TOSHIBA

Lab Report

Supermicro JBOD
“947SE2C-R1K66JBOD”

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

Introduction:

Top-loaded HDD enclosures with 60+ HDDs are becoming more and more popular, as they serve the need storing more and more data on smaller rack- and datacenter footprint. Drives are inserted from the top into an array of several rows of HDDs conveniently. Still due to this construction, all such enclosures face the same maintenance challenge: Installing drives as well as hot-swapping is tricky, since the heavy enclosure needs to be pulled out of the rack with all cables and the entire unit has to be pulled out to open the lid.

Supermicro introduced a smart solution to solve this problem: With the second generation of the “947”-chassis you can open the complete array like a drawer without the need to handle any cables or having to detach any lids. This looks promising, as it allows an easy and smooth installation and maintenance.

Toshiba Electronics Europe GmbH (“Toshiba”) had the chance to evaluate a sample of the 947SE2C-R1K66JBOD model (“JBOD947”) in the European HDD lab in Düsseldorf, Germany and to test the new drawer feature under live conditions.

Picture 1: JBOD947 in the Toshiba lab

TOSHIBA

Dimensions and mechanical features

With 4U height and a chassis length of just 813mm, this JBOD947 is rather short, and will fit into existing 1000mm server racks.
60 SAS/SATA 3.5” HDDs are installed in four rows of 15 each. HDDs are carried in plastic tool-less hot swap trays.
Connectivity is thru two 12Gbps SAS IO-Modules with 6x SFF-8644 ports each. The IO-Modules are managed by Supermicros IPMI accessed by a dedicated LAN port on each IO Module.

Setup in the Toshiba lab

Model:	947SE2C-R1K66JBOD
Firmware:	00.12.00.34
Host OS:	Windows (Windows Server 2019 Standard)
Host bus adapter (HBA):	Broadcom Avago HBA 9500-16e (Host IF: 8x PCIe-Gen4)
RAID controller:	Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)

Tests with enterprise capacity (“nearline”) SAS drives

Model name:	Toshiba MG10SFA22TE
Block size	512 byte emulated
Firmware	0101
Data-rate outer diameter:	284 MB/s

Power consumption

Idle_B:	3.55 W
Sequential write:	8.79 W
Sequential read:	8.30 W
Random write:	6.74 W
Random read:	9.92 W

Basic JBOD functions

Basic function:	OK
SAS IOM detected:	OK
Hot-plug/re-insert:	OK
Enclosure management:	OK, tested via BMC/IPMI at IOM’s LAN ports

Picture 2: Toshiba HDD MG10SFA22TE in tool-less JBOD tray

For precise measurement of the power consumption, we used a high-accuracy professional power analyser (R&S HMC8015).

JBOS off, BMC active ..... 15 W
JBOD on, no drives, SAS link to host on: ..... 140 W
JBOD with drives, maximum
startup power over 500 ms: ..... 1300 W
JBOD with raw drives at HBA: ..... 580 W
Lambda (ratio of active and reactive power): 0.975
Noise at 1 m distance: ..... 54 dB
Ambient Temperature : ..... 26 degC

140W as idle power with no drives is an excellent power value for the JBOD itself. The lambda factor of 0.975 means that the ratio of reactive power generated by the power supply units is extremely low (equating to just 2.5% of the total power). The higher (i.e. better) the lambda factor, the smaller the reactive power. Reactive power does not cost and does not create heat, but power supply rails have to be dimensioned for both active and reactive power – so for large data centers a high lambda is an important factor.

In case of parallel startup (ie. after power down) of many JBODs at the same time, the 1300W as maximum power for a cold startup should be considered for the dimensioning of the power supplies.

