TOSHIBA

Enterprise Hard Drives

Designed for your business

1 Petabyte of Online Storage – 500 Watts

Whitepaper

toshiba-storage.com

TOSHIBA

Report from Toshiba Electronics Europe GmbH Storage Lab

Online data is continually increasing in capacity, so it is vital to develop storage systems that can keep up with this growing flood of data. The key criteria are:

- Cost: due to the immense amount of data, the most important criterion is the cost per capacity ($/TB).
- Physical dimensions: space in data centres is also a significant cost factor. Using the highest capacity hard disk drives in compact 19-inch form factor racks can minimise the space required.
- Power dissipation: as the name implies, online storage always needs to be on. Therefore, power consumption contributes directly to the total cost of operation. In addition, every watt consumed in the storage system has to be compensated for by the data centre's cooling system, which again leads to additional electricity costs.
- Performance: a certain performance is expected from online storage, as no one wants to wait a long time for their data. In the case of backup applications the time window available for backups is limited, so a defined amount of bandwidth must be available in order that the data can be written in the given time. When the worst case happens, and a backup needs to be restored, the backup data must be retrieved as quickly as possible so that businesses can quickly return to normality.

In this study the focus was on the optimisation of cost and power dissipation while minimising the mechanical dimensions of the system. Optimisation of the system's performance was not a goal, but it was measured in order to provide reference values. Should high performance be a primary goal, other solutions, such as SSDs, could be used, but their cost per capacity are many times more than those of HDD-based approaches.

Storage architecture – choice of HDD

Hard disk drives (HDDs) offer by far the lowest cost per capacity unit for online storage, so of course the choice of storage system is the hard disk drive. With regard to $/TB, the current top models are all similar with the $/TB ratios for 12TB, 14TB or 16TB disks lying in a similar range. Therefore, there is no preference when optimising for $/TB. However, when using 16TB HDDs, fewer disks are necessary for a given capacity than with 12TB or 14TB. This has an impact on another optimisation criterion: because fewer disks take up less space, at higher capacities the power dissipation per capacity will also be significantly lower, as shown in Table 1.

Year	Model	Capacity (TB)	Active Max Power (W)	Active W/TB
2013	MG04ACA	6	11.3	1.9
2015	MG05ACA	8	11.4	1.4
2017	MG06ACA	10	10.6	1.1
2018	MG07ACA	14	7.8	0.6
2019	MG08ACA	16	7.7	0.5

Table 1: Power dissipation and capacity of enterprise capacity HDDs
(Source: Toshiba data sheets & product manuals, each for random read/write QD=1 and 64kB blocks, single drive)

