TOSHIBA

Lab Report

In Focus: Performance, Storage Efficiency and Reliability

Achieving 1.2 petabytes of storage: How 60 Hard Disk Drives could be optimized and managed for the media industry

Author: Rainer W. Kaese, Senior Manager
Business Development, Storage Products Division,
Toshiba Electronics Europe GmbH

Introduction

Streaming services have revolutionized the way that people access entertainment and factual information. Ensuring that consumers can access content online wherever and whenever they like requires a significant amount of reliable data storage. When considering the storage and access of decentralized content from all over the world to enable flawless workflows in a multi-cloud environment, the Hard Disk Drive (HDD) might not be the first technology that springs to mind. Nevertheless, their usage in the media & entertainment industry is inevitable as HDDs deliver exabytes of reliable storage at reasonable cost.

Picture 1: Reference and Demo System at IBC 2023 presented by Irina Chan and Roland Frei, Storage Products Division, Toshiba Electronics Europe GmbH

Toshiba Electronics Europe GmbH (Toshiba) has showcased their data storage portfolio at the International Broadcasting Convention (IBC) in Amsterdam for eight consecutive years. The focus for 2023 were the latest 20TB HDD models as based storage components for digital media and entertainment data storage. Together with our partners Promise, AIC, ATTO, Nanya and Open-E, Toshiba ran a live demonstration of a compact reference storage system optimized for the capacity and performance requirements of the media industry. In this lab report, Senior Manager Rainer W. Kaese describes the reference system in detail, summarizes the live demo results presented at the exhibition, and closes with more detailed performance and configuration benchmarks.

The partners

The HDD storage system was integrated with hardware and software from Toshiba's trusted partners:

- **Promise Technology** for the 60-Bay top loader JBOD
- **ATTO** for the Host Bus Adapter Cards to connect JBOD and internal drives
- **AIC** for the head node server
- **Nanya Technology** for the DRAM modules of the headnote server
- **Open-E** for the linux based ZFS software defined storage Open-E JovianDSS
- **ATTO** for the 100GbE Network Interface Cards to connect to the application server

In detail:

Head node hardware:

Server:	2U 12x 3.5" HDD Rackmount (AIC SB202-TU, SKU XP1-S202TU01)
CPU:	Xeon® Gold-5318Y (2x)
Memory:	128GB / 4x 32GB DDR4 3200 RDIMM (Nanya NT32GA72D4NFX3K-JR)

Picture 2:
SB202-TU head node server from AIC

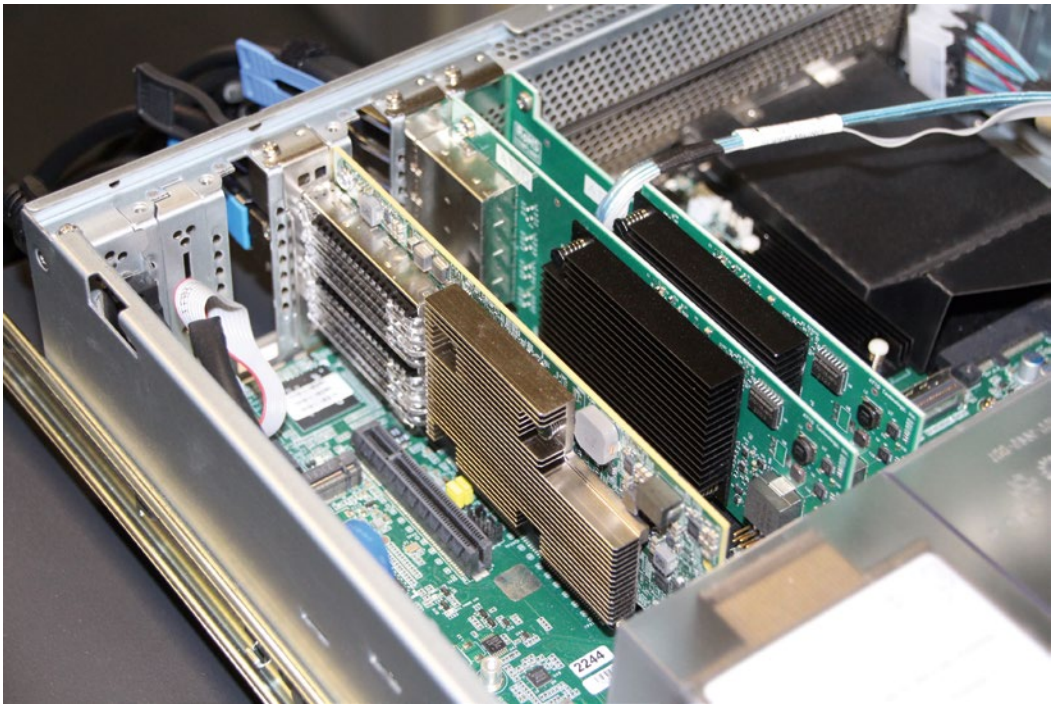
Network:	2x 100GbE (ATTO NIC FFRM-N412)
SAS HBA:	16e (4x SFF-8644) SAS12G HBA (ATTO HBA ESAH12F0) 16i (4x SFF-8643) SAS12G HBA (ATTO HBA ESAH120F)
Cache SSD:	3x Enterprise SSD SAS12G 800GB (KIOXIA KPM51MUG800G)

