REFERENCE ARCHITECTURE

MINIO OBJECT STORAGE USING SEAGATE APPLICATION PLATFORMS

MINIO

SEAGATE

CONTENTS

1 INTRODUCTION

2 ARCHITECTURE

3 TEST PROCESS

4 TEST RESULTS

5 TEST PROCESS DETAILS

6 CONCLUSION

Introduction

Amazon's Simple Storage Service (S3) object protocol has evolved beyond its humble beginnings as a Web 2.0 application programming interface (API) for Web 2.0 applications back in 2006. The design philosophy transcended the more conventional silos of hierarchical file systems with the notions of objects referenced by location-independent URLs and object IDs. S3 also brought with it web technologies to secure data access as a built-in attribute rather than a retrofitted improvement. Today, we find the S3 protocol being leveraged across a broad spectrum of industries and services to provide location-independent networked storage and distribution of services.

While leveraging the S3 API is relatively easy, the implementation and deployment of S3 object storage has posed significant engineering challenges for both adopters and implementers—ones that often require extensive integration testing and tuning across a spectrum of hardware and software vendors. These hurdles have kept public cloud as the default option for many. However, new product offerings such as MinIO's S3 object storage software stack, combined with one-stop hardware solutions such as Seagate® Exos® application platforms (Exos AP), now offer a viable and complementary option to the cost of public cloud.

The combination of MinIO and Seagate Exos application platforms deliver an unprecedented simplification for those wishing to take ownership of their data. Together, these two solutions make it easier to facilitate a self-hosted deployment of vast, scalable, multi-petabyte-capable, on-premises S3 object storage.

MinIO is a highly performant S3-protocol-compliant object storage software implementation released under the AGPL 3.0 open source license.

Seagate Exos AP platforms bring a verified and supported modular hardware solution that's capable of hosting industry leading storage performance and density. By combining compute with storage, the Exos AP platform frees solution architects from the trial-and-error tasks of component integration and optimization.

This paper documents a demonstrated, cost-effective solution built around the Seagate Exos AP 4U100 platform and MinIO.

Architecture

Seagate Exos AP 4U100 Overview

Seagate Exos AP 4U100 provides a total of 96 3.5-inch (LFF) and four 2.5-inch (SFF) drive slots arranged in one of the industry's highest density 4U enclosures. It also hosts two compute modules—referred to as application processors (APs)—each sporting dual-socket Intel Xeon 4110 scalable processors @2.1Ghz and 8 cores per socket, along with up to 256GB of RAM and an onboard boot device that's separate from the storage devices. Additionally, 12Gb/s SAS (Serial Attached SCSI) is supported throughout the enclosure and four mini-SAS expansion ports are provided at the back of each compute module for expansion.

Seagate Exos AP platforms also offer remote management and monitoring via both in-band SES services and out-of-band CLI/WebUI services as provided by its integral BMC.

MinIO

MinIO is a high-performance object storage software solution that delivers S3 compatibility while requiring very little configuration. The server module itself is but a single self-contained binary. Generally, it only needs an initial configuration that describes the network and storage device topology for each participating node to start a fully operational cluster. MinIO is fully compatible with Kubernetes and Docker deployments and bare metal is fully supported on a swath of operating systems, including most current versions of Linux, FreeBSD, and even Windows.

MinIO directly supports **Prometheus** metrics and status reporting without requiring the installation of any additional software or modules.

Hardware Topology

The reference architecture deployed and tested by Seagate featured three Seagate Exos AP 4U100 enclosures configured with two compute nodes each. Note that the Exos AP 4U100 supports a "split mode, shared nothing" mode where the storage and servers in a single chassis are split into two independent systems with each server controller managing 48 HDDs and 2 SSDs. These three enclosures each managed 32 **Exos X16 16TB ST16000NM002G** 12Gb/s SAS disk drives, thus creating a six-node MinIO cluster with a total of 96 Intel Xeon processor cores driving 192 16TB drives. Together, this made for a raw cluster capacity of 3PB of storage within 12U of rack space. In addition, each of the six application processors were fitted with a **Mellanox CX516-A ConnectX-5 Dual-100GbE** network host adapter.