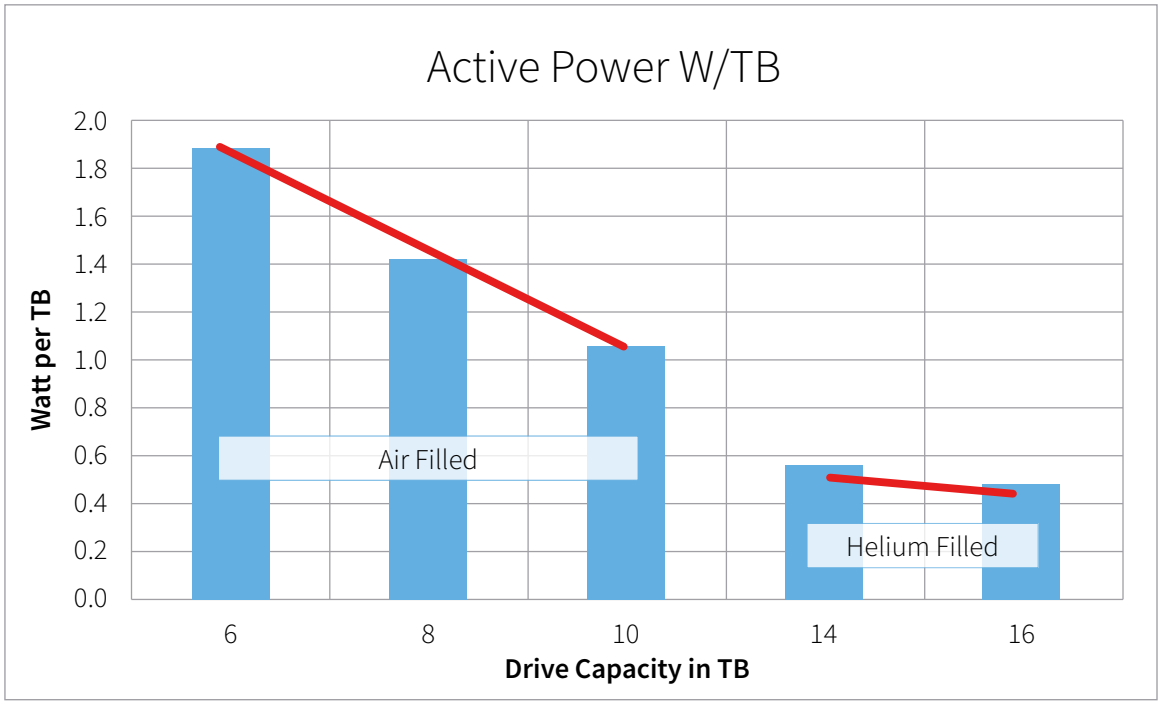

Figure 1: Power dissipation per TB for different HDD generations

The criteria for total power dissipation and space requirements therefore speak for the use of HDDs with the highest currently available capacity, in this case 16TB.

The corresponding 16TB HDDs of Toshiba's MG08 series are available with SAS or SATA interfaces. The SAS interface has two 12GB/s channels, making it suitable for architectures where speed and, above all, high availability are important. This comes at the expense of power dissipation (SAS HDDs consume roughly 1-2W more power than SATA HDDs, due to the higher power consumption of the interface). Since one goal was to optimise for power dissipation, the MG08ACA16TE model with SATA interface was chosen.

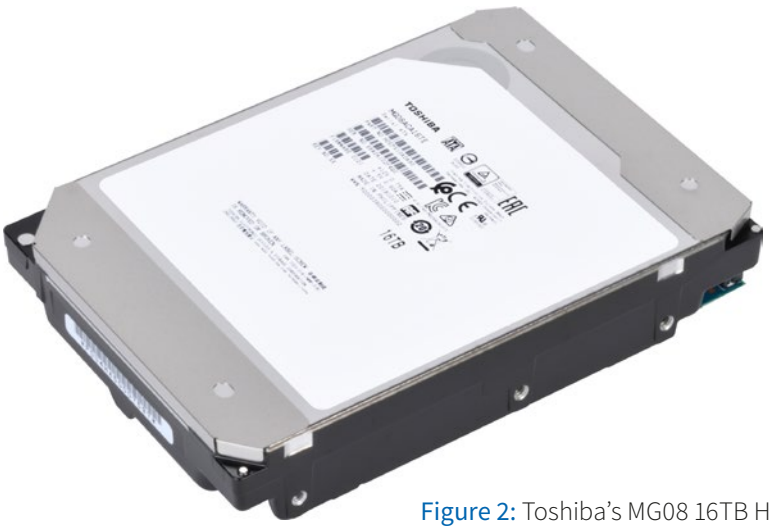

Figure 2: Toshiba's MG08 16TB HDD

The data sheet lists the following power dissipation values for this individual HDD:

Random Read 4kByte blocks, QD=16:	8.60W
Random Write 4kByte blocks, QD=16:	5.83W
Sequential Read:	7.50W
Sequential Write:	6.83W
Idle_A:	4.00W
Spin up Maximum in 500ms:	16.85W

Storage architecture – choice of HDD enclosure

The 45-100 bay top-loader models with 4 height-units (HE) offer the best space utilisation for 3.5" enterprise capacity ("Nearline") HDDs. These are available as a server (with server board) or as a JBOD with single or dual SAS expanders.