Head node software:

ZFS:	Open-E JovianDSS 1.0up30.52540

Disk enclosure:

Model:	4U 60-Bay JBOD (PROMISE VTrak J5960 4U-SAS-60-D BP)
Cables:	4x SFF-8644 to SFF-8644 3m
HDDs:	60x 3.5" Enterprise SAS 7200rpm 20TB (Toshiba MG10SCA20TE, FW 0101)



Picture 3:
ATTO Add-In-Cards (from left to right: 100GbE network, HBA for external, HBA for internal SAS connections)

Picture 4:
Promise J5960 JBOD with open lid

Picture 5:
Toshiba MG10SCA20TE HDD in Promise tray

ZFS configuration:

Pool:	6 Groups of 10 Disks each in RAID-Z2 (double parity) 1x SSD 800GB Read Cache 2x SSD 800GB Write Logs (mirror)
Total Storage:	1200 TB gross, 960 TB net, with max. 90% filling: 864TB usable

Theoretical considerations and practical experiments had shown that six groups of 10 disks in RAID-Z2 configuration are a good basis for performance optimization.

Redundancy levels

This RAID-Z2/RAID6 based configuration is in line with good practice to use (at least) a double redundancy when building data storage based on high capacity HDDs. The rebuild times of a failed 20TB HDD can be very long, so the data would be unprotected during this time in the case of a single redundancy or single mirror configuration. Together with the high workload in an array under rebuild, the failure of an additional disk could result in data loss. Hence, under rebuild, the array should still be protected by the second redundancy data set. This means that RAID-Z2, RAID6 or triple mirroring is a must

when using HDDs of >12TB. Open-E recommends to use them already for HDDs of more than 4TB of capacity.

Open-E storage and RAID calculator

https://www.open-e.com/r/9tsr/

Open-E web storage and RAID calculator is a good reference tool for configuring a ZFS storage pool (Figure 1).

Figure 2 is an overview of the pool configuration used for the live demo.

On this ZFS pool, we created a “Zvol” of 800TB, and configured an iSCSI LUN attached to it (Figure 3).

N.B: Write cache synchronization requests are disabled. This achieves highest writing performance but comes with a risk as most recently cached data could be lost in the event of an un-expected power outage. Hence this configuration should only be used where protection against power outage is in place.

