clients that can in turn lead to read-latency bottlenecks.

I/O and latency bottlenecks can be minimized by using multiple FlexCache volumes per tool and library volume and then using DNS load balancing of server mounts to improve volume loading.

See the next chapter for more details on how to scale tool and library performance.

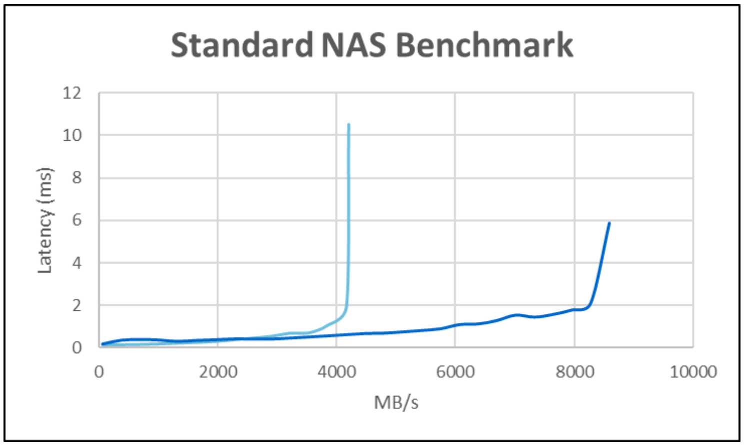
Performance monitoring

Performance monitoring is critical for EDA workloads. The single most important metric is volume latency. SSD-based Cloud Volumes ONTAP file system should provide <2ms latencies. Latencies greater than 3ms are an indicator that the Cloud Volumes ONTAP file system and volumes are under heavy I/O workloads.

High latency conditions can occur to high I/O loads from EDA jobs running in parallel. Cloud Volumes ONTAP file systems might not have the same I/O profile as an on-premises ONTAP appliances like a NetApp A800. NetApp recommends that EDA workloads running on Cloud Volumes ONTAP be tested at scale to determine the I/O limits.

The following graph shows an NAS workload being pushed to its performance limits. The knee of the performance curve is the point at which the controller is operating outside of its performance profile and latencies start to increase rapidly. Operating in this state for long periods of time affects EDA job runtime performance (wall clock).

Figure 7) Standard NAS benchmark.



Note: This graph is a generic NAS performance profile graph for illustration purposes only and does not represent actual Cloud Volumes ONTAP performance.

Quick latency monitoring with ONTAP CLI

There are multiple ways to measure real-time ONTAP performance. During a POC, one way is to simply report latency using the ONTAP CLI. The following command runs until Ctrl-C is pressed to stop reporting. Start the command either before or during EDA job submission to check the real-time volume latency. Extended periods of latency greater than 3ms could be signs of a heavily loaded storage file system.

```
ONTAP cli> qos statistics volume latency show
```

Other approaches to monitoring and reporting real-time latency

- Performance Manager within Cloud Manager
- Cloud-specific monitoring

Capacity reporting

To report on the size and utilization of a FlexCache volume, use the `volume show -fields <fields to report>` command from the ONTAP CLI. There are RESTful equivalents as well.

Here is an example with a few fields populated. You can see in the following example that the awscvo `proj` volume is a FlexCache instance containing the on premises `proj` source volume.

```
ONTAP::> volume show -volume <volname> -fields size,used,available,percent-used,files-used,files
```

In the following example, the on-premises `proj` volume is a FlexGroup volume and the awscvo `proj` volume is a FlexCache volume. Some of the files in the awscvo cache have been read (pre-warmed).

```
onPrem::> volume show -volume proj -fields size,used,available,percent-used,files-used,files
vserver    volume size available used    percent-used files    files-used
---------- ------ ---- --------- ------- ------------ -------- ----------
svm_onPrem proj   1TB  46.47GB   15.85GB 1%           31876696 226372

awscvo::> volume show -volume proj -fields size,used,available,percent-used,files-used,files
```

```
vserver    volume size  available used   percent-used files   files-used
---------- ------ ----- --------- ------ ------------ ------- ----------
svm_awscvo proj   150GB 147.0GB   2.96GB 1%           4669368 792
```

Starting and stopping Cloud Volumes ONTAP instances

You can stop and start Cloud Volumes ONTAP from BlueXP to manage your cloud compute costs. This might be particularly useful during proof-of-concept (POC) testing.

There are multiple ways to accomplish start/stop operation.

1. **BlueXP UI.** Menu options enable manual start/top operation.
2. **BlueXP UI.** Prescheduled start/stop operation.

   You might want to shut down Cloud Volumes ONTAP during specific time intervals to lower your compute costs. Rather than do this manually, you can configure BlueXP to automatically shut down and then restart systems at specific times.
3. **BlueXP APIs.**

For more information, see the documentation.

QuickStart implementation details

This section provides details and explanations for why each option was specified and the thought process behind the best practices for EDA workloads. In some cases, the explanation provides trade-offs to consider when applying these recommendations to specific design workflows or customer requirements.

