Western Digital.

TECHNOLOGY BRIEF

High performance object storage on OpenFlex™ E3000 powered by MinIO

Reference Architecture

WESTERN DIGITAL
D018-000007-AA01

Revision History

Revision	Date	Description	Reference
A00	February 2021	Initial release	

Typographical Conventions

This document uses the typographical conventions listed and shown in the table below.

Table 0-1.Typographical Conventions

Convention	Usage
Bold	Command and option names appear in bold type in definitions and examples. n Directories, files, partitions, and volumes also appear in bold. n Interface controls (check boxes, radio buttons, fields, folders, icons, list boxes, items inside list boxes, multicolumn lists, menu choices, menu names, and tabs) n Keywords and parameters in text
Italics	Variable information appears in italic type. This includes user-supplied information on command lines. n Citations (titles of books, diskettes, and CDs) n Emphasis of words n Words defined in text
`Monospace`	Screen output and code samples appear in monospace type. n Citations (titles of books, diskettes, and CDs) n Examples and code examples, for example, this is a line of code n File names, programming keywords, and other elements that are difficult to distinguish from surrounding text n Message text and prompts addressed to the user n Text that the user must enter n Values for arguments or command options

Western Digital Technologies, Inc. or its affiliates' (collectively "Western Digital") general policy does not recommend the use of its products in life support applications where in a failure or malfunction of the product may directly threaten life or injury. Per Western Digital Terms and Conditions of Sale, the user of Western Digital products in life support applications assumes all risk of such use and indemnifies Western Digital against all damages.

This document is for information use only and is subject to change without prior notice. Western Digital assumes no responsibility for any errors that may appear in this document, nor for incidental or consequential damages resulting from the furnishing, performance or use of this material.

Absent a written agreement signed by Western Digital or its authorized representative to the contrary, Western Digital explicitly disclaims any express and implied warranties and indemnities of any kind that may, or could, be associated with this document and related material, and any user of this document or related material agrees to such disclaimer as a precondition to receipt and usage hereof.

Each user of this document or any product referred to herein expressly waives all guaranties and warranties of any kind associated with this document any related materials or such product, whether expressed or implied, including without limitation, any implied warranty of merchantability or fitness for a particular purpose or non-infringement. Each user of this document or any product referred to herein also expressly agrees Western Digital shall not be liable for any incidental, punitive, indirect, special, or consequential damages, including without limitation physical injury or death, property damage, lost data, loss of profits or costs of procurement of substitute goods, technology, or services, arising out of or related to this document, any related materials or any product referred to herein, regardless of whether such damages are based on tort, warranty, contract, or any other legal theory, even if advised of the possibility of such damages.

This document and its contents, including diagrams, schematics, methodology, work product, and intellectual property rights described in, associated with, or implied by this document, are the sole and exclusive property of Western Digital. No intellectual property license, express or implied, is granted by Western Digital associated with the document recipient's receipt, access and/or use of this document or the products referred to herein; Western Digital retains all rights hereto.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.0 Executive Summary

The growth of data and how to manage and monetize it is the defining characteristic of the modern enterprise. Legacy storage systems such as File and Block struggle to manage the volume and velocity of unstructured data – despite their rich feature sets.

This has given rise to a new, increasingly dominant class of storage – object. Object storage has become the de-facto standard for this class of data, just as this class of data has become the dominant data type. Traditionally object storage was deployed for archival purposes, however, with its stateless HTTP RESTful APIs, scalability and massive performance improvements, it is now deployed against Machine Learning (ML), Internet of Things (IoT), Artificial Intelligence (AI) workloads. These workloads demand that compute and storage to be scaled independently. Composable Disaggregated Infrastructure (CDI) represents the modern architectural approach to datacenter infrastructure, disaggregating compute, storage, and network resources into shared pools that can be composed for on-demand allocation.

The OpenFlex F3200 is a fabric device that leverages the Open Composable Infrastructure (OCI) approach in the form of disaggregated data storage using NVMe™-over-Fabrics (NVMe-oF™).

