MINIO

Supermicro Cloud DC Benchmark

OCTOBER 2021

MinIO S3 Throughput Benchmark on SuperMicro Cloud DC with NVMe Drives

The growth of data and how to manage and monetize it is the defining characteristic of the modern enterprise.

The defining architecture for these enterprises follows the lead of the hyper-scalers where storage and compute are disaggregated, best-of-breed and designed for commodity hardware. This enables computing to become stateless, elastic, and independently scalable from storage.

The standard for this architecture is modern object storage. This is a function of modern object storage's RESTful APIs (S3), scalability, resiliency and throughput performance.

This document describes the performance of MinIO object storage when run on Supermicro Cloud DC servers with NVMe drives.

MinIO is a cloud-native object storage suite designed for high-performance workloads such as AI/ML, advanced analytics and databases. MinIO is software defined and open-sourced under the AGPL v3 license. The object storage suite consists of a server, and optional components such as a client, a management console, a Kubernetes Operator and Operator Console and a software development kit (SDK).

Supermicro Cloud DC servers are designed for the needs of the cloud data center. These all-in-one 2U rackmount systems are compact and efficient. They employ Supermicro's tool-less design for rapid deployment and easy maintenance. They are an excellent platform for hyperconverged storage with up to 12 3.5" hot-swap NVMe/SATA/SAS drive bays, up to 16 DIMM Slots, up to 6TB DDR4-3200 memory, support for Intel® Optane™ persistent memory, Dual 3rd Gen Intel® Xeon® Scalable processors up to 270W TDP or Single 3rd Gen AMD EPYC™ processor up to 280W TDP, and Dual AIOM (Superset OCP 3.0 NIC) for up to 200 Gbps networking.

Our results running on a 4 node MinIO cluster with 40 NVMe drives and a 100 Gbps Network can be summarized as follows:

Our results running on 24 node MinIO cluster can be summarized as follows:

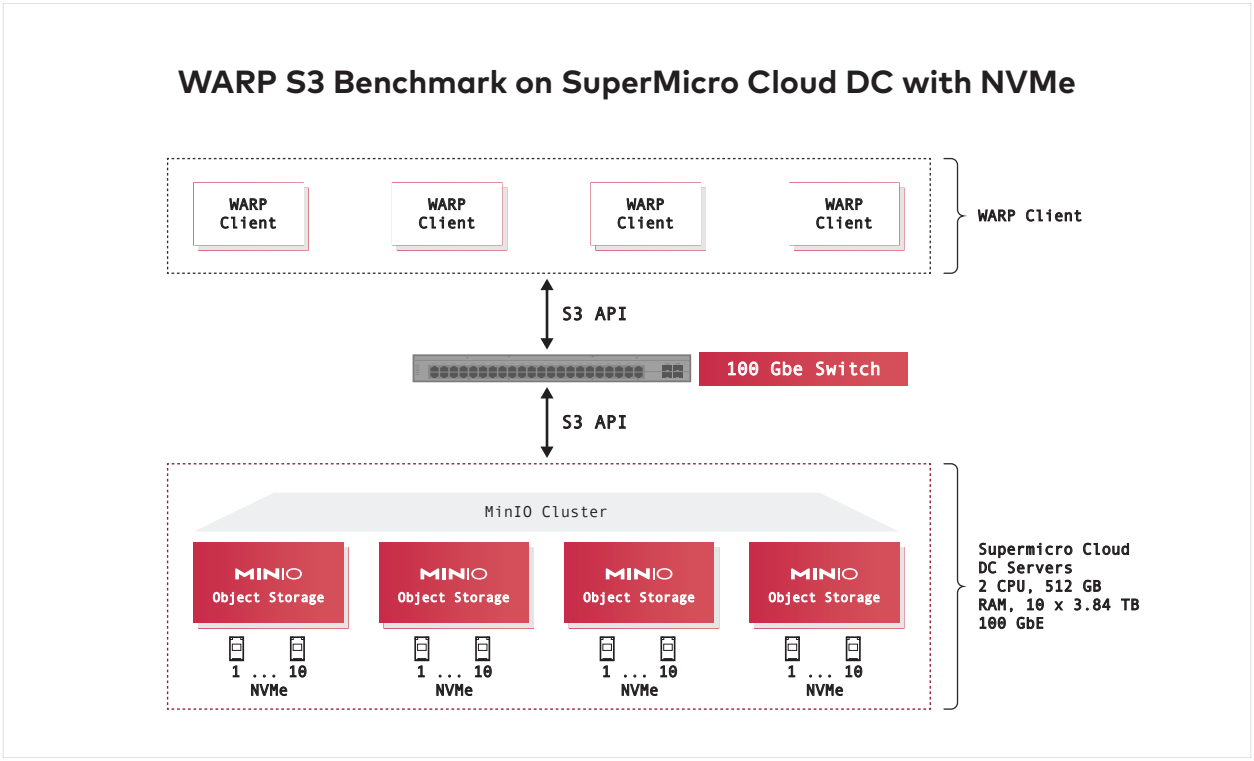
Setup	Avg Read Throughput (GET)	Avg Write Throughput (PUT)
Distributed	42.57 GB/s	24.69 GB/s
Distributed with Encryption	42.54 GB/s	25.96 GB/s
Distributed with TLS	42.35 GB/s	24.42 GB/s
Distributed with TLS and encryption	42.41 GB/s	23.88 GB/s

