TOSHIBA

Whitepaper

What is the best RAID configuration for 4 Drives?

Introduction

RAID systems are popular: companies, small business and private individuals use them for their individual use cases to protect valuable data from storage media failures and for the additional benefits they offer. These may include cost-effectiveness, enhanced performance over the single Hard Disk Drive (HDD) and increased resiliency, depending on the RAID configuration you choose.

RAID stands for Redundant Array of Independent (or Inexpensive) Disks. This technology combines several smaller drives into one larger storage space, with some redundancy like parity or mirroring of data, so that if a drive is failing, the data can be restored from the remaining drives after the failing drive is replaced.

But what is the best RAID configuration for NAS, USB-RAID box, RAID controller, or software defined storage? Which specifications are helpful and how much influence has the individual use case on selection?

This whitepaper provides some answers, based on fact-driven evaluation data from the Toshiba HDD Laboratory. The Toshiba laboratory ensures a correct test environment, where a typical configuration for smaller storage systems and sub-systems had been set up: Four Enterprise HDDs of 4TB, Toshiba Model MG08ADA400E.

Configurations

Four drives can be configured as:

- **RAID5**

The incoming data is distributed and stored in stripes over three disks and a fourth stripe carries the parity information. In case of a drive failing, the data can be retrieved from the parity. This RAID5 configuration offers 75% storage efficiency as the 4 drives (each 4TB) provides 12TB of usable data space. Data can be read from 3 or 4 disks in parallel, so the read speed is expected to be fast. The writing of the data is also done in parallel, but the parity information has to be calculated and written too, so the write speed is expected to be slower. In case of a rebuilding, all parity has to be calculated, which is likely to be a resource-consuming process.

- **RAID10**

This configuration is considered to be a potentially good alternative to RAID5. Instead of saving parity information as redundancy, a RAID10 stripes the data into two disks by simply mirroring the data of each stripe. This avoids the potentially resource-consuming parity calculation of the RAID5 at writing and rebuilding. Data can still be read from four drives, but is written only to two drives at the same time. However, this solution comes with a price: it offers only 50% storage efficiency due to the mirror redundancy.

- **RAID1**

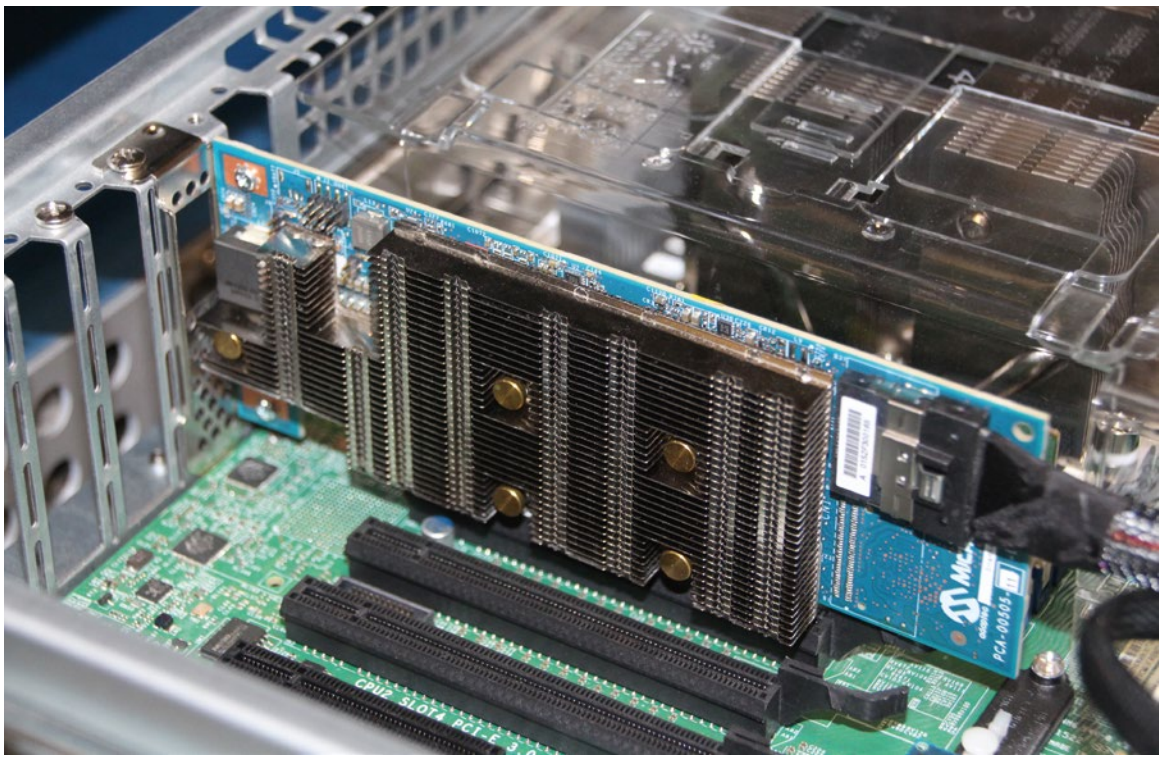
When looking at the configurations above, a question arises. Would it be an option to acquire only two drives at double the capacity - even if four drive slots are available - and operate it as a simple mirror (=RAID1)? The storage efficiency would be the same as RAID10, and data could still be read from two drives. But what are the performance de-merits compared to the other configurations? (It should be noted that especially for NAS systems in home offices and small businesses the performance requirements are rather relaxed). Toshiba evaluated this RAID1 option for the same systems with two units of Toshiba Enterprise HDD MG08ADA800E 8TB.