NVMe™-over-Fabrics (NVMe-oF™) MinIO is a high-performance, Kubernetes-native, storage-as-a-service (STaaS) object storage platform, usually built on servers which has NVMe SSDs, designed for AI and ML workloads.

The purpose of this document is to showcase combined solution of MinIO object storage deployment on servers provisioned with OpenFlex F3200 fabric devices using

NVMe-oF protocol.

2.0 PROBLEM STATEMENT

The data storage challenges of today are vastly different today. Machines produce more data than traditional sources such as video or web traffic. That data is produced in the datacenter but increasingly at the edge as well.Analysts from IDC are expecting data will grow to 175 zettabytes of data by 2025[1]. Almost 80% of that data will be unstructured.

Traditional SAN, NAS approaches, while being feature rich, are overly complex, depend on an outdated API (POSIX) and struggle to deliver against the massive scale required for modern workloads. While there are a number of different object storage options from public Cloud Service Providers (CSP) many enterprises have requirements around performance, security and control that preclude those options. This is where on-premises object storage comes in. Given the standardization around Amazon's S3 API, on-premises object storage is drop-in replacement for the public cloud and gives service providers an extremely cost-effective, and highly scalable environment. Object stores can scale up to hundreds of petabytes in a single namespace without suffering any sort of performance degradation.

As more and more business opportunities are depending on big data, ML, AI, Internet of Things (IoT) workloads, they require data infrastructure designed to scale storage and compute independently to ensure both are provisioned efficiently and effectively. Composable Disaggregated Infrastructure (CDI) ensures compute, storage, and network resources are placed into shared pools that can be composed for on-demand allocation. This enables computing to become stateless, elastic, and scalable independent of storage.

In this document we demonstrate a solution, to growing demand of unstructured data, by deploying object storage platform on a composable infrastructure.

[1] https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/?sh=68560db05459

3.0 TECHNOLOGY OVERVIEW

3.1 OpenFlex™ E3000 enclosure with F3200 fabric device overview

The OpenFlex E3000 is a 3U rack mounted data storage enclosure built on the OpenFlex platform. OpenFlex is Western Digital's architecture that supports Open Composable Infrastructure (OCI) through storage disaggregation. The OpenFlex F3200 is a fabric device that leverage this OCI approach in the form of disaggregated data storage using NVMe-over-Fabrics (NVMe-oF).

NVMe-oF is a networked storage protocol that allows storage to be disaggregated from compute to make that storage widely available to multiple applications and servers.

By enabling applications to share a common pool of storage capacity, data can be easily shared between applications, or needed capacity can be allocated to an application regardless of location. Exploiting NVMe device-level performance, NVMe™-oF™ promises to deliver the lowest end-to-end latency from application to shared storage. NVMe-oF enables composable infrastructures to deliver the data locality benefits of NVMe DAS (low latency, high performance) while providing the agility and flexibility of sharing storage and compute.

Figure 3-1. OpenFlex™ E3000 enclosure with F3200 Series Fabric Devices

The maximum data storage capacity of E3000 is 614TB[1] when leveraging a full set of 10 F3200 fabric devices. F3200 is capable of scaling up to 2 million IOPs and cumulatively we can scale for each E3000 up to 20 million IOPs in a 3U solution.

Composable Infrastructure seeks to disaggregate compute, storage, and networking fabric resources into shared resource pools that can be available for on-demand allocation (i.e., "composable"). Composability occurs at the software level, disaggregation occurs at the hardware level using NVMe™ -over-Fabric (NVMe-oF) will vastly improve compute and storage utilization, performance, and agility in the datacenter.

[1] Raw capacity. One gigabyte (GB is equal to one billion bytes and one terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.

