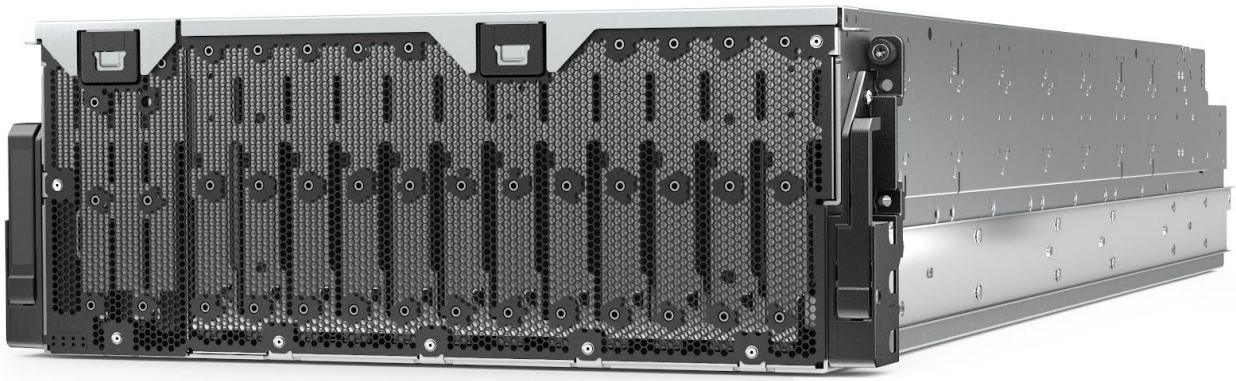
“Rack & Stack” on Prem S3 with MinIO and the Seagate EXOS 4u100 AP

John Suykerbuyk
Lyve Labs Reference Architecture Lead
Seagate PLC

Introduction

MinIO is a High Performance Object Storage released under Apache License v2.0. It is API compatible with Amazon S3 cloud storage service. MinIO is an exceptionally easy to deploy and manage, and performant Amazon S3 API compatible software defined storage software stack.

The Seagate® Exos® AP 4U100 is the datasphere’s highest density combination of compute and storage in a single system. While delivering state of the art density, capacity, and cost effective HDD performance - it also simplifies data center deployment by reducing compute plus storage to a single, widely available commodity component. Solution architects no longer need to mix & match and qualify individual components from multiple vendors, Seagate’s fully qualified Exos 4u100 AP provides a one stop, proven and demonstrable solution.



The hardware test environment

Material List:

Storage Compute Nodes	Exos 4u100 AP with Rockingham Controllers
	Dual socket Xeon 4110 @ 2.1 GHz 16 cores per
	RAM: 256GB
Test Compute Nodes	Server: Intel 1U R1208WFTYS
	CPU: Xeon(R) CPU E5-2640
	RAM: 128GB DDR3 D3-68SA104SV-13
SAS HDD Drives	Exos X16 16TB ST16000NM002G
Network	Mellanox CX516-A ConnectX-5 Dual 100 GbE
	Mellanox N2100 16 Port 100GbE Switch

The software test 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.s3.on.st.4u100

Test permutations exercised

For each configuration for each of GET, PUT, an DEL:

- Concurrency (threads) on each of 4 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 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 disk per Rockingham.
    - 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.