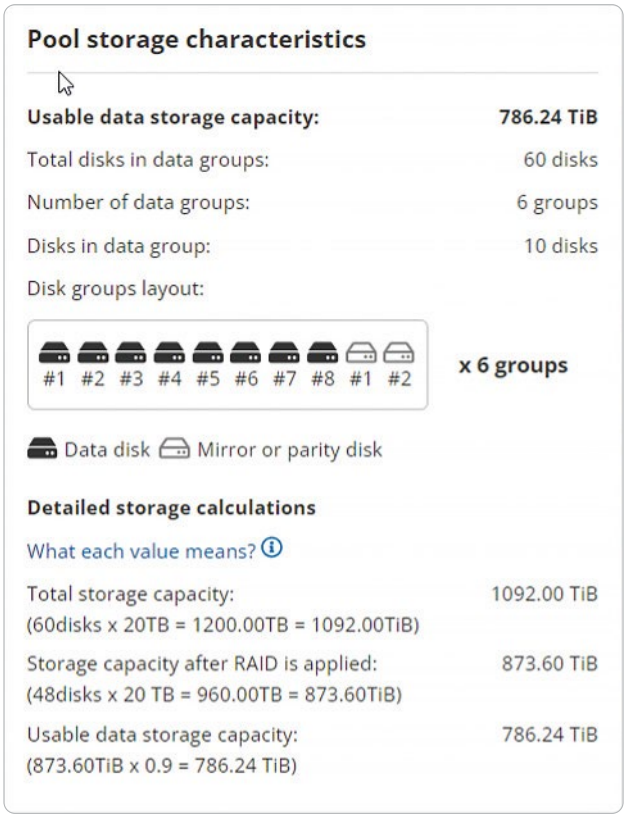


Figure 1: Open-E JovianDSS Storage and RAID Calculator

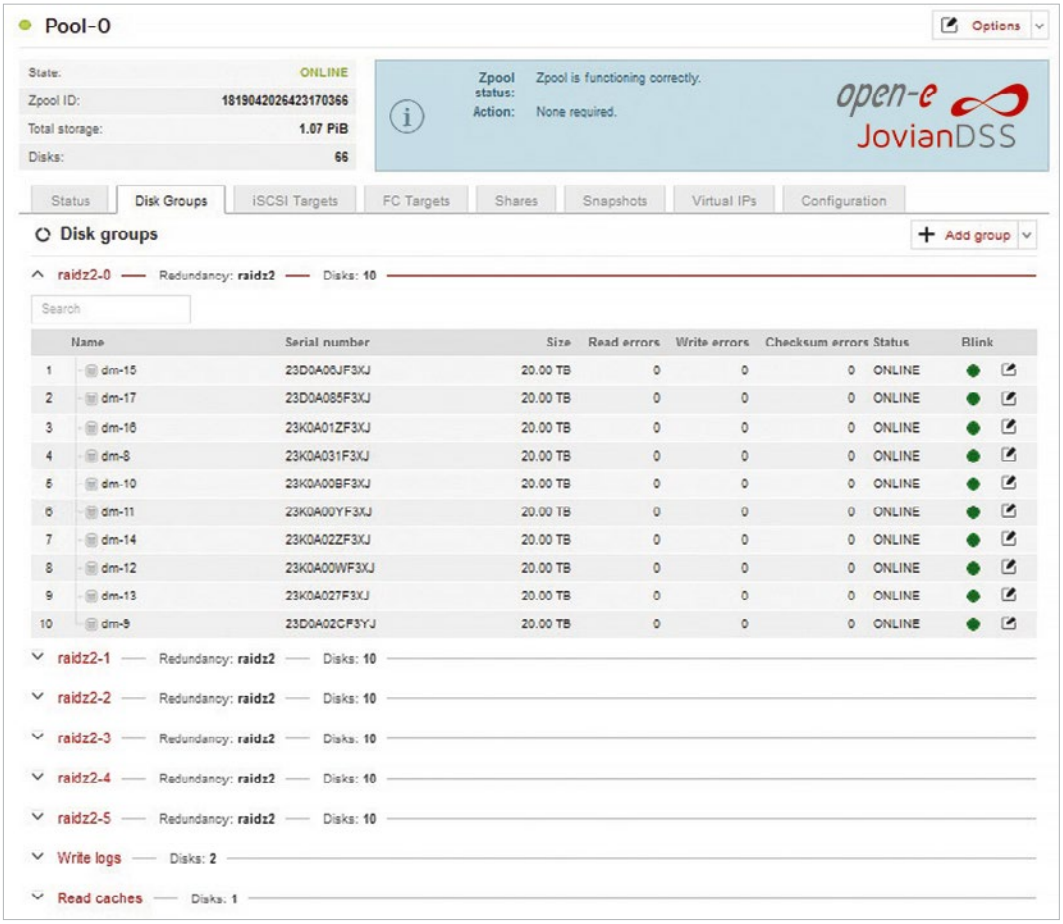


Figure 2:
Zpool configuration in Open-E JovianDSS WebGUI

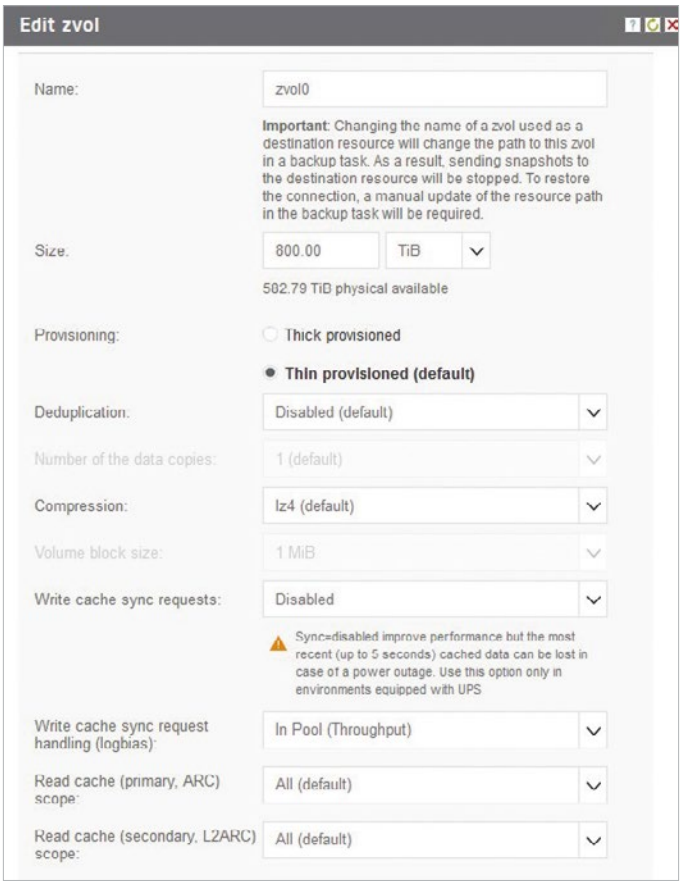


Figure 3: iSCSI block storage configuration in Open-E JovianDSS WebGUI

This iSCSI storage block was connected via a 100GbE network to the application server. A windows logic drive was formatted at the application server.

The performance demo was driven by a „fio“ script of a read-dominated workload from a 100TB test file, followed by writing onto a 100TB storage space.

A reading performance of more than 5GB/s and writing bandwidths of around 2.5GB/s were demonstrated. This matches well with the bandwidth of the 100GbE network infrastructure connection between storage head node and application server.