Western Digital's vision for Open Composable Infrastructures (OCI) is based on four key pillars:

Table 3-1.Key Pillars for Open Composable Infrastructure

Open	Open in both API and form factor
	Designed for robust interoperability of multi-vendor solutions
Scalable	Ability to compose solutions at the width of the network
	Enable self-organizing systems of composable elements that communicate horizontally
Disaggregated	Pools of resources available for any use case that is defined at run time
	Independent scaling of compute and storage elements to maximize efficiency and agility
Extensible	Inclusive of both disk and flash
	Entire ecosystem of composable elements managed and orchestrated using a common API framework
	Prepared for yet-to-come composable elements - e.g., memory, accelerators

Open Composable API Western Digital's new Open Composable API is designed for datacenter composability. It builds upon existing industry standards utilizing the best features of those standards as well as practices from proprietary management protocols.

OpenFlex is Western Digital's architecture that supports Open Composable Infrastructure (OCI) through storage disaggregation – both disk and flash natively attached to a scalable fabric. OpenFlex does not rule out multiple fabrics, but whenever possible, ethernet will be used as a unifying connect for both flash and disk because of its broad applicability and availability.

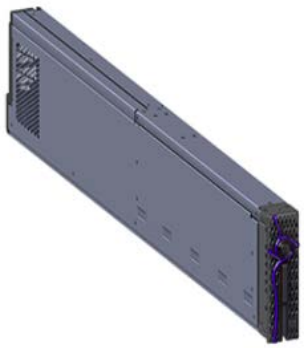


Figure 3-2. OpenFlex F3200 Specifications

Table 3-2. OpenFlex F3200 Specifications

Specification	Value
Max Raw Data Storage Capacity per Device	61.4TB
Data Ingest Capability	2x 50G Ethernet
Data Transfer Rates	12GBps*
Number per Enclosure	Up to 10
How Swappable	Yes

3.2 MinIO Overview

MinIO is a pioneer in the development of high-performance, cloud-native object storage, refining and perfecting many of the features, protocols and APIs that have come to define best in class. Given MinIO's performance characteristics it is often used for machine learning, IoT, AI and big data workloads.

- MinIO's software defined object storage consists of a server, optional client and optional software development kits (SDKs):
- MinIO server is a distributed object storage server released under Apache License v2.0 that includes a array of enterprise-grade features.
- MinIO client ("mc") is a modern and cloud-native alternative to UNIX commands that supports web-scale object storage deployments.
- MinIO SDKs include simple APIs for accessing any Amazon S3™-compatible object storage. MinIO has SDKs for popular development languages such as Golang, JavaScript, .Net, Python and Java.

The features of MinIO's object server are notable for their breadth, depth and focus on enterprise grade challenges. Many of those features are considered best-in-class in the object storage industry:

3.2.1 Erasure Coding

MinIO protects data with per-object, inline erasure coding which is written in assembly code to deliver the highest performance possible. MinIO uses Reed-Solomon code to stripe objects into k data and m parity blocks. It allows read or new write of objects even if there is node failure or multiple device failure corresponds to n-1 drives. It can be configured to any desired redundancy level.

3.2.2 BitRot Protection

MinIO's optimized implementation of the HighwayHash algorithm, ensures that it will never read corrupted data - it captures and heals corrupted objects on the fly. Thus it avoids the silent data corruption or bitrot which is caused by aging drives, current spikes, bugs in disk firmware, phantom writes, misdirected reads/writes, driver errors, accidental overwrites.

3.2.3 Identity and Access Management

MinIO supports the most advanced standards in identity management, integrating with the OpenID connect compatible providers as well as key external IDP vendors.

That means that access is centralized and passwords are temporary and rotated, not stored in config files and databases. Also access policies are fine grained and highly configurable which makes supporting multi-tenant and multi-instance deployments become simple.

3.2.4 Encryption and WORM

MinIO supports multiple, sophisticated server-side encryption schemes to protect data – whether it is in flight or at rest. MinIO uses key-management-systems (KMS) or cryptographic key management system (CKMS) to support SSE-S3. If a client requests SSE-S3, or auto-encryption is enabled, the MinIO server encrypts each object with a unique object key which is protected by a master key managed by the KMS. Given the exceptionally low overhead, auto-encryption can be turned on for every application and instance.