Picture 1: Toshiba HDDs 4TB/8TB

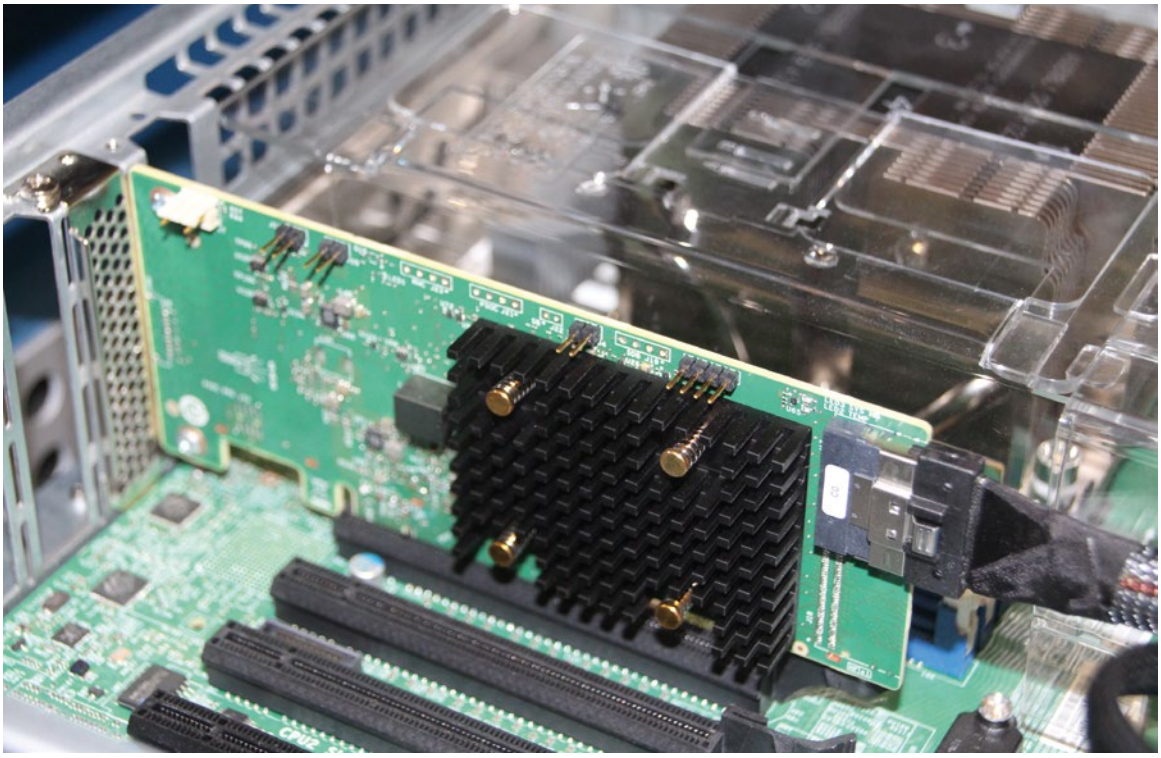
The Systems

RAID Controller

For the evaluation the Toshiba team decided to use two popular models commonly used for smaller configurations of up to 8 drives: The Broadcom “MegaRAID 9560-8i” and the “Adaptec® SmartRAID 3204-8i” from Microchip. They were installed in a PCI-Express Gen4 based Enterprise Server with HDDs connected via a backplane.



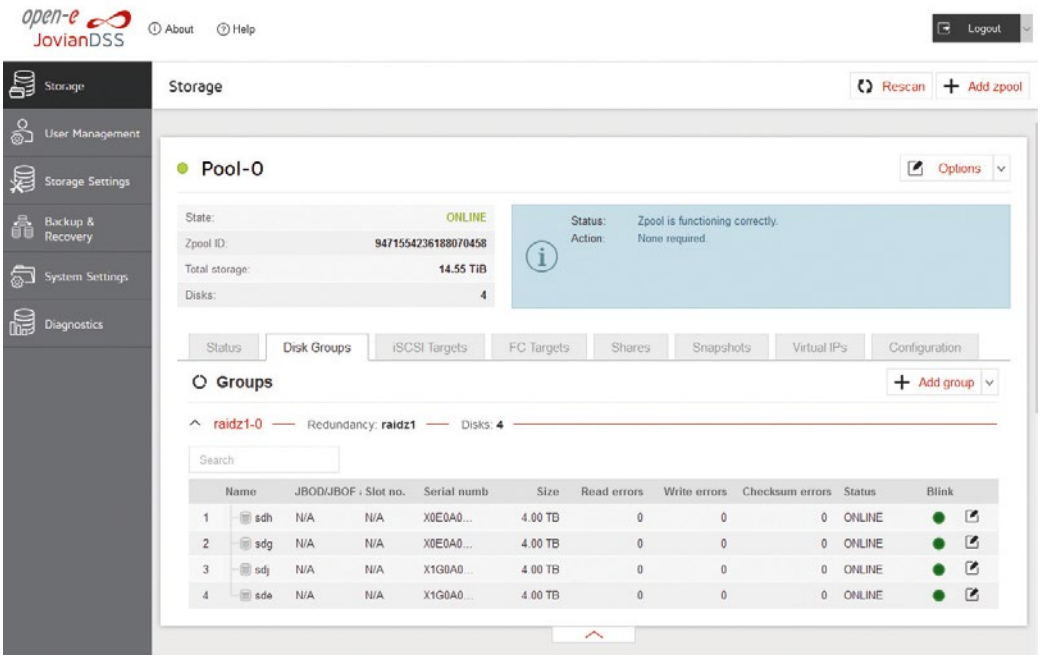
Picture 2: Microchip Adaptec® SmartRAID 3204-8i in server



Picture 3: Broadcom MegaRAID 9560-8i in server

Software Defined Storage

For two or four drives, hardware RAID is the most common and appropriate way to manage. But with high-performance and cost-efficient compute power (CPU, DRAM etc.) RAIDs can also be managed entirely by software, with the advantage of offering additional storage features such as snapshots, backups and more. Toshiba tested all configurations in a Zettabyte File System (ZFS) managed by Open-E JovianDSS software.



