TOSHIBA

1 Petabyte de almacenamiento en línea – 500 vatios

toshiba-storage.com

TOSHIBA

Informe del Lab de Almacenamiento de Toshiba Electronics Europe GmbH

Actualmente es posible proporcionar un petabyte (1000TB) de almacenamiento online en HDD con los últimos discos duros de capacidad empresarial del 16 TB en un JBOD de 4U con un consumo de energía inferior a 500V. Este consumo varía entre 420V (en reposo, sin actividad de lectura/escritura) y 480V (lectura/escritura continua de bloques de diferentes tamaños).

En configuraciones de almacenamiento típicas, como el mirroring o RAID, están disponibles capacidades de almacenamiento neto de entre 480 TB (RAID10/mirroring dividido) y 800 TB (RAID60 / paridad dual dividida) utilizando 60 unidades de 16 TB. En el sistema general, el resultado es un consumo de energía de aproximadamente 1V por TB de capacidad neta (mirroring) y hasta 0,5V por TB (RAIDs de paridad).

Esta es la principal conclusión de las pruebas realizadas por Toshiba Electronics Europe GmbH en su Lab de Almacenamiento y las cuales se han recogido en un informe que pone de manifiesto la responsabilidad que tiene la industria de sistemas de almacenamiento en el desarrollo de discos duros con capacidades cada vez mayores y la optimización simultanea de su ratio de disipación de energía. Este es, de hecho, uno de los objetivos prioritarios de Toshiba.

No en vano, Toshiba estima que la capacidad total de los discos duros de capacidad empresarial (nearline) suministrados en 2019 asciende a alrededor de 500 exabytes (500.000 petabytes). Si todos estos discos duros operaran con modelos de 16TB en JBODs de 60 bahías, el resultado en términos de consumo continuo sería de 225MV, el equivalente a una planta promedio de energía de carbón. Sin embargo, dado que la mayoría de los HDD suministrados en 2019 tenían capacidades mucho más bajas, se deduce que el consumo de energía es mayor. Y, puesto que se prevé que el volumen de datos siga aumentando, el consumo de energía requerido para su almacenamiento tendrá una importancia aún mayor en el futuro.

El análisis de Toshiba parte de una premisa incontestable. El volumen de datos online aumenta de forma continua, por lo que es vital desarrollar sistemas de almacenamiento capaces de mantenerse al día ante esta avalancha. Los criterios clave para ello son:

- Coste: debido al inmenso volumen de datos, el criterio más importante es el coste por capacidad (€/TB).

- Dimensiones físicas: el espacio en los centros de datos también es un factor significativo de coste. El uso de unidades de disco duro de mayor capacidad en bastidores compactos de 19 pulgadas puede minimizar las necesidades de espacio.

- Disipación de energía: como su nombre indica, el almacenamiento online necesita estar siempre activo. Por lo tanto, el consumo de energía contribuye directamente al coste total de operación. Además, cada vatio consumido en el sistema de almacenamiento debe ser compensado por el sistema de enfriamiento del centro de datos, lo que genera nuevamente costes adicionales de electricidad.

- Rendimiento: del almacenamiento online se espera un determinado rendimiento ya que nadie quiere esperar mucho tiempo para acceder a sus datos. En el caso de las aplicaciones de backup, la ventana de tiempo disponible para las copias de respaldo es limitada, por lo que debe estar disponible una cantidad de ancho de banda definida para que los datos puedan escribirse en el tiempo establecido. Cuando sucede lo peor y es necesario restaurar una copia de seguridad, los datos de la copia deben recuperarse lo más rápido posible para que las empresas puedan retornar rápidamente a la normalidad.

En este estudio hemos puesto la atención en la optimización de los costes y la disipación de energía, así como en la minimización de las dimensiones mecánicas del sistema. Si bien la optimización del rendimiento del sistema no era un objetivo, también se midió para proporcionar valores de referencia. En caso de que un alto rendimiento fuera un objetivo fundamental, podrían utilizarse otras soluciones, tales como los SSDs, pero su coste por capacidad es varias veces mayor que el de los enfoques basados en discos duros mecánicos.