Data is written once and enabled WORM (Write Once Read Many). MinIO then disables all APIs that can potentially mutate the object data and metadata to ensure data become tamper-proof.

3.2.5 Scalability and Distribution

MinIO is a multi-tenant and multi-user system and is designed to scale seamlessly from TBs to any size. The tenants are fully isolated from each other with their own instances of MinIO clusters. Each tenant in turn may have multiple users with varying levels of access privileges. Each tenant cluster operates independently of each other. Each cluster is a collection of fully symmetric and distributed servers sets that participate equally in serving the objects. Standard HTTP load balancers or round-robin DNS may be employed. A single cluster may span an entire data center and grow to 100s of petabytes. Within a cluster, racks of homogenous servers are grouped into zones. Zones are the basic unit of expansion and they bring the concept of rack-awareness and failure-domains. A zone can be as small as four servers and as large as multiple racks. A cluster is scaled by adding one or more zones at a time. There is no rebalancing penalty for scaling. Zones also allow heterogeneous expansion of the cluster.

3.2.6 Multi-cloud Gateway

MinIO can be deployed in gateway mode to leverage public cloud resources. Leveraging the same binary, MinIO enables companies to run their applications on premises or in the public cloud with no modification. MinIO make sure that data present in all the cloud looks same with Amazon S3 API.

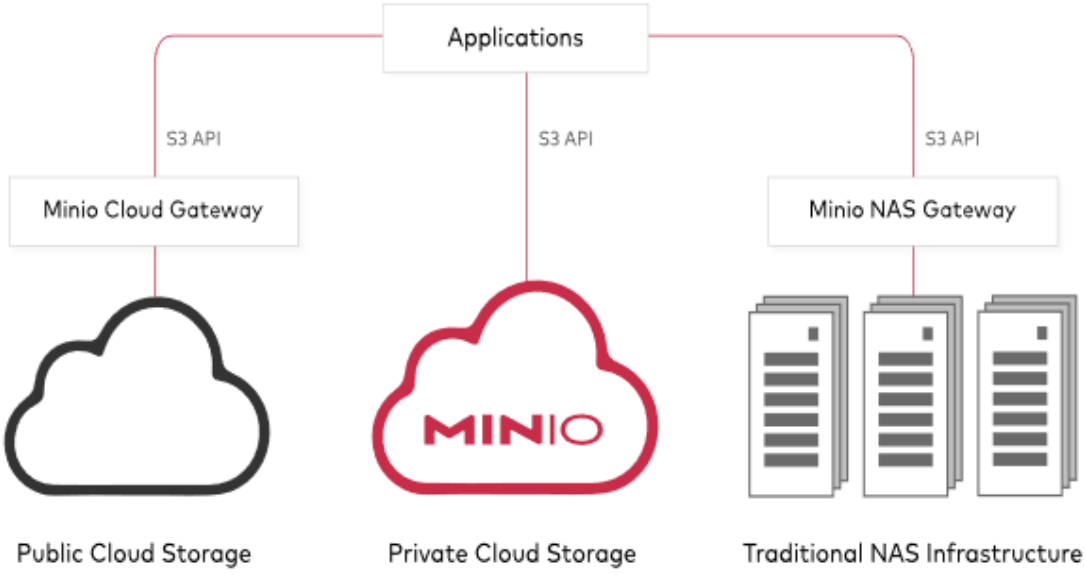


Figure 3-3. Multi-cloud gateway

3.2.7 Active-Active Continuous Replication

To support at-scale replication for critical workloads, MinIO has developed technology that allows for two geographically distinct data centers to withstand a data center failure without any downtime for end clients. MinIO follows strict consistency within the data center and eventual-consistency across the data centers to protect the data. Replication performance is dependent on the bandwidth of the WAN connection and the rate of mutation. As long as there is sufficient bandwidth, the changes are propagated immediately after the commit. Versioning capability enables MinIO to behave like an immutable data store to easily merge changes across the active-active configuration.