1. Benchmark Environment

1.1 Hardware

MinIO utilized SuperMicro Cloud DC with local NVME drives and 100 GbE networking.

Instance	# Nodes	Server type	CPU	MEM	Storage	Network
Server	4	SuperMicro Cloud DC	2	512 GB	10 x 3.84TB NVMe	100 Gbps

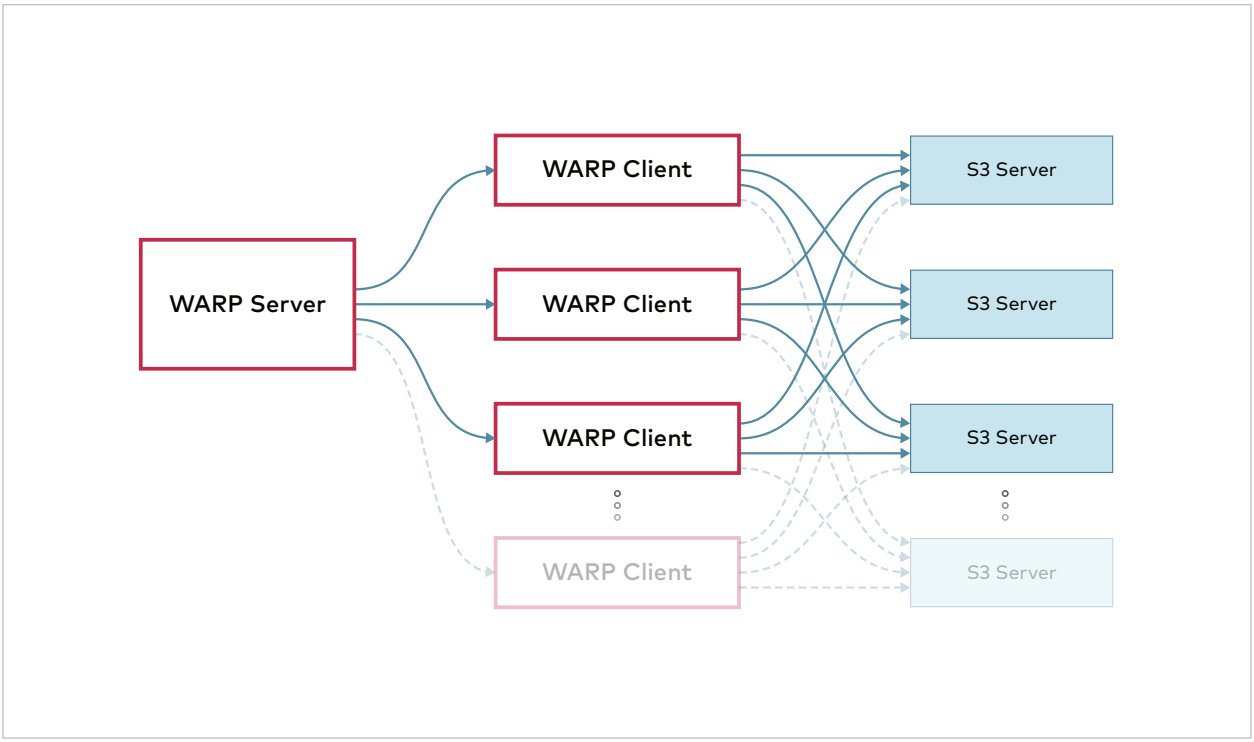


1.2 Software

Property	Value
Server OS	Ubuntu-20.04.2
MinIO Version	2021-08-05T22:01:19Z
Benchmark Tool	WARP v0.5.0 https://github.com/minio/warp

1.3 WARP S3 Benchmark

Supermicro and MinIO ran the WARP S3 benchmark (https://github.com/minio/warp) for our performance tests. This tool conducts benchmark tests from one or more clients to one or more hosts. WARP is an open source S3 performance benchmark tool developed and maintained by MinIO.



It is possible to coordinate several WARP instances automatically. This can be useful for testing performance of a cluster from several clients at once.

WARP can be configured using command line parameters or environment variables. The S3 server to use can be specified on the command line using **`--host`**, **`--access-key`**, **`--secret-key`** and optionally **`--tls`** and **`--region`** to specify TLS and a custom region. The same parameters can be set using the **`WARP_HOST`**, **`WARP_ACCESS_KEY`**, **`WARP_SECRET_KEY`**, **`WARP_REGION`** and **`WARP_TLS`** environment variables.

The credentials must be able to create, delete and list buckets and upload files and perform the operation requested.

By default, operations are performed on a bucket called warp-benchmark-bucket. This can be changed using the **`--bucket`** parameter. The bucket will be completely cleaned before and after each run, so it should not contain any data.