Picture 4: Open-E Jovian DSS GUI

Direct Attached Storage (DAS)

These are typically USB- or Thunderbolt-connected RAID boxes, which are directly attached to host systems. We tested an “RD3640SU3” from “ICY Box”.

Network Attached Storage (NAS)

NAS systems consist of a RAID storage subsystem connected to the network and they offer block- and shared storage within this network. We tested the “TS-464” from QNAP, expanded with a 10 GbE network card to avoid performance bottlenecks in the network connectivity.

See Table 1 for the full test matrix showing all models and test configurations:

Manufacturer	Model	Disk Model	Disk Capacity	Number	Config	Capacity
Adaptec® (Microchip)	3204	MG08ADA400E	4TB	4	RAID5	12TB
					RAID10	8TB
		MG08ADA800E	8TB	2	RAID1	8TB
Broadcom	9560	MG08ADA400E	4TB	4	RAID5	12TB
					RAID10	8TB
		MG08ADA800E	8TB	2	RAID1	8TB
Open-E JovianDSS	ZFS	MG08ADA400E	4TB	4	raid-z1	12TB
					2gr-mirror	8TB
		MG08ADA800E	8TB	2	s-mirror	8TB
QNAP	TS-464 (+10G card)	MG08ADA400E	4TB	4	RAID5	12TB
					RAID10	8TB
		MG08ADA800E	8TB	2	RAID1	8TB
ICY Box (Raidsonic)	RD3640SU3	MG08ADA400E	4TB	4	RAID5	12TB
					RAID10	8TB
		MG08ADA800E	8TB	2	RAID1	8TB

Table 1: Systems and Configurations Matrix

The Methodology

For each of the systems we setup a RAID5 of 4 HDDs MG08ADA400E and waited for full RAID initialization. We then stored 6TB of data in this RAID system and measured performance for:

- sequential writing of 1MB blocks
- sequential reading of 1MB blocks and a random reading
- Mixed read- and write workload of a mix of different block sizes.

```
fio --filename=test --size=6T --direct=1 --rw=write --bs=1024k --iodepth=64
--time_based --runtime=5m --group_reporting --name=job1
--ioengine=windowsaio --thread --numjobs=1 --group_reporting
--output=log_seqwrite.txt --norandommap --randrepeat=0

fio --filename=test --size=6T --direct=1 --rw=read --bs=1024k --iodepth=64
--time_based --runtime=5m --group_reporting --name=job1
--ioengine=windowsaio --thread --numjobs=1 --group_reporting
--output=log_seqread.txt --norandommap --randrepeat=0

fio --filename=test --size=6T --direct=1 --rw=randrw
--bssplit=4k/20:64k/50:256k/20:2M/10 --iodepth=64 --time_based --runtime=5m
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=8
--group_reporting --output=log_randmixed.txt --norandommap --randrepeat=0
```

Table 2: Measurement scripts

After the evaluation of all throughput values (MB/s), we simulated a failing disk by a hot removal, and measured the performance at the degraded array. We reinserted a new drive, measured the performance under rebuilding conditions and continued the rebuild process without any further load to find out how long it needs to rebuild the array.

Afterwards we repeated the same procedure for a RAID10 configuration of 4 drives, as well as for the RAID1 of two 8TB drives (MG08ADA800E). The measurement matrix for one RAID system is shown in Table 3.

Disk Model	Disk Capacity	Number	Config	Capacity	Rebuild Time	Performance Normal			Performance Degraded			Performance Rebuilding		
						SW	SR	M	SW	SR	M	SW	SR	M
						MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
MG08ADA400E	4TB	4	RAID5	12TB										
			RAID10	8TB										
MG08ADA800E	8TB	2	RAID1	8TB										

Table 3: Measurement Matrix (SW = SeqWrite, SR = SeqRead, M = Mixed)

To discuss and understand the results for a multi-drive configuration, the performance of a single drive was evaluated with the reference scripts as well (see Table 4).