3.2.8 Metadata Architecture

MinIO has no separate metadata store. All operations are performed atomically at object level granularity. This approach isolates any failures to be contained within an object and prevents spill over to larger system failures.

3.2.9 Cloud Native

The multi-instance, multi-tenant design of MinIO enables Kubernetes-like orchestration platforms to seamlessly manage storage resources just like compute resources. Each instance of MinIO is provisioned on demand through self-service registration.

3.2.10 Lamda Function Support

MinIO supports Amazon compatible lambda event notifications which enables applications to be notified of individual object actions such as access, creation, and

deletion. The events can be delivered using industry standard messaging platforms like Apache Kafka®, NATS, AMQP, MQTT, Webhooks, or a database such as Elasticsearch®, Redis™, Postgres, and MySQL™.

3.3 MinIo With OpenFlex Solution Design

Western Digital, a pioneer in reliable, high-density industry-standard hardware for software-defined storage projects, is partnering with MinIO, a global leader in STaaS (storage as a service) object storage platform, to provide verified, scalable object storage solution for on-premises clouds. The following sections of this paper provide an overview of object storage solution - built on Western Digital, powered by MinIO.

The solution combines:

- Western Digital OpenFlex: OpenFlex F3200 which leverages open composable infrastructure approach in the form of disaggregated data storage using NVMe-oF (NVMe over Fabric) perfect fit for scaling big data, AI environments.
- MinIO: MinIO is a high performance, distributed object storage system designed for AI and ML workloads.

By combining MinIO server with Western Digital OpenFlex F3200, organizations can benefit from STaaS Platform:

- Robust performance
- Scalable OpenFlex infrastructure
- Data durability
- Data integrity
- Easy to install and upgrade cluster
- Easy to integrate with public cloud
- Easy data management

Please do the below steps to setup MinIO cluster with OpenFlex storage

- Install Ubuntu 20.04 LTS on 4 industry standard x86 servers then setup NTP to have time sync.
- Setup 2 Mellanox® CX5 cards and install latest Mellanox OFED drivers on servers.
- Setup lossless setting between OpenFlex blades, MinIO servers and Mellanox switch.
- Create 8 volumes each from 4 OpenFlex F3200 fabric devices using Western Digital unified web portal.
- Install nvme cli package and connect 8 volumes from one OpenFlex blade to one MinIO server. Totally 32 volumes shared across 4 servers.
- Format all the volumes with XFS filesystem and mount volumes in each server as /mnt/export1 to 8.
- Setup access key, secret key and erasure code setting in the terminal on all the MinIO servers.

```
export MINIO_ACCESS_KEY=<ACCESS_KEY>
export MINIO_SECRET_KEY= <SECRET_KEY>
export MINIO_STORAGE_CLASS_STANDARD=EC:4
```

```
export MINIO_STORAGE_CLASS_RRS=EC:2
```

- Get the latest minio binary and run the following cmd on all the MinIO servers to start the cluster *minio server http://host{1...4}/mnt/export{1...8}*

Please find the diagram which depicts network connections between MinIO nodes and OpenFlex E3000 enclosure with F3200 fabric devices:

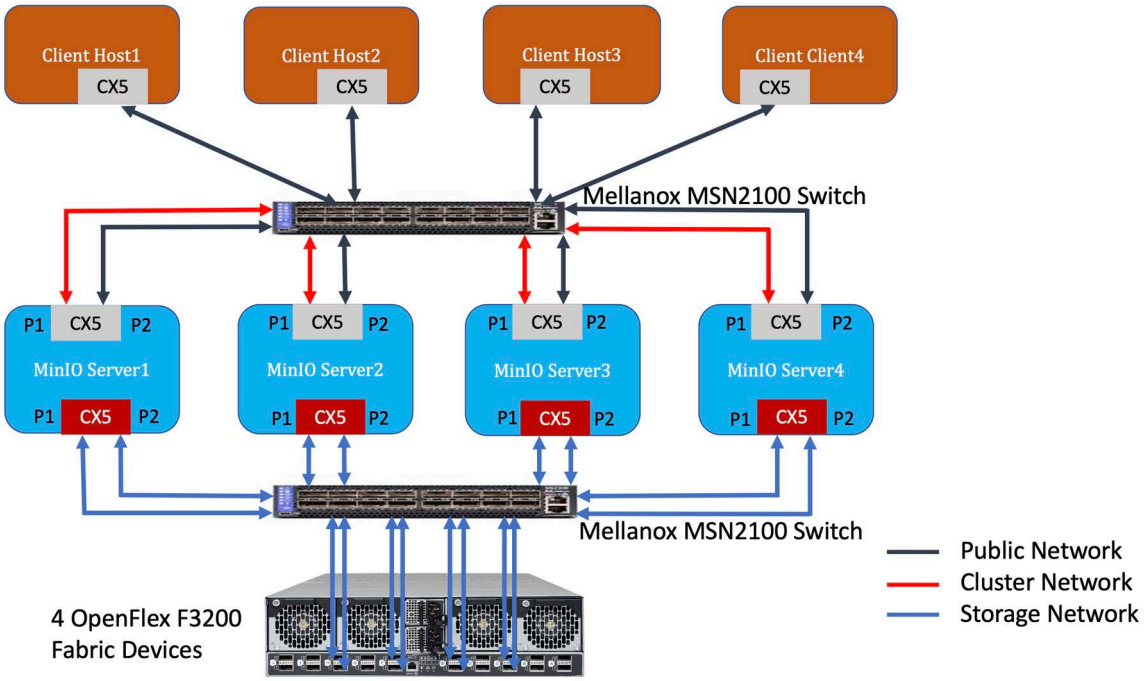


Figure 3-4. Network connections between MinIO nodes and Openflex E3000

Please find the hardware and software configurations used for deploying MinIO cluster in the table below:

Table 3-3. Configuration Details

Configuration Details	
Storage Product	OpenFlex E3000 with 4 x F3200 Fabric Device
Network Storage Protocol	NVMe over Fabric using RoCE v2
Raw Capacity per Fabric Device	61.4 TB
Total Raw Capacity with 4 x F3200 Devices	245.6 TB
Host Processor	Intel® Xeon® Gold 6230 CPU @ 2.10GHz
Host OS	Ubuntu 20.04 LTS (Focal Fossa)
Kernel	5.4.0-48-generic x86_64
CPU	80 with HT enabled

Table 3-3. Configuration Details

Configuration Details	
DIMM	128 GB per MinIO Server
NIC	Mellanox CX5
Total Volumes	32
File System	XFS
MinIO SW Version	minio version RELEASE 2020-09-05T07-14-49Z
Erasure Code	12+4

3.4 S3 Performance Results on MinIO Cluster

MinIO cluster created with 4 nodes having Intel Xeon Gold processors provisioned with storage from 4 OpenFlex F3200 Fabric Devices. WARP benchmark tool, developed and maintained by MinIO, used to run benchmark tests against MinIO cluster.

WARP benchmark ran with 4 WARP clients and 1WARP client was acting as WARP server. Each client is installed with Ubuntu 16.04, 100Gbps CX5 card and provisioned with 128GiB of RAM. Benchmark test ran with objects of various sizes from 4KB to 512MB and number of concurrent threads varied from 100 to 800.