Cloud Volumes ONTAP instance type

This section discusses the trade-offs between single vs HA pair configurations and when to use each.

Cloud Volumes ONTAP uses a cloud-based share-nothing architecture

It is important to understand that Cloud Volumes ONTAP uses a cloud-based share-nothing architecture that is very different than a traditional on-premises AFF or FAS storage appliance in a customer datacenter. An AFF or FAS HA pair system has two independent CPU nodes within a single chassis connected directly to disk shelves. A write to an HA pair generates one write from one node to the shared disk. The two CPU nodes are both directly connected to the disk shelves such that, in the case of a node failure, the second node can take over the disk shelf and continue to serve data.

Cloud Volumes ONTAP is built with cloud compute and disk resources. Each CPU node has its own set of disks, and it does not share disks with the second node. A write operation to a Cloud Volumes ONTAP HA pair is acknowledge after both nodes accept the write. In the case of a node failure, data is served from the secondary node and its own disks. In this model, nodes do not share disks; a write must be written to both nodes.

Cloud Volumes ONTAP performance sizing considers cloud resource network entitlements, max uncached disk limits, and the cumulative entitlements of attached (virtual) disk capacity. Sizing compute node instances must be bandwidth matched to the total disk aggregate bandwidth limits to ensure no artificial bottlenecks are created. As of the writing of this document, the goal of Cloud Volumes ONTAP storage sizing is to provide as many IOPs as possible from the cloud resource configurations available, understanding that the capabilities of a software-defined architecture are defined (in part) by the entitlements of the infrastructure they consume and may not directly match top-end datacenter-class hardware appliances such as an A800 NetApp storage controller.

Reliability and durability considerations

For FlexCache tool and library volumes, HA instances provide improved reliability and durability for volumes that might be shared across multiple workflows, projects, and teams. Higher reliability and durability reduce the change that shared workflows, project, and teams are affected by a failure.

Because FlexCache tool and library volumes are primarily read only, the write performance of an HA configuration is less of a concern.

Note: Cloud Volumes ONTAP HA file system are the recommended choice for FlexCache tool and library volumes.

For durability optimized scratch and release volumes, it is up to the customer to assess the performance-versus-reliability-versus-cost trade-off requirements for their application. This is primarily a trade-off of potentially 2x faster TTT for parallel jobs versus more reliability.

Note: For mixed read/write performance-optimized scratch volumes, Cloud Volumes ONTAP HA instances are the best choice. The customer should discuss the trade-off of performance and cost for their specific workloads.

Cost considerations

Single-node Cloud Volumes ONTAP instances require only one compute node and can deliver as much as 2x the write performance at scale, which can lead to better TTT performance. HA nodes cost 2x that of a single-node file system and delivers approximately half of the write performance and yet provide the potential for better up time.

HA was chosen for FlexCache tool and library volumes because it is assumed that FlexCache volumes will be ~10% the capacity of the on-premises source volumes and are thus less expensive in terms of capacity costs. If cost is a concern, FlexCache tool and library volumes can be configured as single-node instances.

Added services

Data sense and compliance (optional)

Data sense and compliance are cyber security features. There has been no EDA-specific performance impact testing performed. Data sense and compliance does have some known performance overhead. NetApp recommends turning off this option to maximize performance.

If additional security is required, then test this feature both on and off to assess the effect on performance.

Important: Although the anti-ransomware feature has a minimal effect on performance in terms of throughput and peak IOPS, in certain situations, NetApp recommends limiting its configuration. Specifically, for workloads that are read intensive or for which the data can be compressed, NetApp recommends enabling the feature for fewer than 50 volumes per storage node. With workloads that are write intensive and for which the data cannot be compressed, limiting anti-ransomware to fewer than 20 volumes per storage node is suggested. In less ideal scenarios, anti-ransomware analytics might run less frequently for a given volume and general performance might be affected.

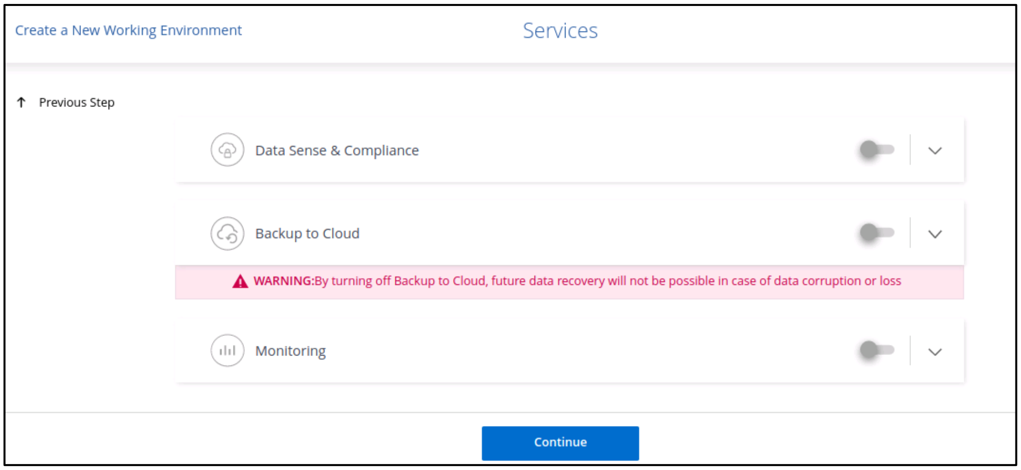
Back-up to cloud (optional)

