White Paper

Benchmarking BeeGFS and OpenZFS with the Western Digital® OpenFlex™ Data24 NVMe-oF™ Storage System

Prepared for Western Digital Platforms by Atipa Technologies.

Abstract

When ZFS was tuned for optimal performance, a resultant 4X-5X increase in read/write bandwidth performance was noted without the storage system being impacted.

February 2025

Table of Contents

About Atipa Technologies

Atipa Technologies is a U.S.- based High-Performance Computing and Storage Solution provider. By adopting industry-leading technologies and utilizing rigorous burn-in procedures, Atipa has delivered trusted HPC solutions for over two decades. Atipa has received more than 60 awards on the Top500 list of the fastest supercomputers in the world, including #13 in November 2013. Atipa's goal is to deliver affordable solutions tailored to each customer's specific needs, enabling groundbreaking research in government agencies and universities small and large. Atipa is an Intel® Platinum HPC Data Center Specialist, AMD Elite Solution Provider, NVIDIA Elite Partner, BeeGFS Gold Partner, and Bright Computing Premier Partner.

About BeeGFS

BeeGFS (formerly known as FhGFS - developed at the Fraunhofer Institute for Industrial Mathematics ITWM) is a parallel file system designed for high-performance computing (HPC) and enterprise environments, developed with a strong focus on performance and ease of installation and management. If I/O intensive workloads are your problem, BeeGFS is the solution.

BeeGFS is built on a client-server architecture, where multiple client nodes access data transparently distributed over multiple server nodes. By increasing the number of servers and disks in the system, performance and capacity of the file system scales seamlessly to the level needed by HPC clusters from tens to thousands of nodes. BeeGFS also provides features like automatic failover and dynamic load balancing, ensuring high availability and optimal performance even in the presence of hardware or network failures.

OpenFlex Data24 NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 NVMe-oF storage platform extends the high performance of NVMe™ flash to shared storage. It provides low latency sharing of NVMe SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs. Unsurpassed connectivity in its class using Western Digital RapidFlex™ NVMe-oF controllers, allows up to six hosts to be attached without a switch (like a traditional JBOF). The OpenFlex Data24 uses Western Digital's RapidFlex Adapters to provide 2, 4, or 6-ports of 100GbE connected to RDMA (RoCEv2) configured host initiators.

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 to respond to application needs.

The OpenFlex Data24 can also be used as a disaggregated storage resource in an open composable infrastructure environment using the Open Composable API. The platform can also be specified with just two RapidFlex adapters for simpler environments and as a direct replacement for SAS external storage.

The OpenFlex Data24 design exposes the full performance of the dual port NVMe SSDs to the network. With 24 Western Digital Ultrastar® DC SN840 3.2 TB[1] devices, the enclosure can achieve up to 71.4 GB/s of 128K bandwidth and over 16.7 MIOPS at 4K block size.