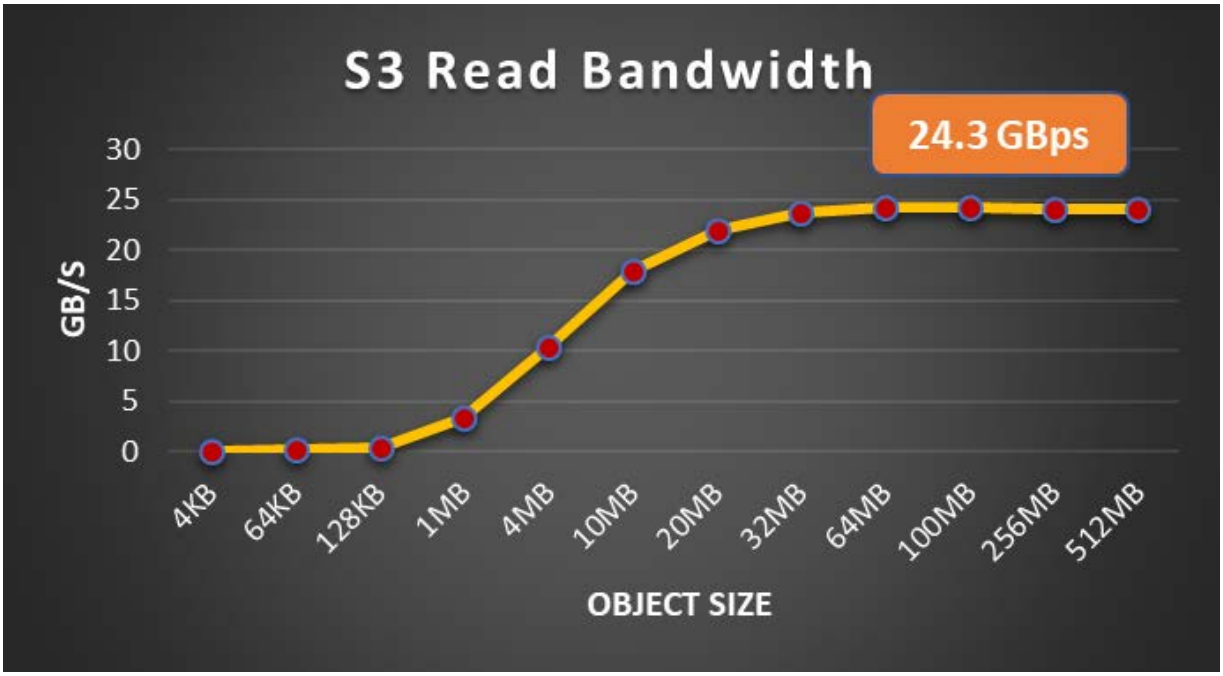
As a result of benchmarking tests with erasure coding 12+4 on MinIO servers, got read bandwidth as 24.3 GBps and write bandwidth as 8.92 GBps.