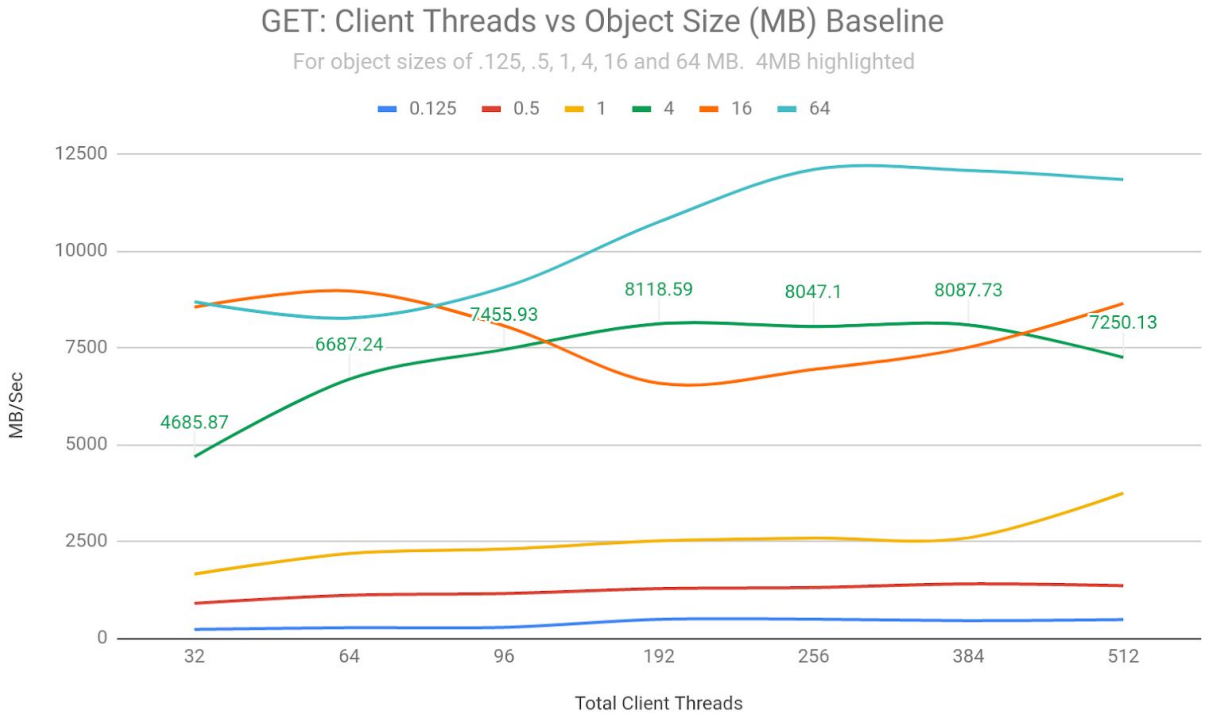
Test Results, best case scenarios.

MinIOs read performance is simply outstanding. Measured throughput on our 100GbE links hovers at around a peak of 22.5 GB/Second. MinIO was able to achieve 12GB/S, which given the way MinIO ingest on one server node and fans out the request to it's peers, represents an ability to saturate our links. What was surprising is when we gave MinIO it's own private network connections for intra-server communication, we did not see an increase of overall client throughput.

Read (GET) MinIO Performance, default configuration

Below we see the "default" out of the box performance with 32 16TB Evans drives, with the default EC:8 erasure coding, default single (100GbE) network routing, default "deadline" disk scheduler, and no added SSD caching to speed up operations.

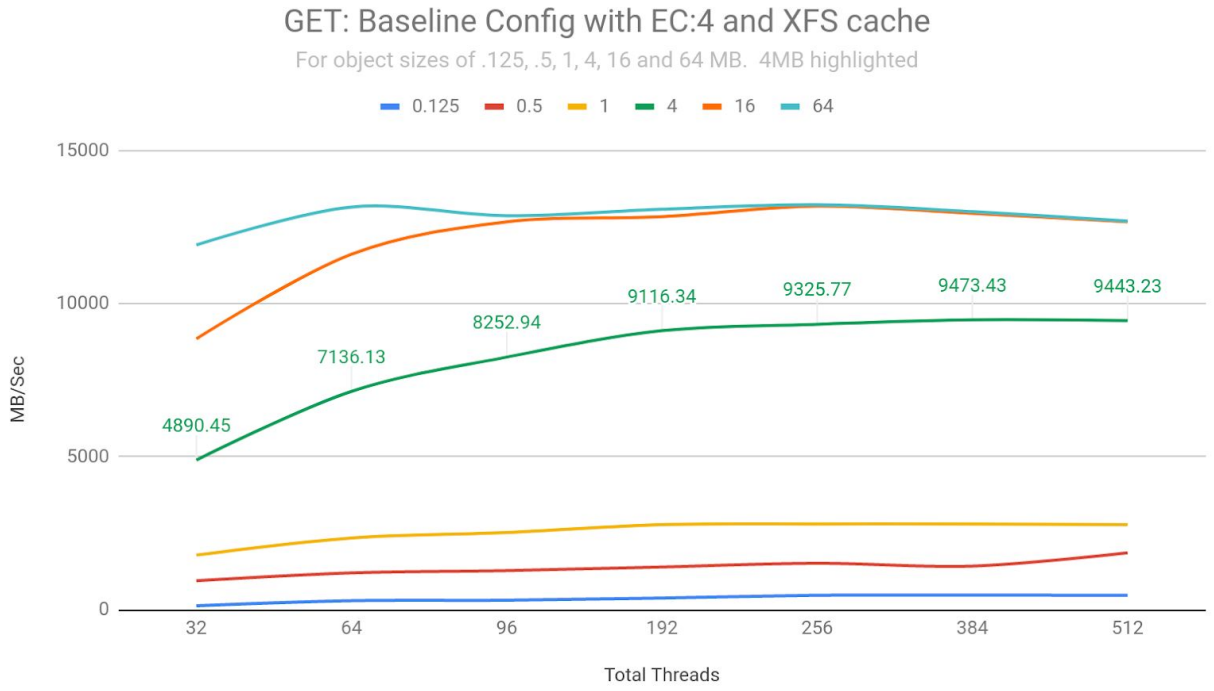
The data point that most aligns with the SmartShelf opportunity is the 4MB, 192 thread data point at 8GB/Sec, far and away exceeding requirements and even the best "read" performance the RA Lab has ever gotten from CEPH.