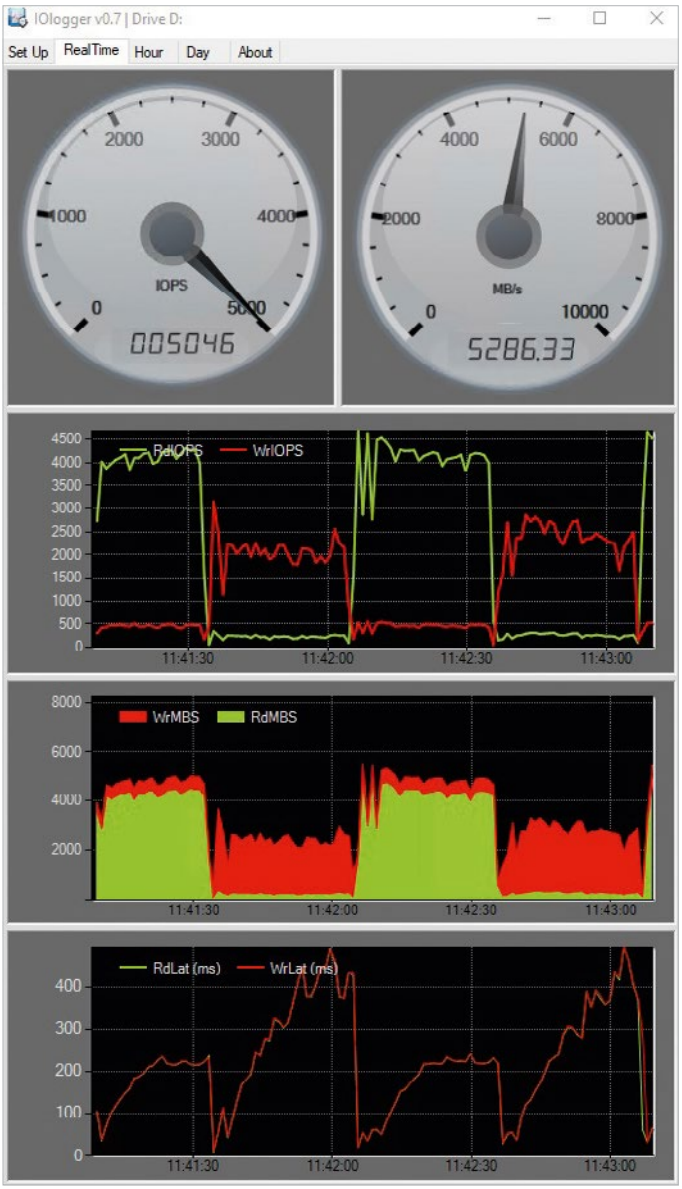


Figure 4: Read- and write- dominated workload demo

Benchmark of different write cache settings

In the Toshiba HDD application laboratory we benchmarked the configuration for different write cache sync request settings, as follows.

Demo script:

```
:a
fio --filename=test --size=100T --direct=1 --rw=rw --rwmix=90 --bs=1m --iodepth=64 --time_based
--runtime=30 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=16
--norandommap --randrepeat=0 --output=seqreadlogical.log
fio --filename=test --size=100T --direct=1 --rw=rw --rwmix=10 --bs=1m --iodepth=64 --time_based
--runtime=30 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=16
--norandommap --randrepeat=0 --output=seqwritelogical.log
goto a
```

Configuration with disabled write cache sync requests

Write Cache Sync Requests: Disabled
Write Cache Sync Request Handling: In Pool / Write log device (does not matter)

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		1570
Sequential read 1M		3830
Random write 4k	1000	
Random read 4k	1270	
Mixed 4k/64k/256k/1M	740	220

This configuration achieves highest write performance, but the cache data protection limitation mentioned earlier should be taken into account.

Configuration with enabled write cache sync requests handled on the SSD based write log device

Write Cache Sync Requests: Enabled
Write Cache Sync Request Handling: Write log device

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		790
Sequential read 1M		3880
Random write 4k	970	
Random read 4k	1270	
Mixed 4k/64k/256k/1M	730	220

Here the sequential write performance drops to approx. 800 MB/s level (which is still high). Write IOPS and all reading performance are in the same range.