Backup to cloud is a SnapMirror-to-cloud-as-a-service capability. The decision to enable or disable this feature should be made based on the type of Cloud Volumes ONTAP file system as follows. Back-up to cloud should have minimal performance effect, so enabling the feature should be discussed with the customer.

- **Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes**
  - Backup to cloud (optional) – turn OFF.
  - The tool and library Cloud Volumes ONTAP file system primarily serves FlexCache volumes, so there is no need for backup. The Source volumes, typically on-premises, should be backed up, not the FlexCache volumes.
- **Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes.**
  - Back to cloud (optional) – turn OFF.
  - Scratch volumes are high-change-rate volumes that contain rapidly changing design data and output from EDA tools. Most on-premises scratch volumes are not backed up. EDA design data is version controlled using tools like Perforce, ICManage, SOS, ClearCase, or Git. As such the critical design data is already safely backed up and versioned.
- **Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch and release volumes**
  - Back to Cloud (optional) – turn OFF.
  - This option can be enabled if needed based on customer consultation. Scratch volumes, specifically for release workloads, might require backup in case of disaster. Note that scratch volumes are typically FlexGroup volumes and are thus not currently supported.

Monitoring (optional)

Enabling this feature enables Cloud Insights monitoring. The performance effect of monitoring has not been assessed for EDA workloads. NetApp recommends turning OFF this feature during proof-of-concept testing. After you have assessed the flow and flow performance, you can enable this option after the fact to assess any performance effects.



WARNING:By turning off Backup to Cloud, future data recovery will not be possible in case of data corruption or loss

Select availability zone (AZ)

What is a single AZ versus a multiple AZ?

The traditional EDA semiconductor datacenter contains compute and storage in the same building. If there is a failure in the datacenter, compute and storage can be affected. This might be a result of a network outage, power outage, or any other issue. The EDA data stored on the NFS filers is typically not replicated in real time to secondary datacenters. Some EDA volumes might be replicated or backed-up to another datacenter for disaster recovery several times per day, but not continuously.

A single AZ is essentially the traditional datacenter as described above. The Cloud Volumes ONTAP filer resides completely within one AZ, so there is no additional latency associated with writing to the second AZ before the write is acknowledge.

A multiple AZ on the other hand is more like a NetApp MetroCluster instance. MetroCluster instances are two NetApp filers in two different datacenters separated by a distance (same city or adjacent city). Every write that comes in on one filer is immediately sent to the second filer. Only after both filers acknowledge the write is the acknowledgement sent back to the computer that initiated the write.

MetroCluster use is uncommon for EDA workflows. MetroCluster is typically used where there is no tolerance for outage and failover from one datacenter to another must be seamless, immediate, and completely fool proof.

AWS Multiple AZ is very similar to a MetroCluster. When a Cloud Volumes ONTAP HA pair is configured, one Cloud Volumes ONTAP file system is in one AZ (AWS datacenter) and one is in another AZ (another AWS adjacent datacenter). Writes to the primary node are propagated to the node in the other AZ. Write latencies are typically longer than single AZ configurations.

Note: Single-node Cloud Volumes ONTAP deployments have only one node, so they are single AZ by definition. HA Cloud Volumes ONTAP deployments have two nodes and can be configured either as single or multiple AZ configurations.

Recommendation for high performance EDA workloads

Because traditional EDA datacenters are essentially single AZs and multiple AZs increase write latency, NetApp recommends using single-AZ deployments for all EDA workflows.

This option is only relevant for Cloud Volumes ONTAP HA pair configurations.

- Cloud Volumes ONTAP optimized for maximum I/O for scale-out tool and library FlexCache volumes.

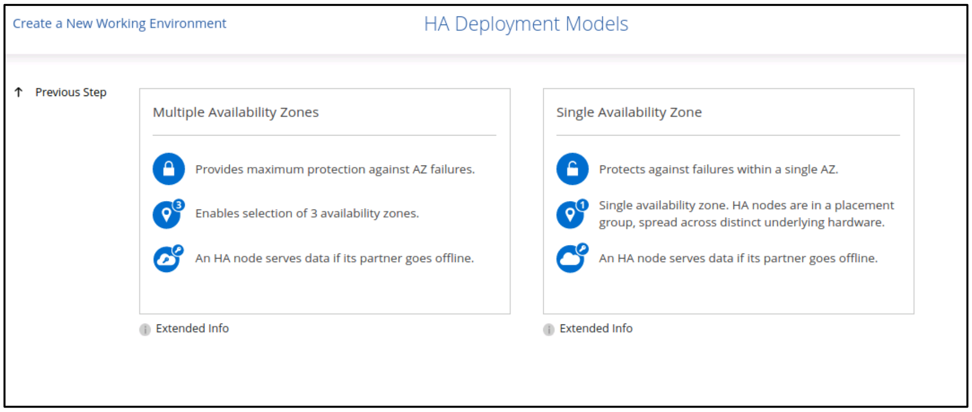
Select **Single Availability Zone**, which is best for FlexCache optimization of a Cloud Volumes ONTAP HA pair.

