SoftNAS Deployment Guide for High-Performance SaaS

Introduction

The goal is to provide the SoftNAS best practices for setting up SoftNAS for High-Performance SaaS.

High-performance is necessary for a SaaS solution, and since all SaaS solutions require data, you must have a high-performance storage solution. To achieve the best performance for cloud storage, you must use a Cloud NAS. A cloud NAS offers high-performance storage access by many clients, servers, or applications.

Top reasons to choose Buurst SoftNAS

1. SoftNAS enables businesses to utilize the massive amount of storage available in the cloud. SoftNAS gives you total control and overall flexibility of data and the associated costs and replicates data across availability zones.
2. High-performance, tier 1 and 2 caching
3. High availability across availability zones
4. Deduplication & Compression
5. Industry-standard file-sharing protocols including iSCSI, NFS, and CIFS
6. Scheduled snapshots via copy-on-write filesystem (ZFS)
7. Thin provisioning
8. Block replication through SoftNAS patent SnapReplicate
9. Data integrity through built-in error detection and correction
10. Cloud storage RAID

SoftNAS Best practices and guidelines

Your business needs price, availability, performance, scalability, and other requirements when picking a cloud storage solution. You need to balance your current and future needs or goals against your budget.

The best approach is first to define your business requirements. What kind of data do you have, what type of access do you need it to have and to whom, when and where, how long do you need it to be available, what are the performance requirements for your data?

Instance Size and Storage

SoftNAS has simplified selecting your instance size by defining performance categories. By matching the performance categories below to your organization's use case, you can make an educated choice. The critical thing to remember is that your highest requirement should always govern your decision. For example, if your use case requires deduplication, encryption, or compression, your primary consideration will likely be processing power (vCPU). If using a large amount of storage, RAM might be the primary consideration.

https://www.softnas.com/wp/products/instance-size-recommendations/

SoftNAS can leverage both block storage (such as the varieties of EBS volumes available on AWS) and object storage such as S3. This flexibility is one of the critical features of SoftNAS. It can, however, introduce some complexities when determining your requirements. As performance is based not only on the instance size selected but also on the storage characteristics, leveraging different storage types in the same solution will affect your performance.

General Purpose SSD

General-purpose SSD volume that balances price and performance for a wide variety of transactional workloads

- Volume Size: 1 GiB - 16 TiB
- Max IOPS / Volume: 10,000
- Max. Throughput/Volume: 160 MiB/s
- Max. IOPS/Instance: 65,000
- Max. Throughput/Instance: 1,250 MiB/s
- Dominant Performance Attribute: IOPS

Provisioned IOPS SSD

Highest-performance SSD volume designed for mission-critical applications

- Volume Size: 4 GiB - 16 TiB
- Max IOPS / Volume: 20,000
- Max. Throughput/Volume: 320 MiB/s
- Max. IOPS/Instance: 65,000
- Max. Throughput/Instance: 1,250 MiB/s
- Dominant Performance Attribute: IOPS

Throughput Optimized HDD

Low-cost HDD volume designed for frequently accessed, throughput-intensive workloads

- Volume Size: 500 GiB - 16 TiB
- Max IOPS / Volume: 500
- Max. Throughput/Volume: 500 MiB/s
- Max. IOPS/Instance: 65,000
- Max. Throughput/Instance: 1,250 MiB/s
- Dominant Performance Attribute: MiB/s

Cold HDD

Lowest cost HDD volume designed for less frequently accessed workloads

- Volume Size: 500 GiB - 16 TiB
- Max IOPS / Volume: 250
- Max. Throughput/Volume: 250 MiB/s
- Max. IOPS/Instance: 65,000
- Max. Throughput/Instance: 1,250 MiB/s
- Dominant Performance Attribute: MiB/s

Simple Storage Service (S3)

AWS object storage, known as Simple Storage Service, or S3, has its characteristics. S3 serves as a repository for Internet data. It provides access to reliable, fast, and inexpensive data storage infrastructure. It is designed to make web-scale computing easy by enabling you to store and retrieve any amount of data, at any time, from within Amazon EC2 or anywhere on the web. Amazon S3 stores data objects redundantly on multiple devices across multiple facilities and allows concurrent read or write access to these data objects by many separate clients or application threads. You can use the redundant data stored in Amazon S3 to recover quickly and reliably from instance or application failures.