For this project a common 60-bay model from AIC was selected that fits into any existing 1000 mm rack due to its compact design. It should be noted that models with more than 60 drives are sometimes very long, so they cannot be inserted into 1000 mm racks and therefore need deeper versions. This JBOD variant was chosen because it allows easy measurement of the HDD power dissipation and the signal wiring (backplane, expander). Finally, a model with single expander was selected. This saves cost, power dissipation, and matches the SATA HDDs chosen that feature only one data channel on the interface anyway. The selected AIC model is named AIC-J4060-02 (JBOD, 4 height units, 60-bay, version 02 = single expander).

Figure 3: AIC's J4060-02 JBOD

Such a 60-Bay JBOD, when completely filled with 16TB HDDs, has a gross storage capacity of 960TB, making almost a petabyte of storage. The JBOD is connected to the host bus adapter (HBA) or RAID controller of the server by one mini-SAS-HD cable.

Configurations

The power consumption of the completely filled 60-bay JBOD was measured on the 220V terminals of the redundant power supplies. All measurements were carried out at an ambient temperature of 24°C.

Firstly, the power dissipation of the powered JBOD was measured but without the HDDs installed:

JBOD up, without drives, SAS link up: 80W

The next step was to install a single drive in the JBOD and take measurements under different workload conditions. Sequential 64kB blocks were written (equivalent to the workload of archiving, video recording and backup) and along with sequential 64kB block reads (equivalent to the workload of backup recovery and media streaming). As a reference, the power consumption during random reading/writing of 4kB blocks was also measured, corresponding to the workload of agile “hot-data” storage in databases. Naturally, this is not the target application for the configuration with one or more HDDs, so these values are for reference only. For all test setups, the power dissipation as well as the resulting performance (IOPS for random, MB/s for sequential) was measured.

In addition to these borderline cases a test with an approximate real workload was carried out. A mix of different block sizes was read and written randomly (4kB: 20 %, 64kB: 50 %, 256kB: 20 %, 2MB: 10 %). In order to achieve the maximum possible performance, all synthetic loads were executed with a queue depth (QD) of 16. In addition to these tests, a standard copy process was started on a logical drive under Windows and the power dissipation measured.

Workload	Power	IOPS	Bandwidth
Sequential write, 64kB blocks	85.0W	n/a	270MB/s
Sequential read, 64kB blocks	86.0W	n/a	270MB/s
Random write, 4kB blocks	83.6W	350	n/a
Random read, 4kB blocks	84.0W	420	n/a
Mixed read/write workload	84.2W	200	70MB/s
Windows copy	85.0W	n/a	110MB/s

The values for the individual drive (difference to the 80W of the unpopulated JBOD) are consistently lower than the values in the data sheet. It is noticeable that, contrary to the data sheet of the individual drive, the values for sequential loads are higher than for random loads. This is due to the higher power consumption of the SAS expanders of the JBOD carrying higher data bandwidths in sequential operation.

With all slots of the JOBD filled with 16TB HDDs, the maximum power dissipation at start-up was logged along with the power consumption in idle mode without read/write activity on the drives.

JBOD with drives, spin-up, max in 500ms: 720W
JBOD idle: 420W

The maximum power consumption when starting the JBOD is below the calculated value (80W + 60x 16.85W = 1100W) because the HDDs are started with a time delay (staggered spin-up). The value for JBOD idle is higher than the calculated value (80W + 60x 4W = 320W) as the controller addresses the HDDs occasionally even in idle mode.

60 HDDs in JBOD mode, parallel loads

In the next step, all 60 HDDs in JBOD mode were directly addressed in parallel by the operating system with synthetic workloads. The workloads described above were carried out and power dissipation, as well as performance, were measured as a reference.