Arquitectura de almacenamiento – elección de HDD

Los discos duros mecánicos o HDDs (Hard Disk drives) ofrecen, con mucho, el coste más bajo por unidad de capacidad para almacenamiento online, por lo que la elección del sistema de almacenamiento es, por supuesto, la unidad de disco duro. Respecto a la relación €/TB, los modelos superiores actuales son similares con una ratio €/TB en la misma franja para discos de 12 TB, 14 TB o 16 TB. Por lo tanto, no existe una preferencia cuando se trata de optimizar la relación €/TB. Sin embargo, cuando se utilizan discos duros de 16 TB son necesarios menos discos para contar con una capacidad dada que si se utilizan discos de 12 TB o 14 TB. Esto tiene un impacto en otro criterio de optimización: dado que menos discos ocupan menos espacio, a mayores capacidades, la disipación de energía por capacidad también será significativamente menor, como se muestra en la Tabla 1.

Año	Modelo	Capacidad(TB)	Pwr (V) Activo Máximo	V/TB Activo
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

Tabla 1: Disipación de energía y capacidad de HDD de capacidad empresarial
(Fuente: Hojas de datos & guías de producto de Toshiba, cada una para lectura / escritura aleatoria QD=1 y bloques de 64kB, unidad única)

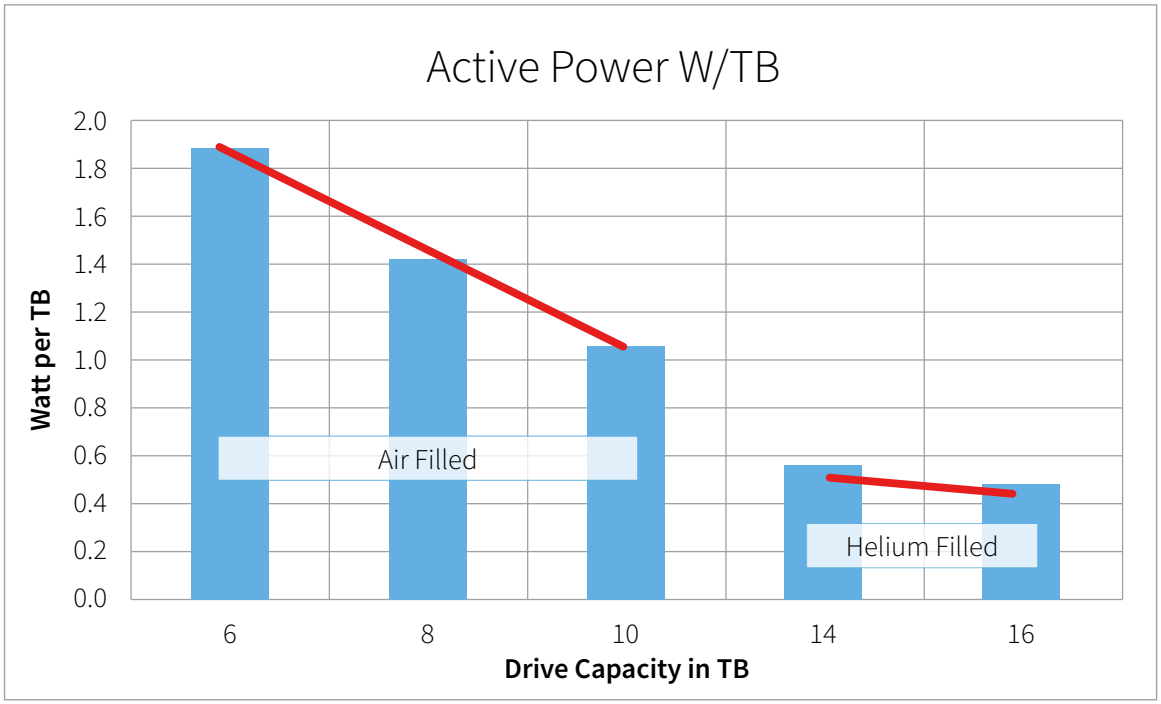

Figura 1: Disipación de energía por TB en diferentes generaciones de HDD

Los criterios para la disipación de energía total y los requerimientos de espacio favorecen, por tanto, al uso de HDD con la capacidad más alta disponible actualmente, en este caso 16 TB.