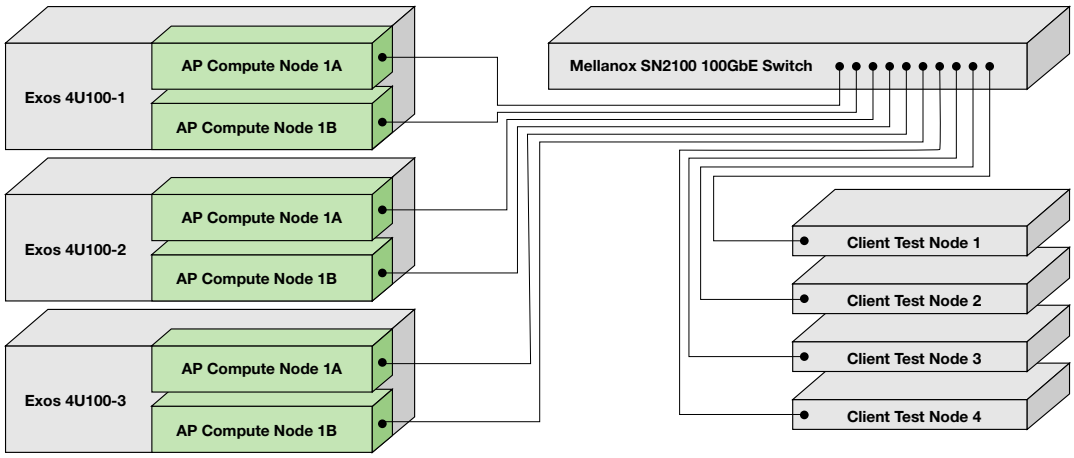
Bill of Materials List:

Storage Compute Nodes	Three Seagate Exos AP 4U100 application platforms, each with two dual-socket server controllers with Intel Xeon 4110 @2.1GHz 8 cores per socket
	RAM: 256GB
Test Compute Nodes	**Server: Intel 1U R1208WFTYS**
	CPU: Intel Xeon E5-2640
	RAM: 128GB DDR3 D3-68SA104SV-13
SAS HDD Drives	192 Seagate Exos X16 16TB ST16000NM002G (32 per node, 64 per system)
Network	2 per system Mellanox CX516-A ConnectX-5 Dual-100GbE NIC
	Mellanox SN2100 16-Port 100GbE Switch

Seagate Exos AP 4U100

Seagate Exos AP 4U100 offers a fully qualified, hardware-complete solution for a variety of software-defined storage applications—including MinIO. It provides a cost-effective answer to the alternative of qualifying an aggregated collection of vendor-unique options and permutations.

Network connection diagram for Seagate AP 4U100/MinIO test cluster



The network topology leveraged one of the two 100GbE ports offered by the Mellanox dual-port HBAs. While we tested configurations that placed server peer-to-peer communications on ports physically separate from the test client paths, we did not see an appreciable change in throughput. A bonded network configuration was not tested.

Test Software Environment

Host OS	SuSE SLES 15 SP2
MinIO Server	RELEASE.2021-01-08T21-18-21Z
MinIO Warp (test driver)	v0.3.29
Test scripts	https://github.com/suykerbuyk/minio.on.sgt.4u100

The Test Process

MinIO Warp was used as the S3 benchmarking tool. In previous testing, the Seagate Reference Architecture Lab has found it to be comparable to COSBench and far less likely to be subject to test configuration artifacts. In addition, Warp is also written in GO and can be distributed as a single monolithic binary to test clients, whereas the Java dependencies of COSBench can present significant issues in bare-metal deployments. Like with COSBench, Warp can be configured such that one control instance can orchestrate several client instances.

Our focus was on atomic testing of the three primary S3 operators: GET, PUT, and DEL. GET and PUT roughly translate to conventional read and write operations. Therefore, PUT operations are a measure of data ingress and GET operations are a measure of data egress.