You can enable server-side-encryption of objects using **`--encrypt`**. A random key will be generated and used for objects.

WARP is run as follows:

warp command [options]

For example, running a mixed type benchmark against 8 servers named s3-server-1 to s3-server-8 on port 9000 with the provided keys:

```
warp mixed --host=s3-server{1...8}:9000 --access-key=minio
--secret-key=minio123 --autoterm
```

This will run the benchmark for up to 5 minutes and print the results.

To test N number of servers, you will need N number of client machines to generate sufficient load.

1.4 MinIO Configuration

The MinIO binary was downloaded onto each server node, and configured as follows:

```
# Remote volumes to be used for MinIO server.
MINIO_VOLUMES=http://data{1...4}/mnt/drive{1...10}/minio4
# Use if you want to run MinIO on a custom port.
MINIO_OPTS="--console-address :9199"
# Root user for the server.
MINIO_ROOT_USER=minio
MINIO_STORAGE_CLASS_STANDARD=EC:2
# Root secret for the server.
MINIO_ROOT_PASSWORD=minio123
MINIO_PROMETHEUS_AUTH_TYPE="public"
MINIO_PROMETHEUS_URL=http://data1:9090
MINIO_PROMETHEUS_AUTH_TYPE="public"
```

1.5 Client Setup

Each client was provided with a hostname matching the pattern client-{1...4}. We ran internal DNS to facilitate inter-node communication.

The WARP tool was downloaded onto each client.

2. Understanding Hardware Performance

2.1 Measuring Single Drive Performance

The performance of each drive was measured using the command dd. dd is a UNIX tool used to perform bit-by-bit copy of data from one file to another. It provides options to control the block size of each read and write.