SoftNAS can leverage S3 Storage within its infrastructures, creating disks, pools, and volumes, much as you would with EBS block storage. It offers similar performance characteristics to General Purpose SSD. However, this performance cannot be improved by stacking disks into a RAID configuration.

S3 Cloud Disk Best Practices

Without proper configuration, a SoftNAS instance leveraging S3-compatible cloud disk extenders can perform poorly. To get the best performance possible for a SoftNAS deployment with S3-compatible cloud disks, keep in mind the following:

Sizing

Sizing a solution involving Cloud Disk Extenders is very much like a solution using a block-based implementation (VMDK or EBS). There is no change in storage space requirements. However, additional system resources may be required to handle the S3-compatible storage's virtualization necessary to present the S3 Cloud Disk as block storage. Stated another way, the number of buckets configured via cloud disk extender influences the number of additional resources required to access the same overall storage capacity.

CPU

If using cloud disk extenders in your instance/s, it is essential to configure your instance with additional processing power (CPU), above and beyond what is required for traditional block-based storage access. Presenting S3 storage as block-based storage requires several other functions to be executed, including, for example, SSL/TLS key exchange and encryption, MD5 block computations, network stack processing, as well as optional encryption options. To avoid performance issues:

- Do not use cloud disk extender on single vCPU instances.
- 4 vCPU instances may be suitable for test scenarios. Four vCPU instances may still prove insufficient if your S3-compatible test/POC environment requires decent performance metrics.
- For a production environment, a minimum of 4 vCPU instances is highly recommended. Many workloads will perform better with additional vCPU.
- For each 75 MB/s of throughput required to perform the same task with block-based storage, an additional two vCPU is highly recommended.
- CPU utilization should be monitored during proof-of-concept and initial production stages to verify that sufficient CPU has been provisioned for the provided workload.
- Email alerts should be monitored, and high CPU utilization indications should be reviewed for the Cloud Disk Extender configuration.

- If operating in a trusted environment and available as an option for the S3-compatible object storage being used, CPU usage can be reduced by using HTTP rather than https.
- CPU usage can be further reduced by disabling optional encryption options.

Example: A customer wants to use S3 object storage to save money over EBS. The current workload operates between 100-150MB/s of throughput and is running on an m4.xlarge instance. Evaluating the current workload, we know that it averages a healthy 50% CPU usage. To provide the same 150MB/s of S3 throughput, the general guideline requests four additional vCPU over and above the current instance's existing four vCPU base. As a result, the CPU recommendation points to an m4.2xlarge instance to provide four additional vCPU.

RAM

As mentioned previously in this document, each instance of the cloud disk extender represents a running process inside the SoftNAS instance for virtualizing the object storage as block storage.

- Cloud Disk Extender should not be used in production on systems with less than 16 GB of RAM.
- A general guideline of 512MB of RAM should be provisioned above the typically required memory for a given workload.
- Remember that half of the RAM is utilized for filesystem caching. Additional resources are needed for the network file services and the base operating environment (~2GB of RAM).

Network

Cloud Disk Extender utilizes the network interface of an instance to access the object storage. Sufficient network bandwidth must be provisioned to reach maximum performance profiles using Cloud Disk Extender. When considering the desired available throughput to the object store, consider the amount of network throughput for network file services (NFS, CIFS, iSCSI, AFP) and SnapReplicate™/SNAP HATM, which, in most configurations and platforms, all come from the same pool of available network bandwidth.

A somewhat safe calculation can determine the available network throughput being used for the instance and divide it divided by 3 to calculate 1/3 for file services, 1/3 for replication, and 1/3 for object storage I/O.

- When calculating, consider that SnapReplicate™ only replicates the write bandwidth, not the read bandwidth.