Read (GET) Minio performance, tuned configuration

Below we see MinIO being run with XFS metadata caching on Nytro SSDs, dual network paths, and the “deadline” disk scheduler.

Compared to the default performance we were able to get the critical 4MB object read/GET performance up from 8GB/Sec to over 9GB/Sec, a delta of about 12% in total.



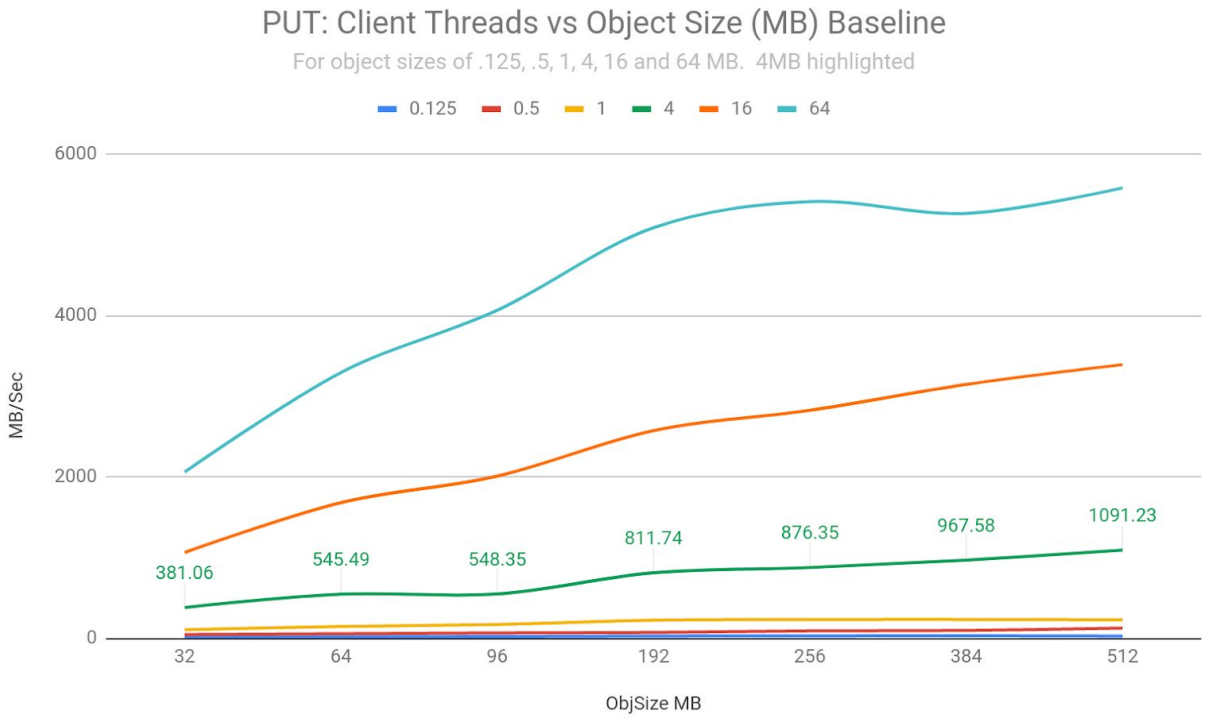
Read (GET) performance summary

Default MinIO 4MB/Obj 192 threads	8.1 GB/Sec
Optimized MinIO 4MB/Obj 192 threads	9.1 GB/Sec

Write (PUT) MinIO Performance, default configuration

Below we see the “default” out of the box performance with 32 16TB Evans drives, with the default EC:8 erasure coding, default single (100GbE) network routing, default “deadline” disk scheduler, and no added SSD caching to speed up operations.