Workload	Power	IOPS	Bandwidth
Sequential write, 64kB blocks	445W	n/a	1900MB/s
Sequential read, 64kB blocks	500W	n/a	2100MB/s
Random write, 4kB blocks	445W	23000	n/a
Random read, 4kB blocks	470W	7600	n/a
Mixed read/write workload	475W	1800	550MB/s

Active power consumption remains consistently below 500W.

Local RAID configuration

In a further step, the 60 HDDs were combined into one virtual drive using a RAID controller, specifically as RAID10 with 5 sub-arrays. On the resulting 480TB net storage, two logical drives of 240TB each were formatted under Windows Server 2016.

Workload	Power	IOPS	Bandwidth
Sequential write, 64kB blocks	425W	n/a	3900MB/s
Sequential read, 64kB blocks	460W	n/a	6200MB/s
Random write, 4kB blocks	445W	9800	n/a
Random read, 4kB blocks	480W	12000	n/a
Mixed read/write workload	465W	2700	790MB/s
Windows copy	430W	n/a	320MB/s

Software-defined storage

Finally, the 60 HDDs were configured to form a storage pool in a software-defined storage environment, a ZFS (zettabyte file system), managed by the JovianDSS software from Open-E.

Figure 4: Open-E's JovianDSS software

The redundancy is implemented by mirroring the data, with a pool made up of 5 sub-arrays and equipped with an 800GB enterprise SSD as a read cache, with another 800GB SSD as write log buffer. The storage capacity of the pool is made available to the server via the iSCSI protocol where, in turn, logical drives of 240TB are installed. Tests as for the logical drive on a local RAID set were undertaken (random read and write, sequential write and read, mixed workload and copy). The performance of a logical drive provided by ZFS via iSCSI depends strongly on the network bandwidth and, above all, on the configuration with SSD read caches and SSD write logs. Therefore, the values for solely synthetic workloads are given here for reference only.

Workload	Power	IOPS	Bandwidth
Idle (with ZFS background tasks)	430W		
Sequential write, 64kB blocks	445W		(250MB/s)
Sequential read, 64kB blocks	440W		(550MB/s)
Random write, 4kB blocks	470W	(2700)	
Random read, 4kB blocks	455W	(7000)	
Mixed read/write workload	480W	1100	330MB/s
Windows copy	450W		230MB/s

Conclusion

A petabyte (1000TB) of online HDD storage can now be served with the latest 16TB enterprise capacity HDDs in a 4U-JBOD with less than 500W power consumption. This power consumption varies between 420W (standby, no read/write activity) and 480W (continuous read/write of different sized blocks).

In typical storage configurations, such as mirroring or RAID, net storage capacities of between 480TB (RAID10/striped mirror) to 800TB (RAID60/striped double parity) are available using 60 x 16TB. In the overall system, this results in a power consumption of roughly 1W per TB net capacity (mirroring) down to 0.5W per TB (parity RAIDs).

Future development

Toshiba Electronics Europe GmbH estimates the total capacity of enterprise capacity (Nearline) HDDs shipped in 2019 at around 500 exabytes (500,000 petabytes). If all these HDDs were operated as 16TB models in 60-bay JBODs, this would result in a continuous power consumption of 225MW (equivalent to an average coal-fired power plant). However, since the majority of HDDs delivered in 2019 had even lower capacities, it can be assumed that the power consumption was even higher. And, since it is expected that the amount of data will increase even more in future, the power consumption required to store this data will play an increasingly important role. Therefore, it is the responsibility of our industry, as well as being one of Toshiba's goals, to develop HDDs with ever higher capacities while also optimising their power dissipation.

Contact us for more information:

www.toshiba-storage.com/contact/

TOSHIBA

Toshiba Electronics Europe GmbH

Hansaallee 181
40549 Düsseldorf
Germany

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

06/2020