Los discos duros de 16 TB de la serie MG08 de Toshiba están disponibles con interfaces SAS o SATA. La interfaz SAS tiene dos canales de 12GB/s, por lo que es adecuada para arquitecturas en las que es importante la velocidad y, sobre todo, la alta disponibilidad. Esto se consigue a expensas de la disipación de energía ya que los discos duros SAS consumen aproximadamente 1-2 V más de energía que los discos duros SATA, debido al mayor consumo de energía de la interfaz. Dado que uno de los objetivos era optimizar la disipación de energía, se eligió el modelo MG08ACA16TE con interfaz SATA.

Figura 2: Disco duro Toshiba MG08 16TB

La hoja de datos enumera los siguientes valores de disipación de energía para este HDD individual:

Lectura aleatoria bloques 4kByte, QD=16:	8.60V
Escritura aleatoria bloques 4kByte, QD=16:	5.83V
Lectura secuencial:	7.50V
Escritura secuencial:	6.83V
Inactivo_A:	4.00V
Giro máximo en 500ms:	16.85V

Arquitectura de almacenamiento – elección de la caja para HDD

En cuanto a la arquitectura, los modelos superiores de cargador de 45 -100 bahías con cuatro unidades de altura (HE) ofrecen el mejor uso de espacio para unidades de disco duro de 3,5" de capacidad empresarial ("nearline"). Están disponibles como servidor (con una placa base) o como JBOD con expansores SAS simples o duales.

Para el proyecto se seleccionó un modelo común de 60 bahías de AIC, que se adapta a cualquier rack existente de 1000 mm debido a su diseño compacto. Se optó un modelo con expansor único. Esto ahorra costes y disipación de energía, y encaja con los discos duros SATA elegidos que, en todo caso, cuentan con un solo canal de datos en la interfaz. El modelo de AIC seleccionado es el AIC-J4060-02 (JBOD, cuatro unidades de altura, 60 bahías, versión 02 con expansor único).

AIC

Figura 3: JBOC J4060-02 de AIC

Un JBOD de 60 bahías como este ofrece, cuando está completamente lleno de discos duros de 16 TB, una capacidad bruta de almacenamiento de 960 TB, lo que supone casi un petabyte de almacenamiento. El JBOD se conecta al adaptador bus del host (HBA) o al controlar RAID del servidor mediante un cable mini-SAS-HD.

Configuraciones

El consumo de energía del JBOD de 60 bahías completamente lleno se midió en los terminales de 220V de las fuentes de alimentación redundante, siempre a una temperatura ambiente de 24°C.

En primer lugar, se midió la disipación de energía del JBOD con alimentación, pero sin los discos duros instalados, con un resultado de 80V. El siguiente paso fue instalar una sola unidad en el JBOD y tomar mediciones en diferentes condiciones de carga de trabajo. Se escribieron bloques secuenciales de 64kb (equivalente a carga de trabajo de archivo, grabación de video y copia de seguridad), junto con lecturas secuenciales de bloques de 64kB (equivalente a carga de trabajo de recuperación de un backup y streaming). Como referencia, también se midió el consumo de energía durante la lectura /escritura aleatoria de bloques de 4KB, correspondiente a la carga de trabajo del almacenamiento ágil de “datos calientes” en las bases de datos. Para todas las configuraciones de prueba, se midió la disipación de energía, así como el rendimiento resultante (IOPS para aleatorio, MB/s para secuencial).

Adicionalmente a estos casos límite, se realizó una prueba con una carga de trabajo próxima a la realidad. Se leyó y escribió aleatoriamente una mezcla de diferentes tamaños de bloques (4 kB: 20%, 64 kB: 50%, 254 kB:20%, 2 MB: 10%). Además de estas pruebas, se inició un proceso de copia estándar en una unidad lógica bajo Windows y también se midió la disipación de energía.