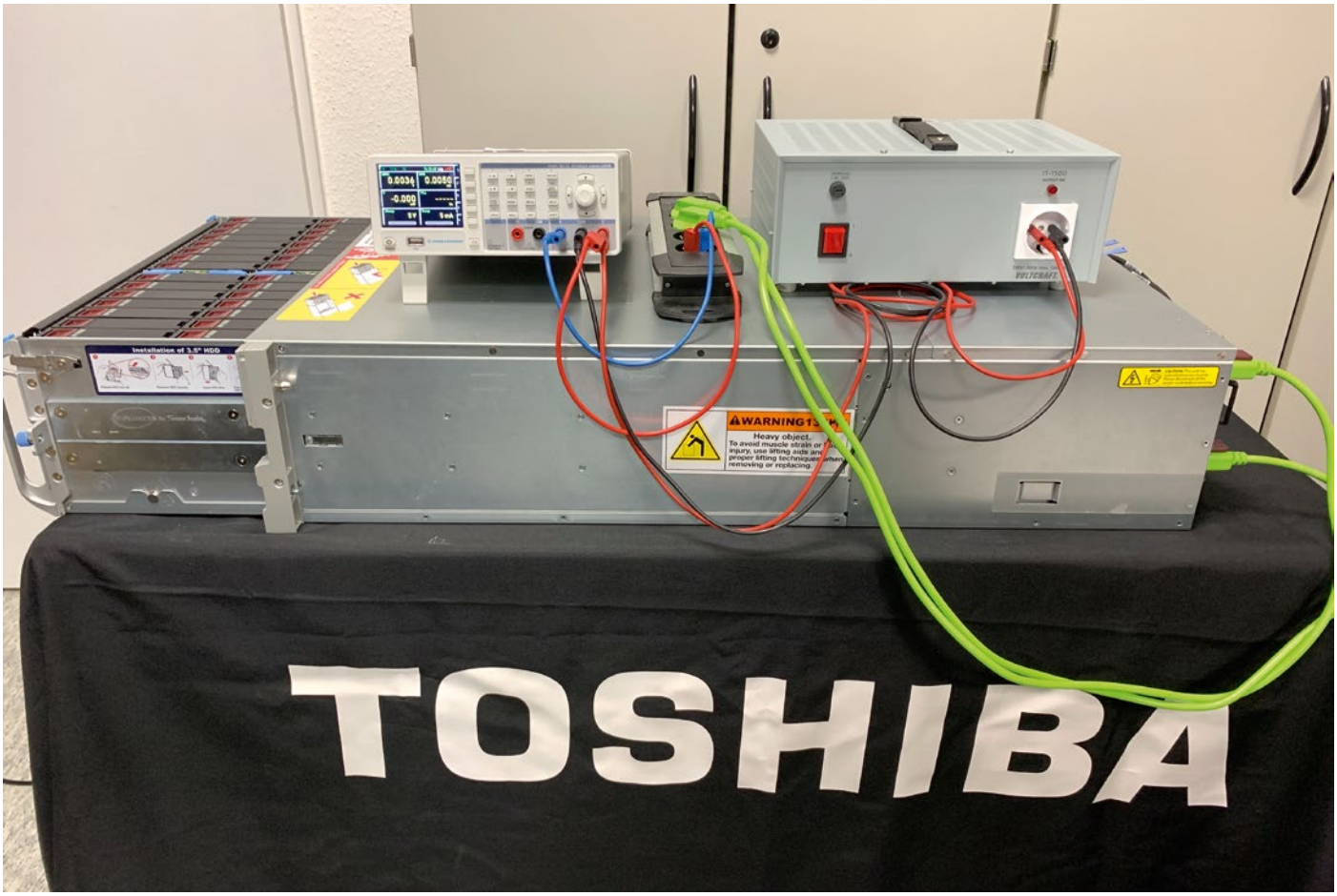
While the JBOD947 starts up at full fan speed with a noise level of >90dB, the fan-noise in idle and active mode is very low. 54dB appears to the most quiet 4U top-loader JBOD measured in Toshiba's lab so far.

Performance measurements in the Toshiba lab

For tests with SAS drives, we have connected both IOM of the JBOD with two mini-SAS-HD cables each to the 4 mini-SAS-HD ports of the 16e HBA and RAID controller.

This configuration gives a theoretical JBOD/HDD access-bandwidth of 4x 4.8GB/s = 19.2GB/s, but this would require multipath setting and wide-path aggregation.