Ultrastar SN840

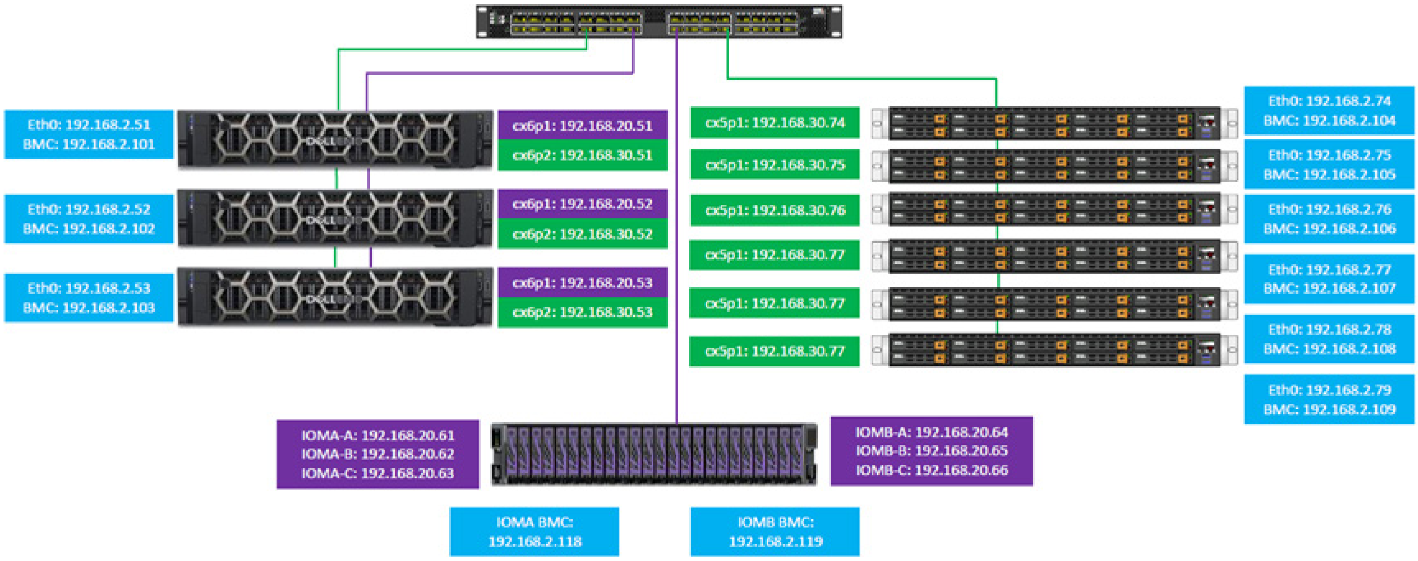
At the core of the OpenFlex Data24 NVMe-oF Storage Platform are Western Digital Ultrastar DC SN840 NVMe SSDs. The Ultrastar DC SN840 is a performance NVMe SSD targeting cloud compute and enterprise workloads that require low latency to data and high availability of data. The DC SN840 is Western Digital's 3rd generation of performance NVMe SSD for data center and extends Western Digital's leadership in dual-port architecture by vertically integrating proven flash controllers. Utilizing 96-layer 3D TLC NAND, it is available in capacities from 1.6TB to 15.36TB in a standard, front-loading 2.5" U.2 form factor.

[1] 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.

The Solution Details

The benchmarked solution is comprised of:

- Six BeeGFS client servers, each configured with a single NVIDIA® Mellanox® CX-5 100GbE Ethernet adapter.
- Three BeeGFS host servers, each configured with two NVIDIA Mellanox CX-6 100GbE Ethernet adapters with multipath connectivity to the OpenFlex Data24
- One single 64-port 200GbE NVIDIA Spectrum SN4000 series switch.
- One single OpenFlex Data 24 with 24 x SN840 7.68TB dual ported SSD.
  - Two IOM modules with 3 x 100GbE AIC per IOM.



Server Infrastructure Detail

	Servers	Clients
Models	Dell® R750	SuperMicro® SYS-2029BT-HNR
System Function	BeeGFS storage server	Load Generation
Boot Drive	Local	Local
PCIe Generation	4	3
CPU	Intel® Xeon® Gold 6354 @3.0GHz	Intel Xeon Gold 6150 @2.70GHz
Memory	512GiB @ 3200 MHz	384GiB @2666MHz
RNIC	Mellanox CX6 x 2	Mellanox CX5 x 1
Fabric Connection	2	1