Disk Model	Firmware	Capacity	Single Disk Performance		
			SeqWrite	SeqRead	Mixed
			MB/s	MB/s	MB/s
MG08ADA400E	0102	4TB	245	245	57
MG08ADA800E	0102	8TB	255	255	62

Table 4: Performance of single drive

Results for Hardware RAID Controllers

Manufacturer	Model	Disk Model	Disk Capacity	Number	Config	Capacity	Rebuild Time	Performance Normal		
								SW	SR	M
								MB/s	MB/s	MB/s
Adaptec® (Microchip)	3204	MG08ADA400E	4TB	4	RAID5	12TB	4h 0min	736	862	85
					RAID10	8TB	5h 20min	504	952	119
		MG08ADA800E	8TB	2	RAID1	8TB	11h 20min	255	491	76
Broadcom	9560	MG08ADA400E	4TB	4	RAID5	12TB	6h 0min	755	748	80
					RAID10	8TB	5h 50min	505	683	117
		MG08ADA800E	8TB	2	RAID1	8TB	11h 30min	260	375	73

Table 5: Nominal performance results for Hardware RAID Controllers (SW = SeqWrite, SR = SeqRead, M = Mixed)

From a sequential performance point of view – which is important for archiving and streaming of data – RAID5 consistently writes faster, as the writing always happens onto three disks in parallel. Reading is equally performant, for some controller models even slightly faster than writing.

RAID10 is slower in writing, as the data is written only to two disks due to the mirror protection technology but reading is rather fast as data can be retrieved from more than two drives.

For RAID1 using larger drives, the writing performance is limited to the speed of one drive, for reading to the speed of two drives.

So, here's the overview:

- Writing RAID5: 3x Reading RAID5: (up to) 4x
- Writing RAID10: 2x Reading RAID10: 3x
- Writing RAID1: 1x Reading RAID1: 2x

We can conclude that, for workloads with a dominance of sequential access, RAID5 is the best choice. It's not only faster, but also provides a higher storage efficiency (75% vs 50%).

In the case of a random or mixed workload, RAID10 gives about 1.5x the performance of the equivalent system with RAID5 configuration, while RAID1 is on the same level.

Rebuild times are in the range of 1 to 1.5 hours per TB of (failed) HDD capacity. That results in a rebuild time of 4~6 hours for the 4TB model and up to 12 hours for the 8TB HDD.

Performance with failed Disk (“Degraded”) and during Rebuilding Phase

Manufacturer	Disk Capacity	Number	Config	Capacity	Performance Normal			Performance Degraded			Performance Rebuilding		
					SeqWrite	SeqRead	Mixed	SeqWrite	SeqRead	Mixed	SeqWrite	SeqRead	Mixed
					MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
Adaptec® (Microchip)	4TB	4	RAID5	12TB	736	862	85	738	321	54	570	189	38
			RAID10	8TB	504	952	119	501	509	82	311	147	40
	8TB	2	RAID1	8TB	255	491	76	263	258	56	191	76	30
Broadcom	4TB	4	RAID5	12TB	755	748	80	754	351	72	725	362	48
			RAID10	8TB	505	683	117	505	511	82	487	479	80
	8TB	2	RAID1	8TB	260	375	73	263	260	50	252	310	49

Table 6: Hardware RAID controller performance in degraded and rebuilding phase | Remark: Controller Settings: Stripe Size 256kB, Write Back Caching, Drive Cache enabled, Controller Default Rebuild Priority (Broadcom: 30%, Adaptec®: “high”)

Interestingly, with one disk failed, the sequential write performance does not change, while the read performance is reduced approximately by the amount the failed drive had added to the nominal performance. The random/mixed workload performance is reduced accordingly (see Table 6).