Configuration with the write cache synchronizing directly into the HDD pool, without using the SSD write logs

Write Cache Sync Requests: Enabled
Write Cache Sync Request Handling: In Pool

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		152
Sequential read 1M		3850
Random write 4k	360	
Random read 4k	1270	
Mixed 4k/64k/256k/1M	360	107

The write performance drops significantly using this setup, so this configuration should only be used in cases where the write performance is not important. Generally, the proper usage of the SSD based write log is advised.

Benchmark of different HDD pool configurations

When discussing appropriate pool configurations, this question arose: would using a simpler mirror technology to avoid the parity calculation effort of the previous RAID-Z2/RAID6 configuration result in higher performance?

To keep the double redundancy, the pool configuration for a mirror-based array would be 20 groups of three-way mirrors.

Reference: “fio” script used for benchmarking:

```
fio --filename=test --size=100T --direct=1 --rw=read --bs=1m --iodepth=64 --time_based
--runtime=1h --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=16
--norandommap --randrepeat=0 --output=seqreadlogical.log

fio --filename=test --size=100T --direct=1 --rw=write --bs=1m --iodepth=64 --time_based
--runtime=1h --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=16
--norandommap --randrepeat=0 --output=seqwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=randread --bs=4k --iodepth=512 --time_based
--runtime=1h --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randreadlogical.log

fio --filename=test --size=1T --direct=1 --rw=randwrite --bs=4k --iodepth=512 --time_based
--runtime=1h --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=randrw --bssplit=4k/20:64k/50:256k/20:2M/10
--iodepth=512 --time_based --runtime=1h --group_reporting --name=job1 --ioengine=windowsaio
--thread --numjobs=64 --norandommap --randrepeat=0 --output=mixedlogical.log
```

Performance for 20 group of three-way mirrors

Write Cache Sync Requests: Disabled
Write Cache Sync Request Handling: In Pool / Write log device (does not matter)

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		1270
Sequential read 1M		2480
Random write 4k	890	
Random read 4k	1360	
Mixed 4k/64k/256k/1M	890	260

Write Cache Sync Requests: Enabled
Write Cache Sync Request Handling: Write log device

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		640
Sequential read 1M		2530
Random write 4k	870	
Random read 4k	1360	
Mixed 4k/64k/256k/1M	890	890

The sequential performance results are lower compared to the previous configuration of 10 groups of RAID-Z2, since fewer disks are accessed in parallel. Random workload is handled with similar performance, and performance for mixed workload is slightly higher.

But as the pool efficiency of a three-way mirror (33%) is far lower compared to RAID-Z2 (80% in this case), this configuration is not recommended for 60 high capacity HDDs.

Performance for 30 groups of two-way mirrors

Finally we evaluated a configuration based on a pool of 30 groups of two-way mirrors (equivalent to RAID10). From practical experience, RAID10 is not the most efficient configuration for optimum capacity (due to the two-way mirrored data, usable capacity is just half of installed capacity) but it is widely used, especially when storages are optimized for (random) performance and general agility.

However, it should be noted that RAID10/two-way mirroring is based on single redundancy. If the HDDs need to be read orrestored in the event that rebuilding does fail, data is lost.

Write Cache Sync Requests: Disabled
Write Cache Sync Request Handling: In Pool / Write log device (does not matter)

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		1680
Sequential read 1M		4280
Random write 4k	1000	
Random read 4k	1450	
Mixed 4k/64k/256k/1M	870	260

Write Cache Sync Requests: Enabled
Write Cache Sync Request Handling: Write log device

Workload	IOPS	Bandwidth (MB/s)
Sequential write 1M		600
Sequential read 1M		4330
Random write 4k	990	
Random read 4k	1310	
Mixed 4k/64k/256k/1M	890	260

As expected, this configuration shows highest performance, but the delta to the RAID-Z2/RAID6 based initial setup is not significant. As pool capacity efficiency and level of data protection are also lower, we recommend using this solution only for very specific cases where the top priority is sequential performance or sustained performance for very long bursts of random data.

For all other use cases, groups of RAID-Z2/RAID60 have been clearly proven as the best configuration in terms of performance, storage efficiency and data protection level.