For each configuration of GET, PUT, and DEL:

- Concurrency (threads) on each of four test clients from:
    - Each Client: 8, 16, 24, 48, 96, 128 threads
    - Total Threads: 32, 64, 96, 192, 256, 512
- Object Sizes:
    - .125 MB, .5MB, 1MB, 4MB, 16MB, and 64 MB
- Disk schedulers:
    - None/noop, deadline, kyber, BFQ
    - None/noop was best for small objects, deadline best overall
- With and without MinIO S3 caching on Nytro® SSDs:
    - Almost no discernible effect
    - With and without XFS metadata caching on Nytro SSDs:
    - 12 256MB partitions on each of four 3.84TB Nytro SSDs, such that each of 32 XFS formatted spinning disks were assigned an SSD partition for metadata caching
    - Approximately 5% performance boost for small objects
    - No measurable effect on 64MB objects
- MinIO cluster sizes:
    - 8, 16, and 32 disks per server controller
    - Fairly linear scaling with disk count
- Network configuration:
    - Single-port 100GbE for both client and server private
    - Dual-port 100GbE
        - One port for client communication
        - One port for server internal communication
        - No substantial difference
- In total:
    - 108 permutations of thread and concurrency per test for each of PUT, GET, DEL
    - 15 variations of platform and performance tuning for each of the 108 permutations above and for each of the three primary operators

Test Results

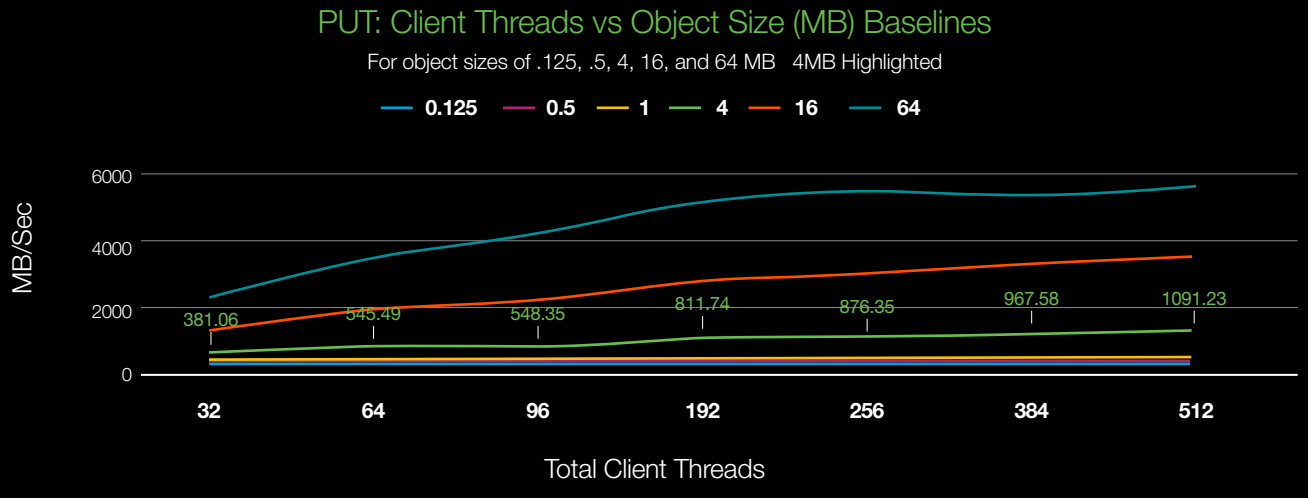
MinIO demonstrated strong read performance. Measured throughput on our 100GbE links hovers at around a peak of 22.5GB/s. MinIO was able to achieve 12GB/s which, given the way MinIO ingests on one server node and fans out the request to its peers, represents an ability to saturate the 100GbE single links.

Read (GET) MinIO Performance

Below we see the default out-of-the-box performance with 32 16TB Seagate Exos X16 SAS drives, with the default erasure coding (8+8), default single (100GbE) network routing, default deadline disk scheduler, and no added SSD caching to speed up operations.

Read (GET) MinIO Performance

Below we see the default out-of-the-box performance with 32 16TB Seagate Exos X16 SAS drives, with the default erasure coding (8+8), default single (100GbE) network routing, default deadline disk scheduler, and no added SSD caching to speed up operations.