When rebuilding the array after a failed drive has been replaced, the productive workload performance heavily depends on the rebuild priority. In the case of Adaptec®, the default rebuild priority setting is rather high, so the performance remaining for productive workload is reduced by about 30%. Broadcom handles the rebuilding with lower priority in the controller’s default setting, hence the productive performance drops by a few percent only. Of course, this leads to longer rebuild times if the workload continues to be at a high level. It should be noted that the rebuild priority setting (tradeoff between performance under rebuild and duration of the rebuilding process) can be adjusted for both controller models internal settings.

Managing the RAID by Software (ZFS)

Manufacturer	Model	Disk Model	Disk Capacity	Number	Config	Capacity	Rebuild Time	Performance Normal			Performance Degraded			Performance Rebuilding		
								SW	SR	M	SW	SR	M	SW	SR	M
								MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
Adaptec® (Microchip)	3204	MG08ADA400E	4TB	4	RAID5	12TB	4h 0min	736	862	85	738	321	54	570	189	38
					RAID10	8TB	5h 20min	504	952	119	501	509	82	311	147	40
		MG08ADA800E	8TB	2	RAID1	8TB	11h 20min	255	491	76	263	258	56	191	76	30
Broadcom	9560	MG08ADA400E	4TB	4	RAID5	12TB	6h 0min	755	748	80	754	351	72	725	362	48
					RAID10	8TB	5h 50min	505	683	117	505	511	82	487	479	80
		MG08ADA800E	8TB	2	RAID1	8TB	11h 30min	260	375	73	263	260	50	252	310	49
Open-E JovianDSS	ZFS	MG08ADA400E	4TB	4	raid-z1	12TB	3h 30min	562	845	22	330	530	18	399	296	16
					2gr-mirror	8TB	4h 40min	610	670	65	491	462	38	437	561	26
		MG08ADA800E	8TB	2	s-mirror	8TB	9h 30min	358	503	28	290	290	17	301	504	12

Table 7: Results for a ZFS based System (SW = SeqWrite, SR = SeqRead, M = Mixed)

In terms of ZFS, a configuration called raid-z1 (Raid with single redundancy) is equivalent to RAID5, a 2-group mirror would be the same as RAID10, and a single mirror configuration equivalent to RAID1. The Toshiba team tested all three configurations for completeness, but managing a single mirror of two disks with a ZFS would probably be too much of a good thing.

As Table 7 reveals, for ZFS and just 4 drives a parity approach (raid-z1) and a striped mirror configuration (2-group mirror) show the same sequential performance: this somehow falls in between the performance level of RAID5 and RAID10 with hardware RAID controllers. The performance drop at degraded systems is rather low and the performance under rebuild compares to a hardware RAID controller with a high rebuild priority setting.

Performance under random/mixed workload is just 1/4th to 1/3rd of Hardware RAID Controllers, but this will increase if (small size) SSDs are added to the system as write- and read cache – which is usually be done for configurations targeting random dominated workloads anyway.

RAID rebuild times for ZFS are lower than for hardware RAID controllers as software defined storage is aware of the amount of actual data on a disk and it will only re-create this data on a replacement disk. Hardware RAID controllers typically rebuild the entire disk, even if just partially filled with user data.

Network Attached Storage (NAS) with RAID configurations

Manufacturer	Model	Disk Model	Disk Capacity	Number	Config	Capacity	Rebuild Time	Performance Normal			Performance Degraded			Performance Rebuilding		
								SW	SR	M	SW	SR	M	SW	SR	M
								MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
Adaptec® (Microchip)	3204	MG08ADA400E	4TB	4	RAID5	12TB	4h 0min	736	862	85	738	321	54	570	189	38
					RAID10	8TB	5h 20min	504	952	119	501	509	82	311	147	40
		MG08ADA800E	8TB	2	RAID1	8TB	11h 20min	255	491	76	263	258	56	191	76	30
Broadcom	9560	MG08ADA400E	4TB	4	RAID5	12TB	6h 0min	755	748	80	754	351	72	725	362	48
					RAID10	8TB	5h 50min	505	683	117	505	511	82	487	479	80
		MG08ADA800E	8TB	2	RAID1	8TB	11h 30min	260	375	73	263	260	50	252	310	49
Open-E JovianDSS	ZFS	MG08ADA400E	4TB	4	raid-z1	12TB	3h 30min	562	845	22	330	530	18	399	296	16
					2gr-mirror	8TB	4h 40min	610	670	65	491	462	38	437	561	26
		MG08ADA800E	8TB	2	s-mirror	8TB	9h 30min	358	503	28	290	290	17	301	504	12
QNAP	TS-464 (+10G card)	MG08ADA400E	4TB	4	RAID5	12TB	6h 15min	571	697	34	572	412	27	422	324	21
					RAID10	8TB	6h 30min	491	535	42	492	476	31	446	394	25
		MG08ADA800E	8TB	2	RAID1	8TB	13h 50min	245	331	26	245	213	19	241	207	16