- Be sure to convert properly between bits and bytes when comparing network throughput (usually expressed in bits) to disk throughput (traditionally expressed in bytes)
- There is inherent overhead in the protocols used on the network (request/response, headers, checksums, control data, etc.) such that full network saturation does not yield the full bandwidth as useful throughput. Consider only anticipating 90% of the link-speed as usable throughput.
- Most clouds (and most data centers) do not provide full link-speed bandwidth on a sustained basis as systems utilize shared resources. Systems designed to run at maximum provisioned capacity (of any metric) should be assigned to dedicated hosts rather than a shared tenancy.

Example: A customer uses NFS, SnapReplicate™, and SNAP HATM and would like to use object storage. Expected throughput is about 40MB/s, with 90% reads. According to the calculation, the network throughput for the source node reads as follows:

- 4MB/s writes to NFS (incoming)
- 36MB/s reads to NFS (outgoing)
- 4MB/s writes to SnapReplicate (outgoing)
- 4MB/s writes to Object Storage (outgoing)
- 36MB/s reads to Object Storage (incoming)
- Total: 40MB/s incoming 44MB/s outgoing

Calculating the total throughput in bytes, this is 320mbps incoming and 352mbps outgoing.

According to the calculation, the network throughput for the target node reads as follows:

- 4MB/s writes from SnapReplicate (incoming)
- 4MB/s writes to Object Storage (outgoing)
- **Total:** 4MB/S incoming and 4MB/S outgoing
- In bytes, this works out to 32mbps incoming 32mbps outgoing.
- A 100 Mbps network connection is certainly not sufficient for this configuration. However, a 1gbps connection should be enough, even considering protocol overhead and avoiding 100%

VPC Endpoints

Customers on AWS within a VPC should be using VPC Endpoints for accessing S3 object stores. Using a VPC endpoint, a higher quality service level is provided to S3 object stores within a region, thereby improving the overall reliability and performance when accessing

S3 object storage. Additionally, a VPC Endpoint can communicate with resources in other services via private IPs, without exposing instances to the Internet.

Azure Block Storage

The SoftNAS product can leverage both block storage (such as the varieties of SSD and HDD disks available under General Purpose storage accounts on Azure) and object storage such as Azure Blob Hot or Cool storage. This flexibility is one of the critical features of SoftNAS. It can, however, introduce some complexities when determining your requirements. As performance is based not only on the instance size selected but also on the storage characteristics, leveraging different storage types in the same solution will affect your performance.

To determine what this effect might be, we must first understand each storage type's performance characteristics.

There are numerous distinctions and abstractions in Azure storage, confusing the layperson in choosing the right storage option for them. However, when using SoftNAS, the simplest way to determine the storage option for you is based on two factors – the storage account, which determines the type of storage available, and the storage type itself, block or object storage. The below information will help you understand Azure block and object storage related to your SoftNAS Instance.

- Block storage provides a fixed-size raw storage capacity. Each storage volume can be treated as an independent disk drive and is only accessible when attached to an OS. It is typically formatted with a file system, such as FAT32, NTFS, EXT3, or EXT4.

The storage account determines the type of storage provided. Block Storage, or **General Purpose**, is further divided by account type, Standard, or Premium. (Blob Storage is limited to one category, Standard.)

Standard Storage Accounts

- Standard Storage Accounts are based on magnetic drives and are an affordable solution for applications or other use cases in which the underlying data is accessed infrequently.

Premium Storage Accounts

Premium Storage is backed by solid-state drives and offers consistent, low latency performance. It is the recommended option for any application in which data must be retrieved quickly and often.

Standard Storage Accounts provide disks with a single performance metric, with per disk limits.

- Max Disk Size: 1023 GB
- Max 8K IOPS per disk: Up to 500
- Max Bandwidth per disk: Up to 60 MB/s

On the other hand, with Premium Storage, you are offered three disk types, corresponding to 3 different disk sizes. When creating a disk for your SoftNAS instance using a Premium Storage account, you will only be able to specify one of the following sizes: 128GB, 512GB, and 1024GB. Whether you create the disk from the Azure portal or type the disk size within the Add Device wizard in the SoftNAS UI, you select the corresponding disk type. The performance characteristics in the table below.