Carga de trabajo	Energía	IOPS	Ancho de banda
Escritura secuencial, bloques 64kB	85.0V	n/a	270MB/s
Lectura secuencial, bloques 64kB	86.0V	n/a	270MB/s
Escritura aleatoria, bloques 4kB	83.6V	350	n/a
Lectura aleatoria, bloques 4kB	84.0V	420	n/a
Cargas mixtas lectura/escritura	84.2V	200	70MB/s
Copia Windows	85.0V	n/a	110MB/s

Los valores para la unidad individual (diferencia con los 80V del JBOD vacío) son consistentemente más bajos que los valores en la hoja de datos. Contrariamente a lo reflejado en la hoja de datos de la unidad individual, los valores para cargas secuenciales son más altos que para cargas aleatorias. Esto se debe al mayor consumo de energía de los expansores SAS del JBOD que transportan datos con un ancho de banda más elevado en la operación secuencial.

Con todas las ranuras del JOBD llenas de discos duros de 16 TB se registró la máxima disipación de energía (720V) en la puesta en marcha junto con el consumo de energía en modo inactivo sin actividad de lectura/escritura (420V).

60 HDDs en modo JBOD, cargas paralelas

En un siguiente paso, los 60 discos duros en JBOD fueron directamente llamados en paralelo por el sistema operativo con cargas de trabajo sintéticas. Se llevaron a cabo las cargas de trabajo descritas antes y se midieron la disipación de energía y el rendimiento.

Carga	Energía	IOPS	Ancho de banda
Escritura secuencial, bloques 64kB	445V	n/a	1900MB/s
Lectura secuencial, bloques 64kB	500V	n/a	2100MB/s
Escritura aleatoria, bloques 4kB	445V	23000	n/a
Lectura aleatoria, bloques 4kB	470V	7600	n/a
Cargas mixtas lectura/escritura	475V	1800	550MB/s

Configuración de RAID local

En un siguiente paso, los 60 discos duros se combinaron en una unidad virtual utilizando un controlador RAID, específicamente como RAID10 con 5 submatrices. En el almacenamiento neto resultante de 480 TB se formatearon dos unidades lógicas de 240 TB cada una con Windows Server 2016.

Carga	Energía	IOPS	Ancho de banda
Escritura secuencial, bloques 64kB	425V	n/a	3900MB/s
Lectura secuencial, bloques 64kB	460V	n/a	6200MB/s
Escritura aleatoria, bloques 4kB	445V	9800	n/a
Lectura aleatoria, bloques 4kB	480V	12000	n/a
Cargas mixtas lectura/escritura	465V	2700	790MB/s
Copia de Windows	430V	n/a	320MB/s

Almacenamiento definido por software

Finalmente, los 60 discos duros se configuraron como un pool de almacenamiento en un entorno definido por software, un ZFS (sistema de archivos zettabyte), administrado por el software JovianDSS de Open-E.

Figura 4: Software JovianDSS de Open-E

La redundancia se implementa haciendo un mirroring de los datos, con un pool compuesto por 5 subarrays y equipado con un SSD empresarial de 800 GB como caché de lectura y otro SSD de 800 GB como búfer de registro de escritura. La capacidad de almacenamiento del pool se pone a disposición del servidor mediante el protocolo iSCSI y donde se instalan a su vez unidades lógicas de 240 TB. Se realizaron pruebas de la unidad lógica en un conjunto RAID local (lectura y escritura aleatorias, escritura y lectura secuencial, cargas mixtas y copia). El rendimiento de la unidad lógica proporcionada por ZFS a través de iSCSI depende en gran medida del ancho de banda y, sobre todo, de la configuración con las cachés de lectura SSD y los registros de escritura SSD. Por lo tanto, los valores para cargas de trabajo sintéticas únicamente se proporcionan como referencia.

Carga	Energía	IOPS	Ancho de banda
Inactivo (con tareas ZFS en segundo plano)	430V		
Escritura secuencial, bloques 64kB	445V		(250MB/s)
Lectura secuencial, bloques 64kB	440V		(550MB/s)
Escritura aleatoria, bloques 4kB	470V	(2700)	
Lectura aleatoria, bloques 4kB	455V	(7000)	
Cargas mixtas lectura/escritura	480V	1100	330MB/s
Copia Windows	450V		230MB/s

TOSHIBA

Toshiba Electronics Europe GmbH

Hansaallee 181
40549 Düsseldorf
Germany

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

07/2020