Table 8: Measurement results for QNAP NAS (SW = SeqWrite, SR = SeqRead, M = Mixed)

The sequential performance values for NAS with RAID configurations (see Table 8) are insignificantly lower than for the hardware RAID controllers, while the random performance is similar to ZFS without caching. Comparing RAID5 and RAID10, RAID5 is about 20% faster in terms of sequential speed and RAID10 is 20% faster with a view to random/mixed workloads.

It needs to be stressed that the sequential performance values of 200 MB/s and larger require a 10GbE interface at minimum. Still most (home-) NAS systems come with a 1GbE interface, which limits the sequential speed to about 100 MB/s, which is lower than a single HDD. Some are equipped with 2.5GbE, this increases

the maximum speed to 250MB/s which is enough for a 2-Bay RAID1 configuration. Only with 10GbE (this was implemented by a separate Add-In Card for the TS-464 model) can we achieve speeds beyond that level. So, any tradeoff / optimization in terms of sequential speed is only relevant for 10GbE or higher speed networks. For 1GbE and 2.5GbE network infrastructures, two HDDs in RAID1 are enough.

Just for workloads with many random read/write accesses, 4 drives of RAID10 may bring a speed advantage of about 1.5x.

Direct Attached Storage (USB/Thunderbolt RAID boxes)

Manufacturer	Model	Disk Model	Disk Capacity	Number	Config	Capacity	Rebuild Time	Performance Normal			Performance Degraded			Performance Rebuilding		
								SW	SR	M	SW	SR	M	SW	SR	M
								MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
Adaptec® (Microchip)	3204	MG08ADA400E	4TB	4	RAID5	12TB	4h 0min	736	862	85	738	321	54	570	189	38
					RAID10	8TB	5h 20min	504	952	119	501	509	82	311	147	40
		MG08ADA800E	8TB	2	RAID1	8TB	11h 20min	255	491	76	263	258	56	191	76	30
Broadcom	9560	MG08ADA400E	4TB	4	RAID5	12TB	6h 0min	755	748	80	754	351	72	725	362	48
					RAID10	8TB	5h 50min	505	683	117	505	511	82	487	479	80
		MG08ADA800E	8TB	2	RAID1	8TB	11h 30min	260	375	73	263	260	50	252	310	49
Open-E JovianDSS	ZFS	MG08ADA400E	4TB	4	raid-z1	12TB	3h 30min	562	845	22	330	530	18	399	296	16
					2gr-mirror	8TB	4h 40min	610	670	65	491	462	38	437	561	26
		MG08ADA800E	8TB	2	s-mirror	8TB	9h 30min	358	503	28	290	290	17	301	504	12
QNAP	TS-464 (+10G card)	MG08ADA400E	4TB	4	RAID5	12TB	6h 15min	571	697	34	572	412	27	422	324	21
					RAID10	8TB	6h 30min	491	535	42	492	476	31	446	394	25
		MG08ADA800E	8TB	2	RAID1	8TB	13h 50min	245	331	26	245	213	19	241	207	16
ICY Box (Raidsonic)	RD3640SU3	MG08ADA400E	4TB	4	RAID5	12TB	6h 30min	230	240	12	233	234	12	210	158	11
					RAID10	8TB	6h 40min	235	215	23	233	233	23	227	235	23
		MG08ADA800E	8TB	2	RAID1	8TB	11h 50min	221	233	25	217	227	25	206	213	24