Here is a sample of a single NVMe drive's Write Performance with 16MB block-size, O_DIRECT option for a total of 64 copies. Note that we achieve greater than 3.3 GB/sec of write performance for each drive.

```
$ dd if=/dev/zero of=/mnt/drive/test bs=16M count=10240 oflag=direct
Drive2-Write
10240+0 records in
10240+0 records out
42949672960 bytes (43 GB, 40 GiB) copied, 12.9619 s, 3.3 GB/s
```

Here is the output of a single HDD drive's Read Performance with 16MB block-size using the O_DIRECT option and a total count of 10240. Note that we achieved greater than 4.6 GB/sec of read performance for each drive.

```
$ dd of=/dev/null if=/mnt/drive/test bs=16M count=10240 iflag=direct
Drive2-Read
10240+0 records in
10240+0 records out
42949672960 bytes (43 GB, 40 GiB) copied, 9.31661 s, 4.6 GB/s
```

2.2 Measuring JBOD Performance

JBOD performance with O_DIRECT was measured using iozone. Iozone is a filesystem benchmark tool that generates and measures filesystem performance for read and write among other operations. Following is an example Iozone command operating with 160 parallel threads, 4MB block-size and O_DIRECT option.

```
iozone -s 1g -r 4m -i 0 -i 1 -i 2 -I -t 160 -b `hostname`-iozone.out
-F /mnt/drive1/tmpfile.{1..16} /mnt/drive2/tmpfile.{1..16}
/mnt/drive3/tmpfile.{1..16} /mnt/drive4/tmpfile.{1..16}
/mnt/drive5/tmpfile.{1..16} /mnt/drive5/tmpfile.{1..16}
/mnt/drive6/tmpfile.{1..16} /mnt/drive7/tmpfile.{1..16}
/mnt/drive8/tmpfile.{1..16} /mnt/drive9/tmpfile.{1..16}
/mnt/drive10/tmpfile.{1..16}
```

We measured 56.1 GB/sec of read throughput and 30.2 GB/sec of write throughput on a single node. Combining the results from all nodes, we calculated a rough estimate of total cluster throughput as 225.0 GB/sec of read throughput and 120.4 GB/sec of write throughput.

2.3 Network Performance

In virtually all cases with MinIO, the network is the bottleneck. MinIO takes full advantage of the available underlying server hardware. In this test, we used separate networks for server-client and server-server communication.

Although the server hardware could support 200 Gbps, we only had access to a 100 Gbit/sec switch. The network hardware on these nodes allows a maximum of 200 Gbit/sec, but we were limited by a 100 Gbit/sec Ethernet switch. 100 Gbit/sec equates to 12.5 Gbyte/sec (1 Gbyte = 8 Gbit).

Therefore, the maximum throughput that can be expected from each of these nodes would be 12.5 Gbyte/sec.

There are 4 nodes, making the theoretical maximum GET throughput 50 GB/sec (400 Gbps) and PUT throughput 25 GB/sec (200 Gbps). Note that such theoretical maximums aren't achievable in the real world given the overhead imposed by TCP and Ethernet technologies. When Ethernet, TCP and IP header overhead are taken into account along with other factors such as preamble size, inter-frame gap and frame size, many switched networks run at about 80%-90% efficiency. In the case of a 100 Gbps network, that translates to real-world available file transfer speeds between 40 GB/sec and 45 GB/sec.

3. Running the 4-node Distributed WARP S3 Benchmark

We ran the client process on the client nodes as:

```
$ warp client
```

GET test

We ran WARP from one node connected to four clients to initiate a series of the benchmark tests that vary object size, number of threads, enabling/disabling multipart, enabling/disabling on-disk encryption and enabling/disabling TLS. Below is an example.

```
warp get --host data{1...4}:9000 --warp-client clt{1...4}
--access-key minio --secret-key minio123 --duration 5m --obj.size
256M  --disable-multipart --concurrent 32 --tls -insecure
```

PUT test

We ran WARP from one node connected to four clients to initiate a series of the benchmark tests that vary object size, number of threads, enabling/disabling multipart, enabling/disabling on-disk encryption and enabling/disabling TLS. Below is an example.

```
warp get --host data{1...4}:9000 --warp-client clt{1...4}
--access-key minio --secret-key minio123 --duration 5m --obj.size
256M  --disable-multipart --concurrent 32 --tls -insecure
```