- Cloud Volumes ONTAP optimized for maximum I/O for scale-out scratch volumes.

  NetApp recommends a single-node Cloud Volumes ONTAP file system, so this option is not relevant.
- Cloud Volumes ONTAP optimized for high availability and maximum I/O and scale-out scratch and release volumes.

  Select **Single Availability Zone**.

A single AZ provides the best write performance in an HA configuration. No multiple AZ performance testing has been performed to evaluate this option.



Provides maximum protection against AZ failures.

Enables selection of 3 availability zones.

An HA node serves data if its partner goes offline.

Protects against failures within a single AZ.

Single availability zone. HA nodes are in a placement group, spread across distinct underlying hardware.

An HA node serves data if its partner goes offline.

Manage NetApp Support Site account

This option enables Cloud Volumes ONTAP file system communication and reporting to the NetApp Support Site (also known as Active IQ). Active IQ provides a single pane of glass for monitoring all NetApp products, performance, renewals, and so on.

BlueXP must also be registered for this feature to work. To register BlueXP, click the link at the top of the BlueXP UI.

No existing NetApp Support Site account.

Active IQ connectivity requires AWS networking configs to enable external communication. The following links help with the setup and configuration of Active IQ (the NetApp Support Site).

- https://docs.netapp.com/us-en/occm/reference_networking_aws.html
- https://docs.netapp.com/us-en/ontap/system-admin/setup-autosupport-task.html

Select charging method (Licensing)

Most if not all EDA accounts opt for the Essential by Capacity license feature versus the more expensive Professional by Capacity license. The only difference between the two licenses is that Professional includes the Back-up-to-cloud license enablement.

FlexCache tool and library Cloud Volumes ONTAP instances do not require backup. Back-up of tool and library volumes should be done on the source volume.

EDA scratch volumes are typically not backed up on-premises, so they are most likely not backed up in the burst-to-cloud use model. The design source code is version managed in tools such as Perforce Helix, ICManage, SOS, ClearCase, or Git tools. The EDA job output typically changes so fast due to typical design iterations that snapshots and backups are not necessary or meaningful.

Scratch volumes for HA scratch or release work might require backup to cloud. For POC work, backup to cloud is not required, and thus the Essential by Capacity license is the right option. For production (non-POC) environments, the customer should be consulted to decide if these volumes require backup to cloud. In this case, the customer should decide if Essential or Professional by Capacity is the right choice.

Cloud Volumes ONTAP sizing considerations for EDA workloads

Modern EDA workloads use 20k to 100k parallel compute cores on a single design. High parallelism, high file count, and high metadata workloads make optimizing for scale-out performance critical.

There is no single EDA workload. As many as 50+ different tools can be used to design chips, from the early design specification stage to design, timing, power and area optimization, and chip finishing. The following recommendations are focused on configuring Cloud Volumes ONTAP for maximum IOP, minimum latency, and maximum parallel compute performance.

Infrastructure cost is always a consideration, but engineer productivity and the use of expensive EDA design tool licenses are a much large expense than IT and cloud infrastructure costs. That said, after you determine your performance requirements, you can then assess the performance-versus-cost trade-offs.

EDA EC2 sizing considerations

- Highly parallel EDA workloads benefit from ONTAP systems with larger processor cores. Selecting EC2 instances with high core counts enable more parallel processing of read/write requests.
- EDA data has a very high file count and contains a mixture of small and large file sizes in deep directory structures. ONTAP performance benefits from having a large amount of system RAM.
- ONTAP performance significantly benefits from an EC2 instance that supports FlashCache (EC2 instances with local NVMe storage).
- Choose an EC2 instance that supports the maximum IOPS to the disk aggregate. If the EC2 instance has less IOPS performance than the aggregate, the EC2 instance limits disk performance.
- EC2 throughput limits should be divided in half for HA-node configurations. This is because every write is written to disk aggregates owned by the first node, but they are also written to the second node and its disk aggregate.

Storage systems can become node (or compute) limited before they become I/O (or storage medium) limited. As a result, it is critical to balance the frontend compute I/O bandwidth and backend storage aggregate bandwidth.

EC2 shared vs dedicated instances

It is an EDA best practice is to use shared EC2 instances because they are less expensive and have equivalent performance to dedicated instances.

What is the difference between dedicated and shared file systems?