Table 9: Measurement results for DAS USB RAID box (SW = SeqWrite, SR = SeqRead, M = Mixed)

Sequential values for 4 drive configurations are limited by the USB host connectivity's maximum transfer speed, which is true for most of the USB attachments. Only high-end Thunderbolt 10Gbps can deliver the full performance potential of 4 HDDs (see Table 9). Consequently a 4 drive configuration does not improve performance. Two HDDs in RAID1 / mirror deliver a maximum of sequential, but also a relatively good random/mixed performance.

4 or more HDD configurations only make sense if more capacity is required than a single/mirrored HDD can deliver. But as HDDs are available up to 20TB and more, two drives of higher capacity make more sense than four or more smaller HDDs.

Summary

Whether to use all four drives in RAID5 (single parity) or RAID10 (striped and mirrored) configuration, or deploy just two drives of double the capacity and run in RAID1 (simple mirror) mainly depends on capacity and

speed requirements. So make sure to check the necessary specifications for your own RAID system before making the final choice. Ask yourself questions like:

- How much capacity do I need?
- What about my typical workload?
  - Highly active applications such as database, virtualization etc. are dominated by random read/write workloads
  - Archiving, streaming and video/surveillance recording are almost exclusive sequential
  - Shared drives/folders in a network create a mixture of random and sequential, but unless users are actively working on the shared resources, the sequential part dominates.
- Are there any limitations imposed by the host/network connectivity?

Different configurations have diverse characteristics in terms of storage efficiency, RAID rebuild times and performance in degraded (disk-failure) condition and whilst being rebuilt, e.g when a defective drive has been replaced. All these aspects need to be considered.

Based on the data of our evaluation of RAID configurations as outlined above, we would offer the following recommendation for the three RAID configurations we tested:

RAID5 of four drives is best for all kinds of storage system solutions where the majority of the workload has a sequential nature. It also provides the best storage efficiency (75%). It is therefore suitable for high net capacity requirements and delivers the shortest possible rebuild times for failing drive replacements.

Raid10 of four drives: this is best for random/mixed workloads, of course trading off at a lower storage efficiency of only 50%. RAID10 is recommended for local server storage sub-systems for generic workloads using hardware RAID controllers or software RAID technology.

RAID1 of two drives of double the capacity: the configuration of choice for reasonable economic NAS systems in homes and small businesses with network connectivity at 1GbE or 2.5GbE, as well as for direct attached RAID systems with USB connectivity.

Further Considerations

Of course there are more options and more degrees of freedom than discussed here: for a higher number of drives (6 up to some dozen), RAID6 (raid-z2) is an option. With RAID6, double parity information is saved; so in case of a disk failure there is still a parity protection in place. With a view to performance, RAID6 is similar to RAID5 with one disk less. For software defined storage like ZFS, even a triple parity is possible (raid-z3).

Sub-arrays of RAID5 and RAID6 can be striped to connect more drives and achieve higher performance (similar to a striping of RAID1 into a RAID10). This would be called RAID50 and RAID60.

Indeed, the configuration options are endless, but our lab team is very motivated. If you have a storage planning challenge ahead, please contact us. We can evaluate and figure out the most appropriate configuration with Toshiba hard disk drives.

One more note: For a system of 60 hard disk drives, the configuration of 6 groups of 10 disks in raid-z2 (or RAID60 with 6 sub-arrays of 10 disks each in RAID6) is optimal for sequential reading and writing performance – that was the result which emerged from a recent set up in our lab . A detailed whitepaper about this will follow. Stay tuned.

TOSHIBA

Toshiba Electronics Europe GmbH

Hansaallee 181
40549 Düsseldorf
Germany

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

Issued 11/2023