Benchmarks

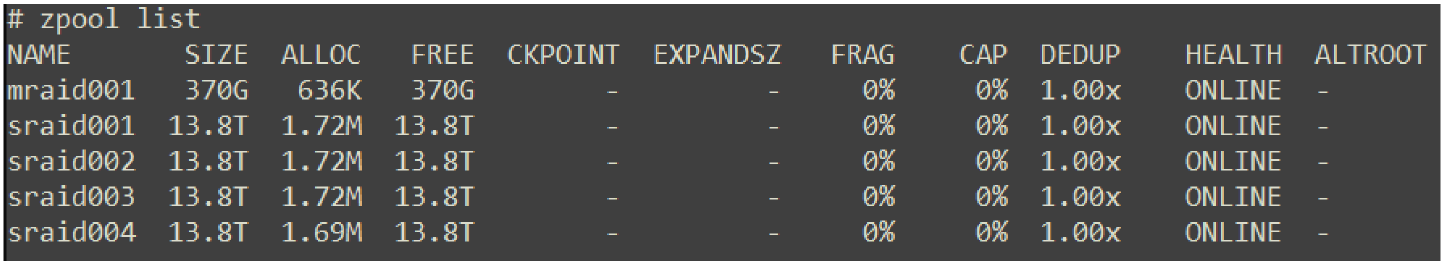
To measure the performance of the BeeGFS parallel file system, IOR was run on 6 BeeGFS client servers.

IOR is a parallel IO benchmark that can be used to test the performance of parallel storage systems using various interfaces and access patterns. The IOR repository also includes the mdtest benchmark which specifically tests the peak metadata rates of storage systems under different directory structures. Both benchmarks use a common parallel I/O abstraction backend and rely on MPI for synchronization.

Each Ultrastar DC SN840 NVMe SSD was partitioned into one 100GB namespace for BeeGFS metadata and four 1.89TB namespaces for BeeGFS object storage. All namespaces were preconditioned by completely filling the namespace four times with sequential writes using a chunk size of 128k.

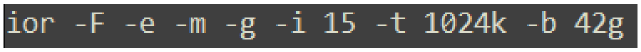
Next, OpenZFS was used to create metadata and storage targets for BeeGFS. OpenZFS distinguishes itself from other file systems by its built-in volume management and robust data integrity. OpenZFS also offers fast and efficient software RAID schemes equivalent to RAID-1, RAID-5, and RAID-6. The Copy-on-Write (COW) transactional model ensures data is always consistent on disk. This means that when data is changed, it is not overwritten, but instead it is written to a new block and checksummed before pointers to the most recent copy of the data are changed. User data is protected against silent data corruption by 256-bit checksums which are stored separate from the data. When a data block is read, OpenZFS calculates its checksum and compares it to the stored checksum. If the block is corrupted, it is repaired transparently and on the fly.

A striped zpool “mraid001” of four ZFS mirrors consisting of the 100GB namespaces was created on each server. The 1.89TB namespaces were configured as four ZFS RAID-Z2 (6+2) pools “sraid00[1-4]” consisting of one 1.89TB namespace per NVMe SSD. All zpools were subsequently formatted as ZFS file systems, resulting in the following ZFS storage pool layout:



```
# zpool list
NAME        SIZE  ALLOC   FREE  CKPOINT  EXPANDSZ   FRAG    CAP  DEDUP    HEALTH  ALTROOT
mraid001   370G   636K   370G        -         -     0%     0%  1.00x    ONLINE  -
sraid001  13.8T  1.72M  13.8T        -         -     0%     0%  1.00x    ONLINE  -
sraid002  13.8T  1.72M  13.8T        -         -     0%     0%  1.00x    ONLINE  -
sraid003  13.8T  1.72M  13.8T        -         -     0%     0%  1.00x    ONLINE  -
sraid004  13.8T  1.69M  13.8T        -         -     0%     0%  1.00x    ONLINE  -
```

Before implementing the BeeGFS parallel file system, the IOR parallel file system benchmark was used to measure and tune the performance of the zpools. Since flash storage benchmarks are inherently prone to run-to-run variability, each IOR run was iterated 15 times using the following command syntax:



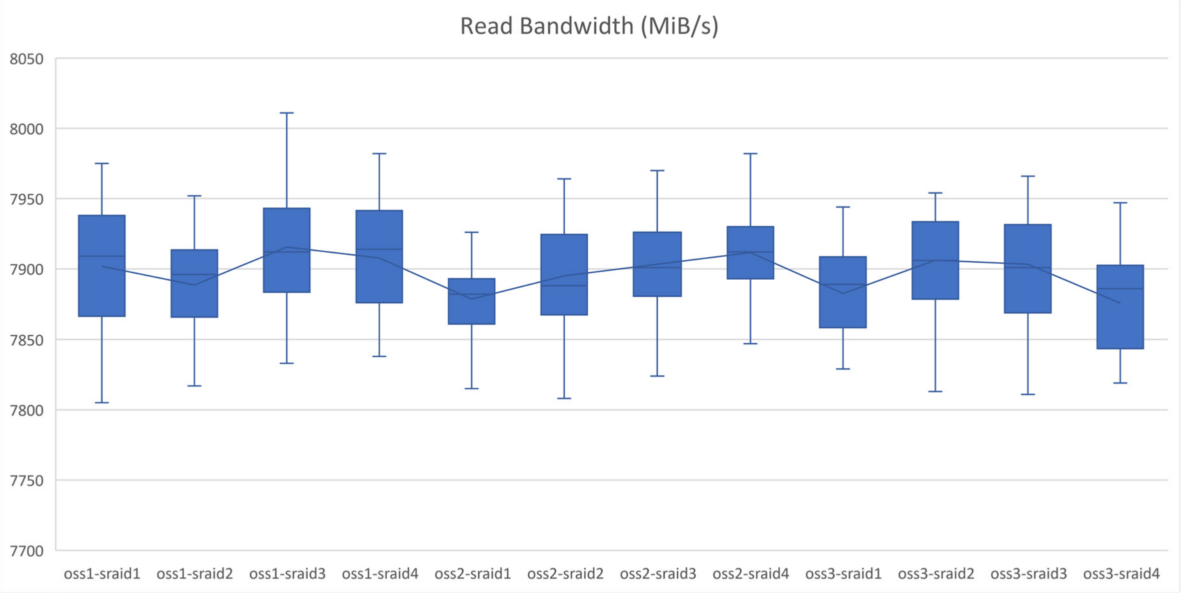
```
ior -F -e -m -g -i 15 -t 1024k -b 42g
```

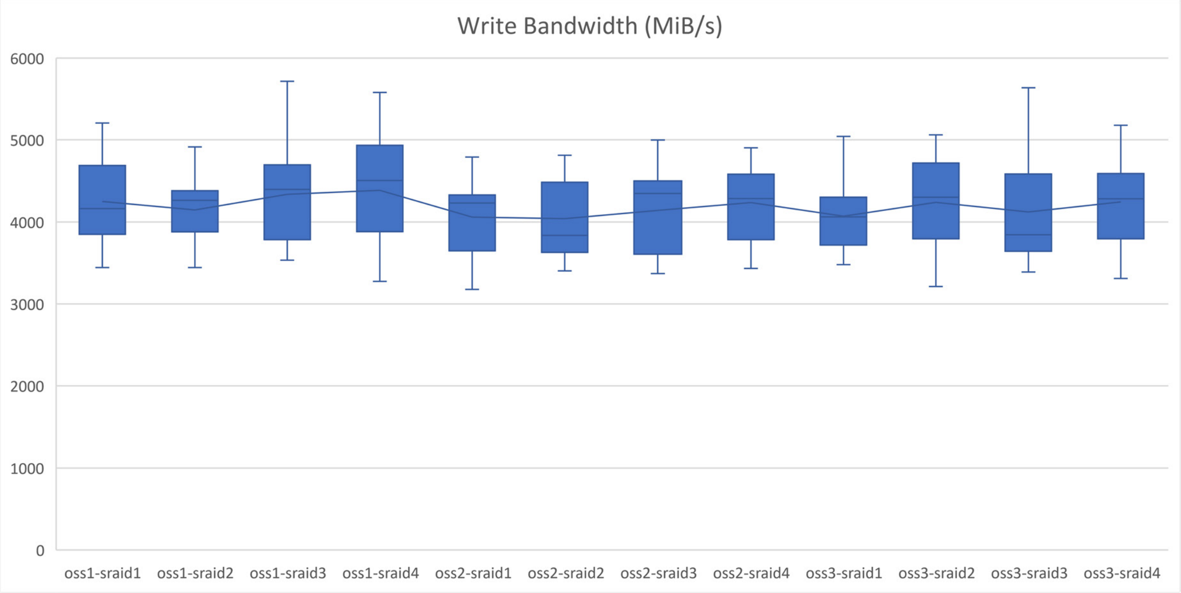
The resulting sequential read and write bandwidths were analyzed using “box and whisker” plots showing the minimum, maximum, first, second, and third quartile, and average of each 15-run data set. The IOR command was launched on all 36 cores on each object storage server using OpenMPI’s mpirun, resulting in a 1.5TB total aggregated file size read and written during each IOR iteration.