Shared tenancy means that multiple EC2 instances from different customers can reside on the same piece of physical hardware. The dedicated model means that your EC2 instances are only run on hardware with other instances that you've deployed; no other customers use the same piece of hardware as you.

Dedicated instances

Dedicated instances are Amazon EC2 instances that run in a virtual private cloud (VPC) on hardware that's dedicated to a single customer. Dedicated instances that belong to different AWS accounts are physically isolated at a hardware level, even if those accounts are linked to a single payer account. However, dedicated instances might share hardware with other instances from the same AWS account that are not dedicated instances.

Dedicated hosts can incur higher costs than shared tenancy and don’t support burstable instance types.

Cloud Volumes ONTAP disk selection and throughput requirements

EDA disk performance requirements

All-flash (SSD-based) disk performance has demonstrated its value for demanding EDA and HPC workloads. Sub 1ms latencies have improved performance of EDA workload and has almost completely removed the need for performance-tuned storage, as was common practice for 10k SAS and SATA drives.

SATA drives are still common in EDA, but only for backup and cold data storage. Active workloads are almost exclusively on high performance All Flash storage platforms on-premises. The same is true for cloud-based NFS storage.

Matching disk-storage throughput performance with EC2 instance throughput is critical for maximizing disk performance.

Underlying disk selection

NetApp currently recommends GP3 disks, but there are other EBS options available, and more are being evaluated and qualified. In this document, we attempt to explain the thought process behind how EBS disks are chosen in a generalized way. As newer EBS disk options become available, the contents of this section may become dated. Use the following as an example.

Table 3) Disk selection.

	Durability	Volume size	Max IOPS per volume (16KiB I/O)	Max throughput per volume
GP3	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	1GiB - 16TiB	16,000	1,000MiBps
io1	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	4GiB - 16TiB	64,000	1,000MiBps

Source: AWS EBS Volume Types.

General purpose SSD (GP3)

General purpose SSD (GP3) volumes offer cost-effective storage that is ideal for a broad range of workloads. These volumes deliver a consistent baseline rate of 3,000 IOPS and 125 MiBps included with the price of storage. You can provision additional IOPS (up to 16,000) and throughput (up to 1,000 MiBps) for an additional cost.

The maximum ratio of provisioned IOPS to provisioned volume size is 500 IOPS per GiB. The maximum ratio of provisioned throughput to provisioned IOPS is .25 MiBps per IOPS.

The following volume configurations support provisioning either maximum IOPS or maximum throughput:

32GiB or larger: 500 IOPS/GiB × 32GiB = 16,000 IOPS

8GiB or larger and 4,000 IOPS or higher: 4,000 IOPS × 0.25MiBps/IOPS = 1,000MiBps

Source: AWS General Purpose SSD (GP3) Information.

Provisioned IOPS SSD (IO1)

Provisioned IOPS SSD volumes can range in size from 4GiB to 16TiB, and you can provision from 100 IOPS up to 64,000 IOPS per volume. IO1 disks can achieve up to 64,000 IOPS only on instances built on the Nitro system. On other instance families, you can achieve a performance of up to 32,000 IOPS. The maximum ratio of provisioned IOPS to requested volume size (in GiB) is 50:1 for io1 volumes. For example, a 100 GiB io1 volume can be provisioned with up to 5,000 IOPS.

On a supported instance type, the following volume sizes allow provisioning up to the 64,000 IOPS maximum:

IO1 volume: 1,280GiB in size or greater (50 × 1,280GiB = 64,000 IOPS)

Provisioned IOPS SSD volumes provisioned with up to 32,000 IOPS support a maximum I/O size of 256KiB and yield as much as 500MiBps of throughput. With the I/O size set at the maximum, peak throughput is reached at 2,000 IOPS. Volumes provisioned with more than 32,000 IOPS (up to the maximum of 64,000 IOPS) yield a linear increase in throughput at a rate of 16KiB per provisioned IOPS.

Example:

A volume provisioned with 48,000 IOPS can support up to 750 MiBps of throughput (16KiB per provisioned IOPS × 48,000 provisioned IOPS = 750MiBps). To achieve the maximum throughput of 1,000 MiBps, a volume must be provisioned with 64,000 IOPS (16KiB per provisioned IOPS × 64,000 provisioned IOPS = 1,000MiBps).

Source: AWS Provisioned IOPS SSD Information.

Sizing aggregates and disks - Performance

Each cluster is built on groups of EBS volumes that are presented to ONTAP as SCSI and then grouped into aggregates. Sizing aggregates and EBS volumes effectively is based on the performance of the EC2 instance chosen as well as the size of the workload per cluster.

NetApp best practices for performance recommend building aggregates with at least four disks in the aggregate to promote write parallelism to disk.

In our testing of EDA workloads, we found that the optimal number of disks is six per aggregate.