Premium Storage Account	P10	P20	P30
IOPS per disk	500	2300	5000
IOPS per disk	100 MB/s	150 MB/s	200 MB/s
Throughput per disk	128 GB	512 GB	1024GB

Azure Blob Storage: Hot and Cool**

If deciding to add Azure object storage (otherwise known as Blob storage), you will need to have a Blob storage account set up, or you will not be able to call upon the storage within the SoftNAS UI. When creating your Blob Storage account, you will also have another decision to make - whether you will leverage hot or cool storage for Azure. Buurst offers full support for both options:

Azure Cool Storage

Object storage allows economical, safe-keeping of less frequently accessed file data.

Azure Hot Storage

Object storage that optimizes frequently accessed stored data to enable continuous IO.

Note: You cannot mix hot and cool storage disks in a RAID configured pool. A decision must be made on the storage type for each pool. As storage type is determined at the blob storage account level, you must be aware of the type of account created. Buurst recommends labeling them with Hot or Cool in the names to avoid confusion.

Region and Availability Zones

Amazon EC2 allows placement of instances in multiple locations. Your instance can be placed in a region, which corresponds roughly to a physical location or area, such as US East (N. Virginia), Canada (Central), EU (Ireland), and so forth. There are numerous options to choose from. Alternatively, it can be placed in an Availability Zone, essentially a logical grouping of servers or racks of servers separated into an artificial 'location' within a given region. Instances paired across Availability Zones are isolated from failures in other Zones, ensuring redundancy in case of large-scale failures.

Software RAID Considerations

SoftNAS provides a robust set of software RAID capabilities for non-durable disk drives when there is no hardware protection. Software RAID is best used in scenarios where raw disk devices are attached directly to SoftNAS. Software RAID is **not** recommended for object storage (S3), AWS EBS Volumes, and disks behind hardware RAID controllers. There are use cases in which RAID 0 (and nothing beyond RAID 0) can provide some redundancy and performance benefits. Consult with SoftNAS support before deciding to leverage software RAID 0.

Software RAID options include RAID 1 and RAID 10 mirrors, RAID 5 (single parity), RAID 6 (dual parity), and even RAID 7 (triple parity) support. It also includes hot spare drive capabilities and the ability to hot-swap spares into operation to replace a failed drive. RAID 10 (striped mirrors) and RAID 6 (dual-parity) are generally recommended for the best balance of read/write I/O performance and fault tolerance. Use RAID 10 for the most performance-sensitive storage pools (e.g., SQL Server, Virtual Desktop Server) and RAID 6 for high-capacity, high-performance applications (e.g., Exchange Server) as it provides the highest write IOPS.

SoftNAS is atop the ZFS filesystem. Please take a few moments to become familiar with ZFS Best Practices for more details on a storage pool, RAID, and other performance, data integrity, and reliability considerations.

Considerations for RAID Level 10:

- Minimum 4 disks.
- This is also called a "stripe of mirrors."
- Excellent redundancy (as blocks are mirrored)
- Excellent performance (as blocks are striped)
- For higher operating budgets, RAID 10 is the BEST option for any mission-critical applications (especially databases).

Best Practices for ZFS RAIDz

Note: Do not use RAIDz1 for disks 1TB or greater in size (use RAIDz2/3 or mirroring instead for better protection)

- Mirrors trump RAIDz every time. Far higher IOPS result from a RAID10 mirror pool than any RAIDz pool, given an equal number of drives. This is especially true when using raw disks in situations requiring high write IOPS (typical of VM workloads).
- For 3TB+ size disks, 3-way mirrors begin to become more and more compelling
- Never mix disk sizes (within a few %) or speeds (RPM) within a single vdev
- Never mix disk sizes (within a few %) or speeds (RPM) within a zpool, except for L2ARC & ZIL devices
- Never mix redundancy types for data vdevs in a zpool (use all RAID10 mirrors, RAIDz2, etc. instead of mixing redundancy types)
- Never mix disk counts on data vdevs within a zpool (if the first data vdev is 6 disks, all data vdevs should be 6 disks)
- With multiple JBODs, try to spread each vdev out so that the minimum number of disks are in each JBOD. Given enough JBODs for the chosen redundancy level, it is possible to end up with no SPOF (Single Point of Failure) in the form of JBOD. If the JBODs themselves are spread out amongst sufficient HBAs, it becomes possible to even remove HBAs as a SPOF.
- Use RAIDz2/3 over RAIDz1 plus a hot spare, because increased redundancy provides better data protection (and RAIDz3 is like having online hot spares since it can sustain 2 drive failures).