Initial IOR runs resulted in substantial performance variability and occasional performance “cliffs”. More preconditioning was therefore done directly on the ZFS storage pools. We attribute the need for additional preconditioning to ZFS’s variable record size causing a mismatch between the initial precondition on the unformatted namespaces and the ZFS write patterns invoked by IOR.

Baseline ZFS Sequential Read/Write Performance

The baseline sequential read/write performance before any file system tuning is displayed in the following diagrams:

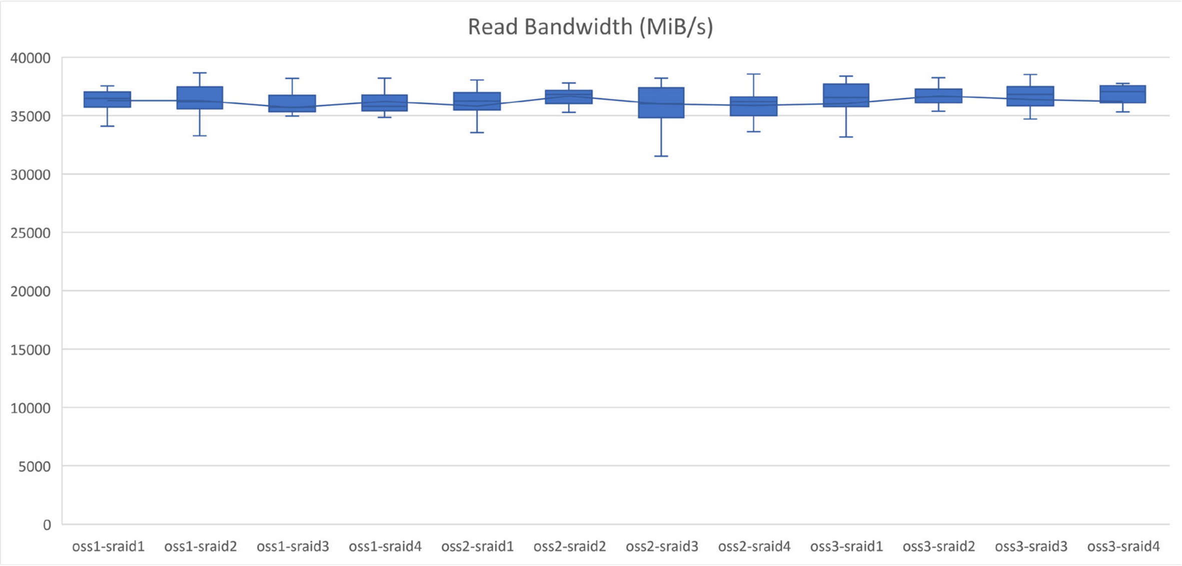


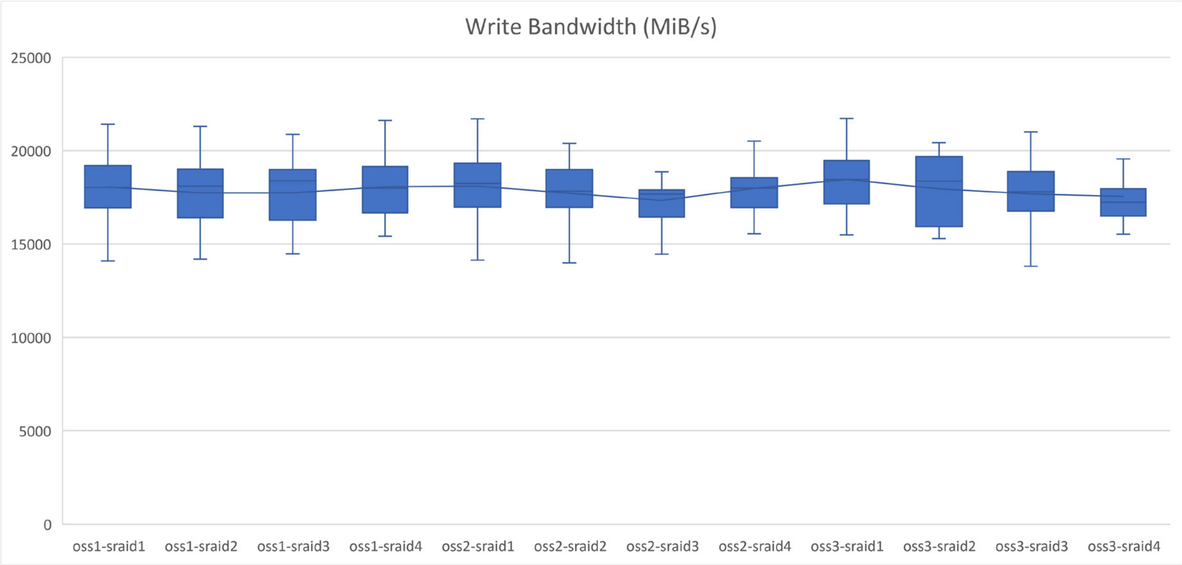


Optimized ZFS Sequential Read/Write Performance