As for configuration with HBA, multipath has to be enabled manually in Linux/Windows. A RAID controller (such as the



Picture 3: Power measurement setup in the Toshiba HDD lab

Adaptec® Ultra-3254 Model) will detect the configurations automatically and invoke a correct multipath setting. Manual multipath and SAS-link aggregation only works for SAS drives. Configuration with SATA drives will still benefit from the aggregation of IOPS of the 60 HDDs, but the sequential bandwidth is usually limited to one mini-SAS-HD link (4.8GB/s).

We`ve tested several drive configurations with "fio" flexible IO tester software - measuring sequential, random and mixed workload performance and related power consumption.

Tests have been done for individual drives connected via HBA and in RAID configurations as physical drives, as well as for logic drives. For logic drives we also measured the performance and power of a copy (i.e. reading and writing) of a large file.

The configurations were:

1. 60 drives configured as RAID0, meaning all HDDs combined to one large array without redundancy, exercised as "physical drive" in Windows Server. Measurements were done accessing the full 1320TB of capacity. This configuration is meant to benchmark the raw performance of the hardware. Due to missing redundancy it is not relevant as productive solution.

2. 60 drives configures as RAID10, meaning 30 two-way mirror sub-arrays of RAID1 combined to on large array. This is a practically used configuration for storage solutions optimized for mixed workload of sequential- but also random and read/write mixed type of access. We created a windows logic volume of 660TB and measured based on a (more realistic) file size of 1TB.

3. 60 drives passed individually to the Linux/Centos operating system via a host bus adaptor, with multipath setting.

1. All drives as RAID0, Windows physical drive:

OS: Windows Server 2019
HBA/controller: Adaptec® SmartRAID Ultra 3254-16e/e (16xPCIe4)
HDD: 60x Toshiba MG10SFA22TE
Configuration: Dual IOM 4x mini-SAS HD cables (3 m long)

Workload	Power (W)	IOPS	Bandwidth (MB/s)
Sequential write 4M	700		13200
Sequential read 4M	690		14500
Random write 4k	580	38700	
Random read 4k	780	9700	
Mixed 4k/64k/256k/2M	750	6100	1800
Idle (raid background)	580		
Temperature ambient	26°C		
Temperature HDD min.	30°C		
Temperature HDD max.	46°C		

Sequential bandwidth of > 13 GB/s is close to the physical limit of the system. Also the IOPS scale with the number of drives. Write IOPS are much higher, because of no need for time consuming seeking of the data in case of random reading. This is also the reason for the high power consumption in at random reading.

Script (all drives as RAID0, Windows physical drive):

```
fio --filename=\\.\Physicaldrive1 --direct=1 --rw=write --bs=4m --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=1
--norandommap --randrepeat=0 --output=seqwritephysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=read –bs=4m --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=1
--norandommap --randrepeat=0 --output=seqreadphysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randwrite --bs=4k --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randwritephysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randread --bs=4k --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randreadphysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randrw --bssplit=4k/20:64k/50:256k/20:2M/10
--iodepth=256 --time_based --runtime=300 --group_reporting --name=job1 --ioengine=windowsaio
--thread --numjobs=64 --norandommap --randrepeat=0 --output=mixedphysical.log
```

2. All drives as RAID10, Windows logic volume:

OS: Windows Server 2019
HBA/controller: Adaptec® SmartRAID Ultra 3254-16e/e (16xPCIe4)
HDD: 60x Toshiba MG10SFA22TE
Configuration: Dual IOM 4x mini-SAS HD cables (3 m long)

Workload	Power (W)	IOPS	Bandwidth (MB/s)
Sequential write 4M	700		7000
Sequential read 4M	680		12900
Random write 4k	580	6800	
Random read 4k	600	27200	
Mixed 4k/64k/256k/2M	620	4500	1300
Windows copy (Rd/Wr)	600		450/450
Idle (raid background)			
Temperature ambient	26°C		
Temperature HDD min.	29°C		
Temperature HDD max.	45°C		

Write bandwidth is lower, as due to the mirrored configuration of the RAID10, data is written only to 30 HDDs in parallel, while a reading process still accesses all 60 disks. The high number of read IOPS is due to the fact that the operation is not exercising the entire 800TB of disk space, but only a test file of 1TB (which is a more realistic scenario).