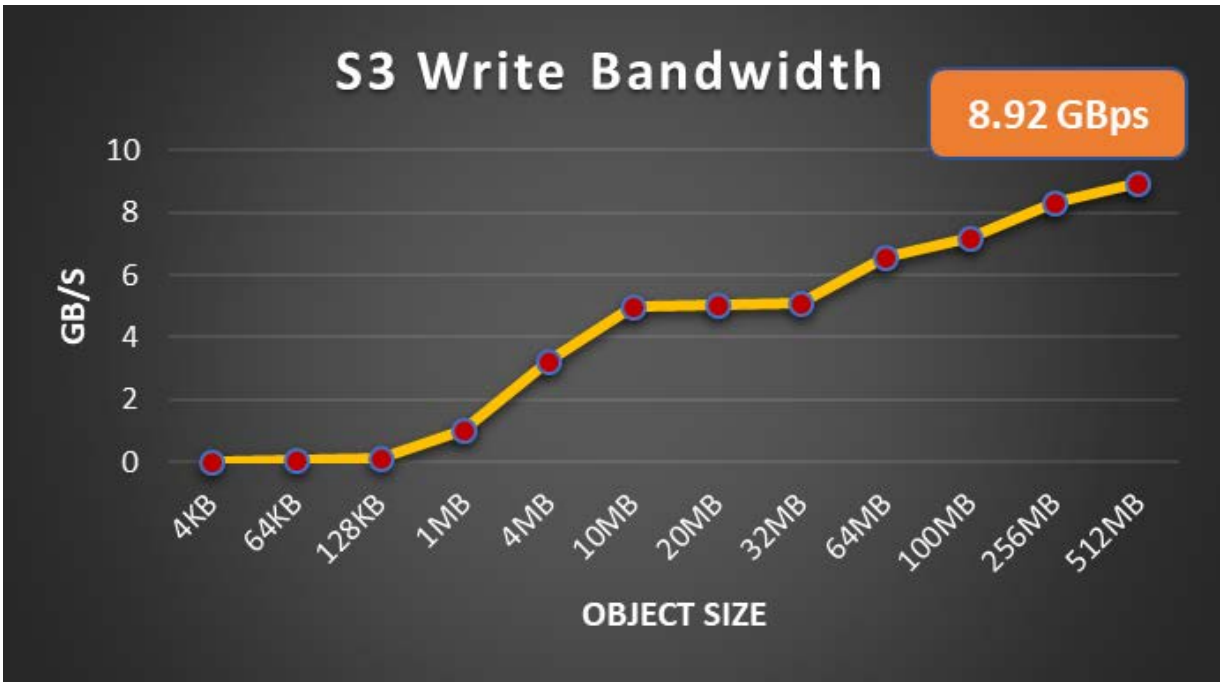
Table 3-4. Performance summary

Avg. Read/GET Bandwidth	24.3 GBps
Avg. Write/PUT Bandwidth	8.92 GBps

Performance numbers will increase if number of F3200 devices are increased. OpenFlex E3000 enclosure can have maximum of 10 OpenFlex F3200 fabric devices.

Below Chart depicts S3 read, write bandwidth for various object sizes on MinIO cluster with 4nodes.


S3 Read Bandwidth
24.3 GBps
GB/S
30
25
20
15
10
5
0
4KB
64KB
128KB
1MB
4MB
10MB
20MB
32MB
64MB
100MB
256MB
512MB
OBJECT SIZE


S3 Write Bandwidth
8.92 GBps
GB/S
10
8
6
4
2
0
4KB
64KB
128KB
1MB
4MB
10MB
20MB
32MB
64MB
100MB
256MB
512MB
OBJECT SIZE

3.5 Use Cases and Application Workloads

MinIO high performance object storage storage-as-a-service (STaaS) platform on NVMe over Fabrics based OpenFlex™ F3200 fabric fuels modern applications which needs low latency high performance storage.

Few of them are,

- Big data applications
- AI and ML workloads
- On-premise private cloud deployments
- Amazon S3 storage compatible applications
- Hybrid cloud infrastructure
- Healthcare data archival
- Data analytics

4.0 CONCLUSION

As massive amounts of unstructured data challenge the enterprise, the storage infrastructure has to be scalable without degrading its performance. Legacy SAN, NAS are not able to cope up with incoming rate of unstructured data and are not suited to modern applications. This is given rise to the adoption of modern object storage as the primary storage class for modern enterprise application workloads.

MinIO is one of the fastest growing object storage system with enterprise grade feature set. Combining it with OpenFlex fabric storage, which provides disaggregated data storage using NVMeOF protocol, customer can build a scalable high performance object storage for datacenters.

This object storage storage-as-a-service (STaaS) platform provides,

- High performance object storage
- Low TCO by reducing storage over-provisioning
- Data protection and integrity
- Dynamically scalable infrastructure
- Resilient and high-availability object storage

5.0 Resources and Additional Links

Distributed MinIO Quickstart Guide:
https://docs.min.io/docs/distributed-minio-quickstart-guide.html

OpenFlex Composable Infrastructure from Western Digital

https://www.westerndigital.com/products/storage-platforms/openflex-composable-infrastructure

6.0 Contact Information

Western Digital
5601 Great Oaks Parkway, San Jose, CA 95119
Phone: +1-408-801-1000
Fax: +1-408-801-8657
Email: oemproducts@wdc.com

Western Digital®

Western Digital
5601 Great Oaks Parkway
San Jose, CA 95119
U.S.A.

For service and literature:
support.wdc.com
www.westerndigital.com
800.ASK.4WDC North America
+31.88.0062100 EMEA & EU

D018-000007-AA01
February 2021