ZFS was tuned for optimal performance, resulting in a 4X-5X increase in read/write bandwidth. The optimized sequential read/write performance is displayed in the following diagrams.

The optimized ZFS volumes are the building blocks for the BeeGFS parallel file system. The striped zpool mirrors “mraid001” on two of the servers were configured as metadata targets, while each of the RAID-Z2 pools “sraid00[1-4]” on all 3 servers were configured as storage targets.





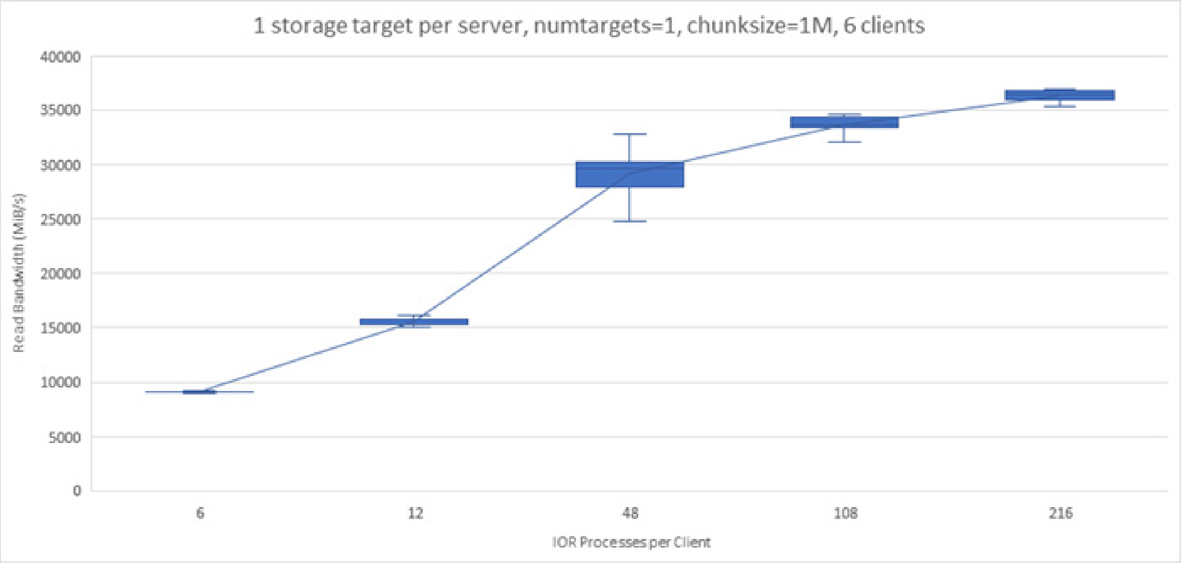
For more information on how the OpenFlex Data24 NVMe-oF Storage Platform can benefit your environment and improve business operations, visit: https://www.westerndigital.com/products/data-center-storage/data-center-platforms.