Sizing aggregates and disks - capacity

The next consideration is the size of each disk per aggregate based on the workload size and capacity overheads for ONTAP. The calculation is simple and has four components:

- Number of disks
- Active workload size
- Aggregate snapshot reserve
- WAFL overhead

The number of disks for EDA remains six, the active workload size depends on the workload and your individual cache size. The aggregate snapshot reserve size is set at 5% by default and WAFL overhead should be set at 12-14% of the total aggregate space.

An example of the calculation is as follows:

> Disk size = (workload size * (1 + ((aggregate snap reserve + WAFL overhead)/100))) / 6
>
> For example, if the active workload size is 4096GiB, disk size = (4096GiB * (1 + ((5+14)/100)))/6 = (4096 * 1.19)/6 = 4875/6 = 813GiB per disk

The goal of sizing the aggregates in this way is to saturate the IOPs and throughput of the EBS bandwidth based on the EC2 instance type so that the EBS bandwidth does not become a bottleneck for the workload.

Cost versus performance

The cost of storage is based on the disk type chosen, the size of the disk, and the provisioned IOPS and throughput on top of the default provided.

Table 4) Cost of storage versus performance.

	Capacity cost	Throughput cost	IOPS cost	Cost of m5d.12xlarge example
io1	$0.125/GB-month	N/A	$0.065 per provisioned IOPS-month	$610 + $468 = $1078/month

	Capacity cost	Throughput cost	IOPS cost	Cost of m5d.12xlarge example
GP3	$0.08/GB-month	125MBps free and $0.04/provisioned MBps/month over 125	3,000 IOPS free and $0.005/provisioned IOPS-month over 3,000	$390 + $18 = $408/month

Source: AWS EBS Pricing.

High Write Speed mode (HWSM)

High Write Speed mode reduces write latency by acknowledging the client's write operation as soon as data has been written to ONTAP memory (RAM). This is great for increasing write throughput and reducing write latency by up to 90%, which is extremely valuable for high-throughput and low-latency write workloads.

Normal Write Speed mode (the Cloud Volumes ONTAP default) acknowledges a write only after the data has been written to persistent disk and is no longer in volatile memory.

The trade-off is a significant performance gain with the potential for data loss (HWSM) versus lower write performance with no data loss (Normal Write Speed mode). How to choose?

This setting reduces the durability of data based on the time it takes for ONTAP to flush consistency points from memory into disk. This can be a risk with single-node systems used for production or persistent data that cannot be replaced from other sources. The risk is relatively low for high availability configurations however, because it would require both nodes in the pair to be lost within the maximum CP flush time of 10 seconds in order to have any data loss.

Post Cloud Volumes ONTAP creation setup

Modify aggregate setup details

Through the testing of EDA workloads, six disks per aggregate have been shown to be the optimal configuration for performance. In AWS, IOPs/throughput for each disk are aggregated to product the total IOPs/throughput value for the aggregate.

Example:

Aggregate with one disk = 1 * (Max IOPS per disk)

Aggregate with six disks = 6 * (Max IOPS per disk) <= Results in 6x more IOPS per aggregate.

Cloud Volumes ONTAP sggregates can use up to a maximum of six disks of the same size and type.

EC2 network performance MUST support (or match) the maximum aggregate IOPS. If not, then the EC2 throughput will be the bottleneck, not the aggregate.

Active-active vs active-passive

In Cloud Volumes ONTAP, there are two potential configurations for data access with an HA pair; active/passive and active/active.

All clusters created by BlueXP are by default active/passive and therefore have a single aggregate that is the home node (node 1). This means that all data is natively accessed through node 1 through the `data_1` LIF (logical interface), and node 2 is only used by clients during failover events.

Active/active refers to configuring of at least one aggregate on node 2 and also accessing data through the `data_2` LIF. This allows clients to take full advantage of both nodes in the HA pair for reads and local caching (FlashCache). However, if either of the nodes are utilized at more than 50%, then there will be a performance degradation during failover as well as the natural degradation of service for all traffic out of a single node with a single local cache.

Active/active can potentially improve performance. However, running in active/active mode means that you need be aware of the issues affecting failover, and thus you should run each node at 50% of performance utilization or less.

For more information, see What is a LIF (Logical Interface)?

NetApp recommends using active/passive configurations so that there is less performance degradation during a failover event, and only one namespace per ONTAP cluster is utilized.

Volumes provision (creating a new volume)

Cloud Volumes ONTAP instance for tool and library volumes

Create FlexCache volume(s) for tool and library volumes. More than one volume might be required depending on the number of on-premises source volumes that must be cached.

This link describes how to set up a cluster and SVM peering relationship between ONTAP clusters and how to provision a FlexCache volume.

FlexCache CLI command

```
ONTAP cli> volume flexcache create -origin-vserver [ORIGIN_SVM_NAME] -origin-volume [ORIGIN_VOLUME_NAME] \
                                   -vserver [DESTINATION_SVM_NAME] -volume [DESTINATION_VOLUME_NAME] \
                                   -size [VOLUME SIZE] -junction-path /[DESTINATION_VOLUME_NAME] \
                                   -aggr-list [AGGR_NAME] -aggr-list-multiplier 8 \
                                   -policy [EXPORT_POLICY]
```

Cloud Volumes ONTAP instance for highest performance scratch volumes

Create one or more FlexGroup volume for a scratch workload.