3. All drives parallel as single physical devices (multipath):

OS: Linux (Centos 7.9)
HBA/controller: Broadcom HBA9500-16e
HDD: 60x Toshiba MG10SFA22TE
Configuration: Dual IOM 4x mini-SAS HD cables (3 m long)
Multipath setup for disks

Workload	Power (W)	IOPS	Bandwidth (MB/s)
Sequential write 4M	670		8700
Sequential read 4M	670		6700
Random write 4k	600	36400	
Random read 4k	600	14800	
Mixed 4k/64k/256k/2M	620	9200	2700
Temperature ambient	26°C		
Temperature HDD min.	30°C		
Temperature HDD max.	46°C		

The sequential performance is limited to the dual-path configuration of two links. Random and mixed performance is excellent as all drives contribute individually. This bare metal performance is a good base for a software defined storage system.

Coolest temperature only 4 degC above air intake indicates a good cooling fan design and speed. Hottest temperature 16 degC higher for the HDDs in the rear rows of the JBOD is okay, but as HDDs long term reliability starts to suffer from about 42 degC onwards, we recommend to make sure that the intake air temperature is kept below 20 degC.

Script (all drives as RAID10, Windows logic volume):

```
fio --filename=test --size=1T --direct=1 --rw=write --bs=4m --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=1
--norandommap --randrepeat=0 --output=seqwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=read –bs=4m --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=1
--norandommap --randrepeat=0 --output=seqreadlogical.log

fio --filename=test --size=1T --direct=1 --rw=randwrite --bs=4k --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=randread --bs=4k --iodepth=256 --time_based
--runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=randreadlogical.log

fio --filename=test --size=1T --direct=1 --rw=randrw --bssplit=4k/20:64k/50:256k/20:2M/10
--iodepth=256 --time_based --runtime=300 --group_reporting --name=job1 --ioengine=windowsaio
--thread --numjobs=64 --norandommap --randrepeat=0 --output=mixedlogical.log
```

Script (all drives parallel as single physical devices):

```
fio --direct=1 --bs=4m —-iodepth=256 —time_based —runtime=1h --ioengine=libaio
--group_reporting --rw=write --output=seqwrite.log -- name=/dev/dm-{2..61}

fio --direct=1 --bs=4m —-iodepth=256 —time_based —runtime=1h --ioengine=libaio
--group_reporting --rw=read --output=seqread.log -- name=/dev/dm-{2..61}

fio --direct=1 --bs=4k —-iodepth=256 —time_based —runtime=1h --ioengine=libaio
--group_reporting --rw=randwrite --output=randwrite.log --name=/dev/dm-{2..61}

fio --direct=1 --bs=4k —-iodepth2566 —time_based —runtime=1h --ioengine=libaio
--group_reporting --rw=randread --output=randread.log --name=/dev/dm-{2..61}

fio --direct=1 --bssplit=4k/20:64k/50:256k/20:2M/10 —-iodepth=256 —time_based —runtime=1h
--ioengine=libaio --group_reporting --rw=randrw --output=mixed.log --name=/dev/dm-{2..61}
```

Summary

Supermicros JBOD 947SE2C-R1K66 is an excellent building block for Petabyte scale high capacity high performance datacenter online storage. With Toshiba's 22TB SAS Enterprise capacity drives it delivers 1320 TB of raw storage accessible with a bandwidth of more than 10 GByte/s and up to 10.000 (read-) IOPS.

Even at an access speed this high, the power consumption stays in the 600~800W range, and the HDDs are kept reasonably cool, supporting long lifetime and low failure rates.

A gamechanger is the amazingly smooth physical access to the drives under maintenance – the chassis carrying the drive can be opened like a drawer, no lids have to be removed and all cable stay in their position. Even hot-swapping becomes less risky as the peril of damaging valuable data is significantly reduced.

Note of thanks to our partners

This lab report is the result of dedicated and passionate collaboration. "I would like to thank all our partners for the support on this project. Supermicro provided the JBOD 947SE2C-R1K66 to us, Microchip supported with the raid controller Adaptec® SmartRAID Ultra 3254-16e /e and finally Broadcom contributed the Host-Bus-Adapter HBA 9500-16e. Together with our Toshiba Hard Disk Drives, I was able to test Supermicro`s JBOD under realistic data center settings in our laboratory and show its impressive results."

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

Issued 12/2023