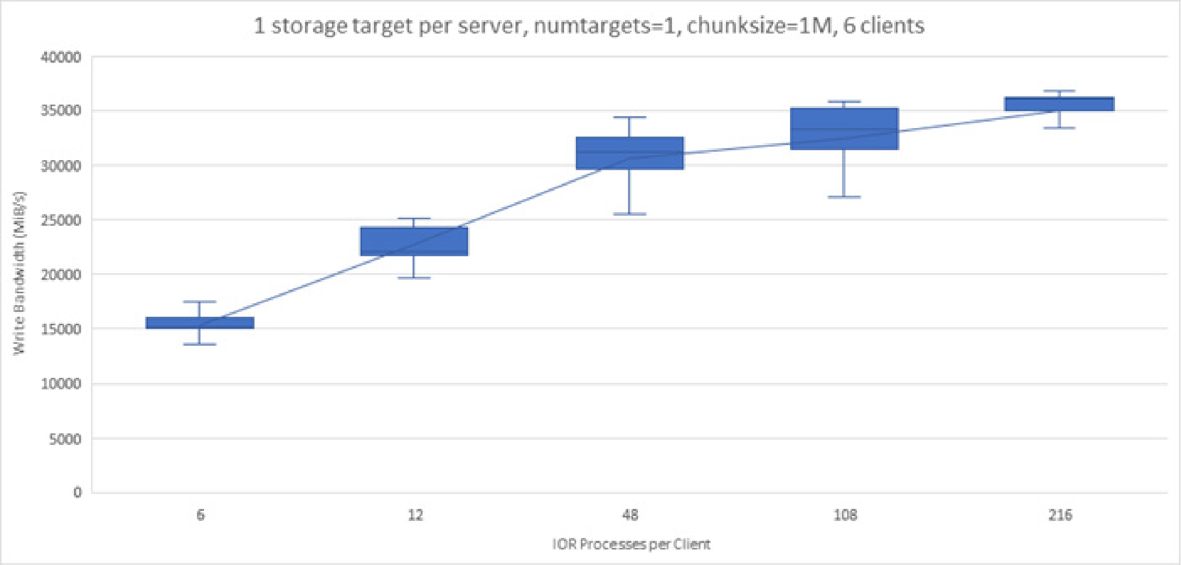
BeeGFS Sequential Read/Write Performance

The BeeGFS file system was setup configuring the ZFS pools “mraid001” and “sraid00[1-4]” as metadata and storage targets, respectively. With a single IOR thread, the resulting BeeGFS filesystem achieved 1.6GB/s read and 3.1GB/s write throughput. When all 36 cores on a single client were utilized, the read and write throughput maxed out at 11GB/s, fully saturating the 100GbE network link.

The read and write throughput achieved from all 6 client servers was analyzed as a function of the number of IOR processes per client, depicted in Figures 5 and 6. Throughout the benchmarks, each storage server used 1 storage target. Test runs employing 2 or more storage targets per storage server showed no discernible impact on the read/write throughput when compared to using a single storage target per server.

Figure 5 illustrates the increase in aggregate read throughput from 9GB/s using 1 IOR process per client to 36GB/s using 36 IOR processes per client (utilizing all available client cores). Similarly, Figure 6 showcases the aggregate write throughput ascending from 15GB/s with 1 IOR process per client to 35GB/s using 36 IOR processes per client.





Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com