Based on SPECstorage 2020 EDA benchmark testing, the best FlexGroup volume performance is achieved when FlexGroup volumes use the following recommendations:

- FlexGroup volumes should be constrained to a single-node, single-aggregate, and single-LIF configuration. This provides cross node latency, and throughput performance does not limit overall volume performance.
- If you are provisioning multiple volumes, place the new volume on the least performance- and capacity-loaded node, aggregate, or LIF pair.
- Ensure volume capacity is sized to make sure that large-file issues are addressed.

FlexGroup CLI command

```
ONTAP cli> volume create -vserver [SVM_NAME] -volume [VOLUME_NAME] \
                         -size [VOLUME SIZE] -junction-path /[VOLUME_NAME] \
                         -aggr-list [AGGR_NAME] -aggr-list-multiplier 8 \
                         -policy [EXPORT_POLICY] \
                         -files-set-maximum true [-supporttiering true]
```

The `aggr-list-multiplier` specifies the number of constituent volumes per node or aggregate pair. NetApp recommends eight constituent volumes for EDA workloads. If an 80TB volume is provisioned, each constituent is 10TB in size. A large file is defined as any file that is 1% to 5% of the member constituent volume. In this example, a file large is any file larger than (80TB/8 x 5%) or 0.5TB (500GB). In general, it is better to provision larger thin-provisioned FlexGroup volumes than smaller volumes.

Increase iNode count

EDA environments might have millions of files; therefore, you might need to increase the inode value to the maximum level. NetApp introduced a new volume option in ONTAP 9.9.1 to make setting the maximum file value simpler.

When you set the **files-set-maximum** value on a volume to **true**, ONTAP automatically adjusts **maxfiles** to the largest possible value. Once set to **true**, the value cannot be reset. Set the value to **true** only if you want to set the maxfiles to the largest possible value.

Table 5) Inode defaults and maximums according to FlexVol size.

FlexVol volume size	Default inode count	Maximum inode count
20MB*	566	4,855
1GB*	31,122	249,030
100GB*	3,112,959	24,903,679
1TB	21,251,126	255,013,682
7.8TB	21,251,126	2,040,109,451
100TB	21,251,126	2,040,109,451

*FlexGroup member volumes should not be any smaller than 100GB in size.

Mount volumes

NFSv3 continues to be the standard for EDA workloads because it is more performant than NFSv4.x. The following two mount commands are for optimizing the read-only (or mostly) tool and library volumes and for providing the fastest possible performance for read/write scratch volumes.

See the section “Where to find additional information” for further guidance on optimizing NFS mounts for EDA/HPC workloads.

Mount command for tool and library volumes

The following script provides the recommended NFS mount option for tool and library volumes:

```
%> mount -t nfs \
         -o "nocto,actimeo=600,hard,rsize=262144,wsize=262144,vers=3,tcp,mountproto=tcp" \
         <ontap svm>:/<volname> /mnt/<volname>
```

To improve file and metadata loading times as well as decrease the load on the ONTAP storage system, NetApp recommends using aggressive compute-server caching for tool and library volumes that are mostly read-only in nature. The NFS mount options `nocto,actimeo=600` have been shown to reduce the metadata I/O of EDA workloads by as much as 90%. It can also dramatically improve the load time for tools and libraries, which is becoming an increasing challenge as semiconductor technology library sizes grow dramatically in number and size as the industry adopts 7nm, 5nm, and 3nm design nodes (transistor size).

EDA workflows often make heavy use of Makefile or Makefile-like dependencies tracking. As a result, it is not uncommon to see 60-80% of metadata I/O as the file system returns date and time stamps, file existence checks, and so on. In addition to Makefile-like dependency checks, EDA tools and scripting tools like Perl and Python often scan long include-file or directory search paths from modules to load. The result is additional metadata I/O transactions used in the search for files and directories. This can put an unnecessary load on the NFS filer.

Aggressive NFS server-side caching works based on the assumption that tools and library volumes are static or unchanging. New tools and libraries are typically installed in directories next to older versions and are seldom overwritten or modified. It is considered a poor practice for EDA workloads to pull in the latest libraries mid-analysis. EDA flows typically specify a set of tools and libraries for a given job, and then those versions are used throughout the entire run.

For more information on this subject, see the blog post What does actimeo mean during NFS mount in Linux?

Mount command for highest performance scratch volumes

The following mount option is optimized for high performance NFS read/write transactions. Unlike read-only tool and library volumes, aggressive file caching is typically not recommended.