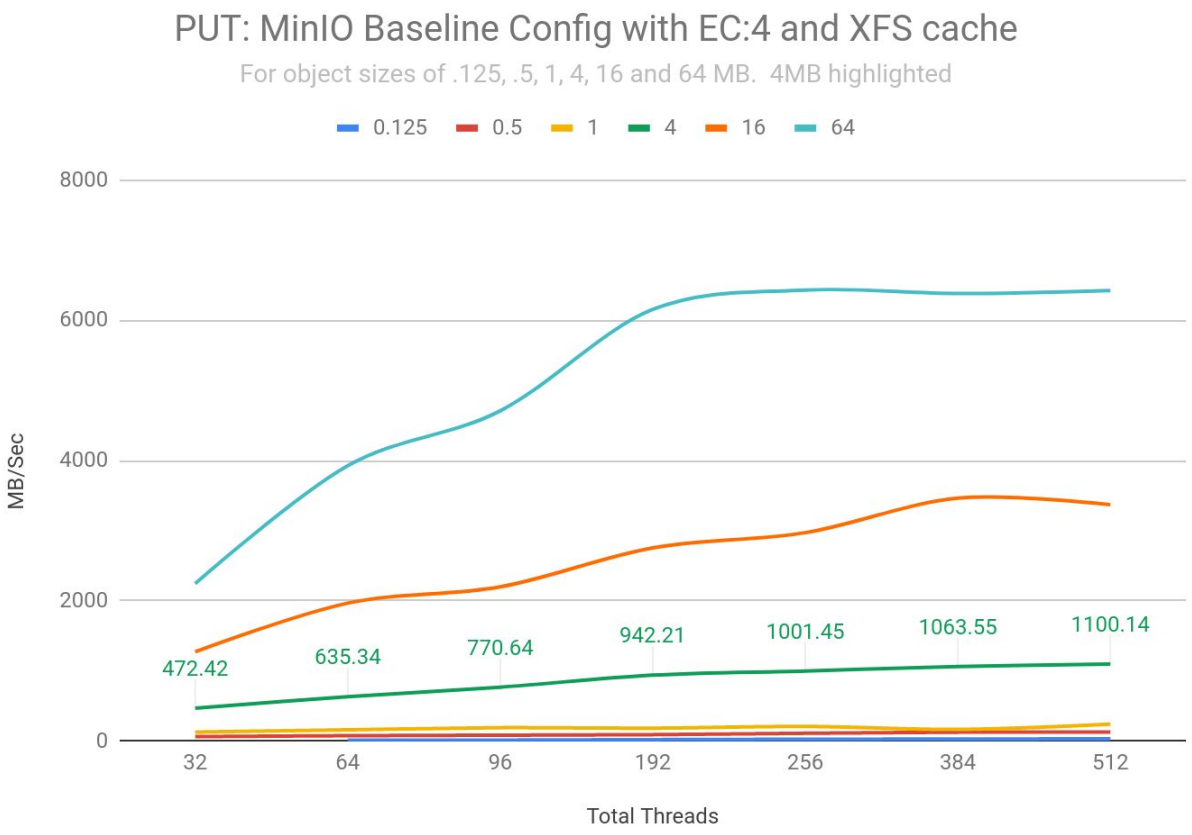
The data point that most aligns with the SmartShelf opportunity is the 4MB, 192 thread data point at 811MB/Sec.



Write (PUT) Minio performance, tuned configuration

Below we see MinIO being run with XFS metadata caching on Nytro SSDs, dual network paths, and the “deadline” disk scheduler.

Compared to the default performance we were able to get the critical 4MB object read/GET performance up from 811MB/Sec to over 940/Sec, a delta of about 15% in total.



Write (PUT) performance summary

Default MinIO 4MB/Obj 192 threads	811 MB/Sec
Optimized MinIO 4MB/Obj 192 threads	942 MB/Sec