Test Process Details

Test configuration and topology

A simple VM was used as the test orchestration point and the control point for provisioning tasks. Leveraging a VM allows us to archive the entire runtime environment should we ever need to go back to it.

We chose SuSE SLES 15 SP2 as the common operating system. For a fully operational SuSE yast automatic installation script that's suitable for PXE-based provisioning of Exos AP 4U100, see **SuSE autoyast network provisioning script.**

For creating instances of the MinIO server on each of the 4U100 server controllers, we used the **MinioDiskPrep.sh** script. This provides functions that prepare the basic server configuration for Minio and partition and format all HDDs and SSDs involved in each test scenario. It was designed to run all disk prep stages in parallel and can completely reconfigure and redeploy the MinIO backend in less than 25 seconds for six server nodes when issued in tandem.

We leveraged four generic Intel 1U servers to serve as MinIO Warp test clients while running the MinIO Warp S3 benchmark tool. Each instance of the warp client was started remotely via a simple forked command:

```
warp client $(ip a | grep 172.20.2 | awk -F ' ' '{print $2}' | sed 's/\/.*//g'):8000
```

The MinIO Warp Server script (**warp.test.sh**) had the responsibility of stepping through the three-way matrix of test data to be collected for each permutation of setup and disk configuration.

Specifically, it:

1. Copied over a copy of the **testmon.sh** script
2. Started testmon.sh on each MinIO server
3. For each object size of .125MB, .5MB, 1MB, 4MB, 16MB, and 64MB it:
    - Tested with 8, 16, 24, 48, 64, 96, and 128 client threads per each of the four test nodes (32, 64, 96, 192, 384, 512 total threads per test)
    - Tested each of four disk schedulers (none, deadline, kyber, bfq)
    - For each of PUT, GET, and DEL S3 operations
4. In total, about 108 permutations for each of the three primary S3 operations across 15 platform configurations

Conclusion

The combination of MinIO and the Seagate Exos AP 4U100 platform is an easy-to-deploy and simple-to-configure on-premises S3 object storage solution. The combined solution demonstrated up to 12GB/s read performance (GET) for 64MB objects and up to 5GB/s write performance (PUT) for 64MB objects using a six-node MinIO cluster created from three Seagate Exos AP 4U100 systems. Results for other object sizes and concurrency loads are also provided in the test results.

The MinIO server is a completely self-contained 50MB binary that does not require anything other than copy/paste/run for installation and, more importantly, has no external dependencies. It incorporates built-in Prometheus/Grafana dashboard reporting, as well as S3 bucket monitoring. All operational parameters can be passed at launch time with simple environmental variables or can be configured with the complimentary `mc` command.

Seagate Exos AP platforms offer cost-effective, fully qualified, and verified single-SKU solutions to complex storage environments, thus negating long and expensive cycles of vendor qualification, support, and integration challenges.

Resources:

Seagate + MinIO Solutions: https://www.seagate.com/solutions/partners/minio

Ready to Learn More? Visit us at **seagate.com**

seagate.com

© 2021 Seagate Technology LLC. All rights reserved. Seagate, Seagate Technology, and the Spiral logo are registered trademarks of Seagate Technology LLC in the United States and/or other countries. Exos and Nytro are either trademarks or registered trademarks of Seagate Technology LLC or one of its affiliated companies in the United States and/or other countries. All other trademarks or registered trademarks are the property of their respective owners. When referring to drive capacity, one gigabyte, or GB, equals one billion bytes and one terabyte, or TB, equals one trillion bytes. Your computer's operating system may use a different standard of measurement and report a lower capacity. In addition, some of the listed capacity is used for formatting and other functions, and thus will not be available for data storage. All coded instruction and program statements contained herein are, and remain copyrighted works and confidential proprietary information of Seagate Technology LLC or its affiliates. Any use, derivation, dissemination, reproduction, or any attempt to modify, reproduce, distribute, disclose copyrighted material of Seagate Technology LLC, for any reason, in any manner, medium, or form, in whole or in part, if not expressly authorized, is strictly prohibited. Seagate reserves the right to change, without notice, product offerings or specifications. TP734.2-2110US October 2021

SEAGATE