Windows Workloads

One approach that works well for a broad range of applications is to use a combination of SAS and SATA drives - using SSD for read cache/write log (always configure write logs as mirrored pairs in case a drive fails). SATA drives provide very high densities in a relatively small footprint, perfect for user mass storage, Windows profiles, Office files, MS Exchange, etc. SQL Server typically demands SAS and/or SSD for the best results due to the high transaction rates. Exchange can be relatively heavy on I/O when it's starting up, but since it reads most everything into memory, high-speed caching does little to help run-time performance after initial startup.

Active Directory

The integration of SoftNAS into Active Directory enables domain users to more securely share files and data in a corporate environment. Authentication is managed by Active Directory (AD) via Kerberos. Kerberos tickets are issued to users authenticated to AD. When a user accesses a CIFS share managed by SoftNAS, the ticket is verified with AD to

ensure it is authentic and valid before allowing access to the shares. Windows user IDs and groups (e.g., Domain Users) are transparently and dynamically mapped from AD into SoftNAS and Linux, making access seamless for Windows users.

Security

- Change the default password.
- Apply the latest software updates.
- Restrict the firewall source.

SoftNAS general performance principles

Cloud-based Deployments

Note: Do not use local SSD or ephemeral disks attached directly to an instance for the write log, as these instance devices are not guaranteed to be available again after reboot. Instead, use volumes with Provisioned IOPS for the Write Log (it's okay to use local SSD devices for Read Cache).

Disk Controller Considerations

There are several ways to get the most performance from these cache devices by following a few disk controller best practices:

Pass-through Controller

In this configuration, the disk controller is passed through to the SoftNAS® VM. Pass-through enables SoftNAS® OS to interact with the disk controller directly. This provides the best possible performance but requires CPUs and motherboards which support Intel VT-d and disk controllers supported by CentOS operating system.

Note: For servers with the disk controller built into the motherboard, it is common to install a virtual platform and then boot from USB, freeing up the disk controller for pass-through use.

PCIe Flash Cache Cards

There are flash memory plug-in cards with high-speed NAND memory available in PCIe form. These make high-speed memory available at high speeds through the PCIe bus. Be sure to choose a PCIe flash memory card that is supported by the hardware's virtualization vendor.

Raw Device Mapping

Some SSD devices can be mapped directly to the SoftNAS® VM using Raw Device Mapping (RDM). Raw device access allows SCSI commands to flow directly between the SoftNAS CentOS operating system and the SSD device for peak cache performance and IOPS and reduce context-switching between the SoftNAS® VM running CentOS and the virtualization host.

Disk controller pass-through is preferred to RDM on systems with processors and configurations that support it.

Disk Speed and RAID

Virtual Devices and IOPS

IOPS (I/O per second) is mostly a factor of the number of virtual devices (vdevs) in a zpool. They are not a factor of the raw number of disks in the zpool. This is probably the most critical thing to realize and understand and is commonly not. A vdev is a "virtual device." A Virtual Device is a single device/partition that acts as a source for storage on which a pool can be created. For example, in VMware, each vdev can be a VMDK or raw disk device assigned to the SoftNAS® VM.

A multi-device or multi-partition vdev can be in one of the following shapes:

- Stripe (technically, each chunk of a stripe is its vdev)
- Mirror
- RaidZ
- A dynamic stripe of multiple mirrors and/or RaidZ child vdevs
- ZFS stripes writes across vdevs (not individual disks). A vdev is typically IOPS bound to the speed of the slowest disk within it. So if with one vdev of 100 disks, a zpool's raw IOPS potential is effectively only a single disk, not 100.