HDD failure rates and reliability calculations

To assess the probability of disk failure and its impact, we considered a service lifetime of five years for the storage system, which is in line with the typical manufacturer's warranty for enterprise components such as HDDs. The MTTF (Mean Time To Failure) of the Toshiba MG10SCA20TE HDD is listed as 2.5 million hours, which equals to an expected statistical annual failure rate (AFR) of 0.35%: meaning out of 10 000 drives in operation, 35 drives may fail per year.

A simple calculation tells us that, out of 60 installed drives, a theoretical count of 0.21 drives may fail in one year. Over five years this will add up to a probability of 1.05 drives failing. In other words, one drive failure over storage must be expected.

With a double parity, this is no issue. When the failed drive is replaced and the data is rebuilt on the replacement drive, the data is still protected by the second parity even during the heavy load

of this rebuilding process. Even if the failed drive is not replaced at all, the data is still protected by the second parity.

But failure probabilities are statistical values and in practice they can happen at any time. If a second drive fails within five years of runtime the consequences depend on which drive failed. If it is a drive in another group of RAID-Z2, the outcome is the same – no issue. If it is in the same drive group where one drive has failed already (and is still under rebuild or has not been replaced), the data will still exist, but it is not parity protected anymore. However, the probability of a second drive failing in the same group is extremely low. The probability for real data loss due to a third drive failing in the same group while the two parity drives are either missing or under rebuilding is negligible.

Summary

We have demonstrated that for 60 HDDs of high capacity, a ZFS pool of 6 groups RAID-Z2 (double parity) of 10 disks each (equivalent to RAID60) is the best configuration, not just for highest performance, but also as regards storage efficiency (the difference between total installed gross storage and usable net storage without redundancy overhead), reliability and data protection level.

The complete SDS reference system of 60 HDDs Toshiba 20TB Enterprise SAS was set up in a Promise 4U 60 bay JBOD connected via ATTOs HBA to AIC's server with Nanya DRAM as head node operating Open-E JovianDSS software to create a ZFS based storage system with 100GbE network connectivity over ATTOs network interface cards.

960TB of usable net storage were implemented based in 60 x 20TB (1200TB total). The access speed for this storage system reaches up to 4 GB/s reading and 2 GB/s for writing sequential data, matching with modern datacenter infrastructures based on 100GbE networks to distribute the bandwidth to the applications. With a 4k raw video stream at 8 bit color depth and 60 frames per second (requiring up to 1.5 GB/s of bandwidth) this solution is capable of handling two to three uncompressed raw 4k media streams, or up to 30 compressed ones.

At the same time, the double parity configuration of RAID-Z2/RAID6 provides unrivalled data protection against HDD failures.

So, the live demonstration has shown that the optimized configuration of Toshiba HDDs, supported by the outstanding teamwork with our partners, delivers excellent results regarding performance, storage efficiency and reliability. The result underlines that our HDDs are more than ready to master the workflows in a multi-cloud environment that are typical for the entertainment industry.

Note of thanks to our partners

This evaluation result is a great example of outstanding cooperation among trusting partners in the IT eco-system. The partners did not only provide their high quality hardware and software, but also contributed with productive round table discussions and valuable practical advice to support the progress and success of the activity. The resulting demo system was well recognized and appreciated by the audience at the IBC show in Amsterdam. I would like to thank especially Janusz Bak and Paweł Brzeżek from Open-E, Chloe Tseng and Greyson Chen from AIC, Michelle Seifert and Matt Mercurio from ATTO, Elke Behrendt and Mansoor Tariq from Promise, Serkan Albayrak from Nanya and the entire Toshiba team for their excellent support.

Rainer Kaese, Senior Manager Business Development, Storage Products Division, Toshiba Electronics Europe GmbH

Toshiba Electronics Europe GmbH

Hansaallee 181
40549 Düsseldorf
Germany

info@toshiba-storage.com
toshiba-storage.com