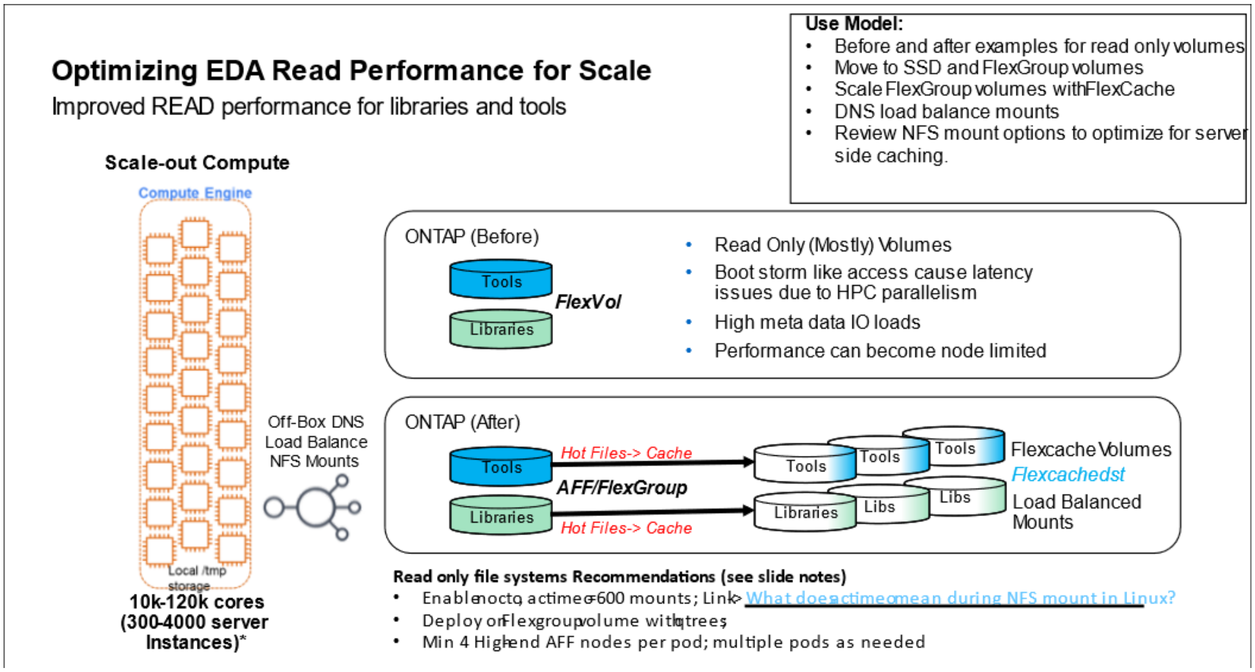
The following script is the recommended NFS mount option for scratch volumes.

```
%> mount -t nfs \
         -o "hard,rsize=262144,wsize=262144,vers=3,tcp,mountproto=tcp" \
         <ontap svm>:/<volname> /mnt/<volname>
```

DNS mount balancing for improved network loading

Scaling and DNS load balancing NFS mounts can improve read performance at scale. Multiple tool and library FlexCache volumes can be created and NFS mount distributed across all available FlexCached tool and library volumes.

Figure 8) Optimizing EDA read performance.



Optimizing EDA Read Performance for Scale

Improved READ performance for libraries and tools

Use Model:

- Before and after examples for read only volumes
- Move to SSD and FlexGroup volumes
- Scale FlexGroup volumes with FlexCache
- DNS load balance mounts
- Review NFS mount options to optimize for server side caching.

ONTAP (Before)

- Read Only (Mostly) Volumes
- Boot storm like access cause latency issues due to HPC parallelism
- High meta data IO loads
- Performance can become node limited

Read only file systems Recommendations (see slide notes)

- Enable nocto, actimeo=600 mounts; Link-> What does actimeo mean during NFS mount in Linux?
- Deploy on Flexgroup volume with qtrees
- Min 4 High-end AFF nodes per pod; multiple pods as needed

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- TR-4617: Electronic Design Automation best practices
  https://www.netapp.com/media/19368-tr-4617.pdf
- Cloud Volumes ONTAP Documentation on-line
  https://docs.netapp.com/us-en/cloud-manager-cloud-volumes-ontap/task-deploying-otc-aws.html#launching-a-single-node-cloud-volumes-ontap-system-in-aws
- TR-4383: Performance Characterization of NetApp Cloud Volumes ONTAP for Amazon Web Services
  https://www.netapp.com/pdf.html?item=/media/9088-tr4383pdf.pdf
- TR-4571: NetApp ONTAP FlexGroup volumes Best practices and implementation guide
  https://www.netapp.com/pdf.html?item=/media/12385-tr4571pdf.pdf
- TR-4067: NFS in NetApp ONTAP Best practice and implementation guide
  https://www.netapp.com/pdf.html?item=/media/10720-tr-4067.pdf
- Cloud Volumes ONTAP Sizing YouTube Video
  https://www.youtube.com/watch?v=GELcXmOuYPw

Version History

Version	Date	Release notes
1.0	November 2022	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright information

Trademark information

TR-4946