Deduplication

A common misunderstanding is that ZFS deduplication is free, enabling space savings on a ZFS filesystems/zvols/zpools. In actuality, ZFS deduplication is performance on-the-fly as data is read and written. This can lead to a significant and sometimes unexpectedly high RAM requirement.

Every block of data in a deduplicated filesystem can end up having an entry in a database known as the DDT (DeDupe Table). DDT entries need RAM. It is not uncommon for DDTs to grow to sizes larger than available RAM on zpools that aren't even that large (a couple of TBs). If the hits against the DDT aren't being serviced primarily from RAM (or fast SSD configured as L2ARC), performance quickly drops to abysmal levels. Because enabling/disabling deduplication within ZFS doesn't do anything to the data already

committed on disk, it is recommended not to enable deduplication without fully understanding its RAM and caching requirements. It may be challenging to get rid of later after many terabytes of deduplicated data are already written to disk, and suddenly the network needs more RAM and/or cache. Plan cache and RAM needs around how much Total deduplicated data is expected.

Note: A general rule of thumb is to provide at least 2 GB of DDT per TB of deduplicated data (actual results will vary based on how much duplication of data is required).

Please note that the DDT tables require RAM beyond whatever is needed for caching of data, so be sure to take this into account (RAM is very affordable these days, so get more than may be necessary to be on the safe side).

Extremely Large Destroy Operations - When destroying large filesystems, snapshots, and cloned filesystems (e.g., in excess of a terabyte), the data is not immediately deleted is scheduled for background deletion processing. The deletion process touches many metadata blocks, and in a heavily deduplicated pool, must also look up and update the DDT to ensure the block reference counts are correctly maintained. This results in a significant amount of additional I/O, which can impact the total IOPS available for production workloads.

For best results, schedule large destroy operations for after business hours or on weekends so that deletion processing IOPS will not impact the IOPS available for regular business day operations.

SoftNAS features for better performance

- As with any storage system, NAS performance is a function of many different combined factors:
- Cache memory (the first level read cache or ARC)
- 2nd level cache (e.g., L2ARC) speed
- Disk drive speed and the chosen RAID configuration
- Disk controller and protocol

Cache Memory (first level)

Solid-state disk (SSD) and PCIe flash cache cards offer high-speed read caching and transaction logging for synchronous writes. However, not all SSDs are created equal, and some are better for these tasks than others. In particular, pay close attention to the specifications regarding 4K IOPS.

For read caching (L2ARC), both read and write IOPS matter, as do the device's sequential throughput specifications. For running a database, VMware VMDK, or other workloads

that produce large amounts of random, small (e.g., 4KB) reads and writes, ensure the SSD and flash cache devices provide high IOPS for 4K reads/writes.

For the write log (ZIL), extremely fast write IOPS is most important (the ZIL is only read after a power failure or other outage event to replay synchronous write transactions that may not have been posted before the outage, so write IOPS is most critical for use as a ZIL). ZFS always uses a ZIL (unless the variable set "sync=disabled"). By default, the ZIL uses the devices which comprise the storage pool. An "SLOG" device (called a "Write Log" in SoftNAS®) offloads the ZIL from the main pool to a separate log device, which improves performance when the right log device is chosen and configured correctly.

Snapshot

Buurst recommends the best practice of creating a snapshot of your current machine image state before applying any updates. This should allow you to roll back to our previous SoftNAS configuration/version by restoring the snapshot if necessary. The below links will help you to create the required snapshots.

Platform-Specific Notes

Amazon Simple Storage Service (S3)

- Only use S3 Cloud Disks in the same region as the EC2 instance.
- Always utilize VPC Endpoints to directly access S3 storage without contention through the public Internet.

Transactional Pricing

Amazon S3 (and many other object storage providers) have a multi-faceted approach to object storage pricing. While capacity is one component of cost, there are also charges for requests to create new objects and access existing objects. These transaction charges can add up that the perceived cost savings of S3 vs. EBS are non-existent or even become expenses rather than savings. A SoftNAS Solutions Architect can assist customers in evaluating if S3-compatible storage is appropriate for specific applications.