In the above commands:

- *minio.ip.or.dns.{1...4}* are the MinIO servers. The ellipses expands the hostnames as *minio.ip.or.dns.1 minio.ip.or.dns.2 minio.ip.or.dns.3 minio.ip.or.dns.4*
- *warp.ip.or.dns.{1...4}* are the warp clients. The ellipses expands the hostnames as *warp.ip.or.dns.1 warp.ip.or.dns.2 warp.ip.or.dns.3 warp.ip.or.dns.4*

3.1 Results

The throughput of the combined cluster of 4 Supermicro Cloud DC servers each backed by 10 NVMe drives, as measured from the clients using WARP is presented below:

Setup	Avg Read Throughput (GET)	Avg Write Throughput (PUT)
Distributed	42.57 GB/s	24.69 GB/s

3.2 Interpretation of Results

The average network bandwidth utilization during both read and write operations fluctuated between 95% and 100% during testing. This represents client traffic as well as inter-node traffic.

The network was almost entirely choked during these tests. Higher throughput can be expected on higher bandwidth networks.

Note that the write benchmark is slower than read because benchmark tools do not account for write amplification (traffic from parity data generated during writes). In this case the 100 Gbps network is, again, the bottleneck as MinIO gets close to hardware performance for both reads and writes.

3.3 Impact of encryption

The results of the same tests with encryption enabled are presented below:

Setup	Avg Read Throughput (GET)	Avg Write Throughput (PUT)
Distributed	42.57 GB/s	24.69 GB/s
Distributed with Encryption	42.54 GB/s	25.96 GB/s
Distributed with TLS	42.35 GB/s	24.42 GB/s
Distributed with TLS and encryption	42.41 GB/s	23.88 GB/s

The impact of encryption is negligible for reads and writes. Minimal reduction in throughput can be accounted for by the overhead of decrypting/encrypting the objects while reading and writing. The apparent improvement in Write/PUT performance by adding encryption is anomalous but not unheard of. WARP results contain insignificant differences that are sporadic and vary between test runs. Adding TLS reduces write throughput further, although the reduction is negligible.

With encryption and TLS enabled, the overall speed of reads is unchanged while that of writes is still very high in comparison to the maximum available bandwidth. Therefore we strongly recommend turning on TLS and encryption for all externally exposed production setups.

4. Conclusion

Based on the results above, we found that a 4 node cluster of Supermicro Cloud DC servers each with 10 NVMe drives provides an outstanding platform for running distributed, high-performance workloads on MinIO. Further, our testing found negligible performance decreases from enabling encryption and TLS - ensuring that even high-performance workloads can run encrypted at all times.

Finally, while performance will increase on a near linear basis with additional servers, bandwidth will often be the bottleneck and architects should build with those constraints in mind.

5. Appendix A

Please see below for hardware specifications:

Setup	4 x SYS-620C-TN12R
CPU	2 x P4X-ICX8368-SRKH8 - ICX 8368 2P 38C/76T 2.4G 57M 11.2GT 270W 4189 D2
Memory	16 x MEM-DR432L-HL03-ER32 - SK Hynix 32GB DDR4-3200 2Rx8 (16Gb)ECC REG DIMM
NVMe M.2(OS)	2 x HDS-SMN1-MZ1LB960HAJQ07 - Samsung PM983 960GB NVMe PCIe3x4 V4 M.2 22x110mm (1.3 DWPD)
NVMe(OSD Drives)	10 x Intel SSDPF2KX038TZ 3.84TB Drive NVMe PCIe 4.0 U.2 15mm 1DWPD
AOC	1 x AOC-S100G-b2C - BCM57508 NetXtreme-E 10Gb/25Gb/40Gb/50Gb/100Gb/200Gb Ethernet
AOC II	1 x AOC-ATG-i2TM - AIOM 2-port 10GBase-T Intel X550 RoHS