



1 Pétaoctet de stockage en ligne - 500Watts

Un livre blanc de Toshiba Electronics Europe GmbH Storage Lab

Auteur : Rainer W. Kaese, Senior Manager Business Development Storage Products, Toshiba Electronics Europe

Le nombre d'informations en ligne augmente constamment. Il est donc essentiel de développer des systèmes de stockage capables de suivre ce flot croissant de données. Il existe plusieurs critères clés sont:

- **Le coût** : en raison de l'immense quantité de données, le critère le plus important est le coût par capacité (\$/To).
- **Les dimensions physiques** : l'espace disponible dans les centres de données est également un facteur de coût important. L'utilisation de disques durs de plus haute capacité dans des racks compacts de 19 pouces peut réduire l'espace requis.
- **La consommation d'énergie** : le stockage en ligne doit toujours être actif. Par conséquent, la consommation d'énergie contribue directement au coût total de fonctionnement. En outre, chaque watt consommé dans le système de stockage doit être compensé par le système de refroidissement du centre de données. Ce qui entraîne à nouveau des coûts d'électricité supplémentaires.
- **La performance** : une certaine performance est attendue en matière de stockage en ligne, car personne ne veut attendre trop longtemps ses données. Dans le cas des applications de sauvegarde, le temps est limité. Donc une quantité définie de bande passante doit être disponible afin que les données puissent être écrites dans le temps imparti. Dans le cas d'une sauvegarde qui doit être restaurée, les données doivent être récupérées le plus rapidement possible afin que les entreprises puissent revenir à la normalité au plus vite.

Dans cette étude, l'accent est mis sur l'optimisation des coûts et de la consommation d'énergie tout en minimisant les dimensions mécaniques du système. L'optimisation des performances du système n'était pas un objectif, mais elle a été mesurée afin de fournir des valeurs de référence. Si l'objectif principal était la performance maximale, d'autres solutions, telles que les SSD, pourraient être utilisées. Mais leur coût par capacité de stockage est beaucoup plus élevé que celui des disques durs.

TOSHIBA

Architecture de stockage – le choix du disque dur

Les disques durs (HDD) offrent de loin le coût le plus bas par unité de stockage en ligne. Donc bien sûr il faut les privilégier. En ce qui concerne le ratio \$/To, les modèles haut de gamme actuels sont tous similaires pour les disques de 12 To, 14 To ou 16 To. Par conséquent, il n'y a aucune préférence lors de l'optimisation pour \$/ To. Cependant, lorsque vous utilisez des disques durs de 16 To, moins de disques sont nécessaires pour une capacité donnée qu'avec 12To ou 14 To. Cela a un impact sur un autre critère d'optimisation : moins de disques prennent moins d'espace, à des capacités plus élevées, la consommation d'énergie par capacité de stockage sera également considérablement plus faible, comme le montre le tableau 1.

Year	Model	Capacity (TB)	Active Max Pwr (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

Tableau 1: consommation d'énergie et capacité des disques durs d'entreprise

(Source: fiches techniques et manuels de produits Toshiba, chacun pour lecture / écriture aléatoire QD = 1 et 64 Ko blocks, lecteur unique)

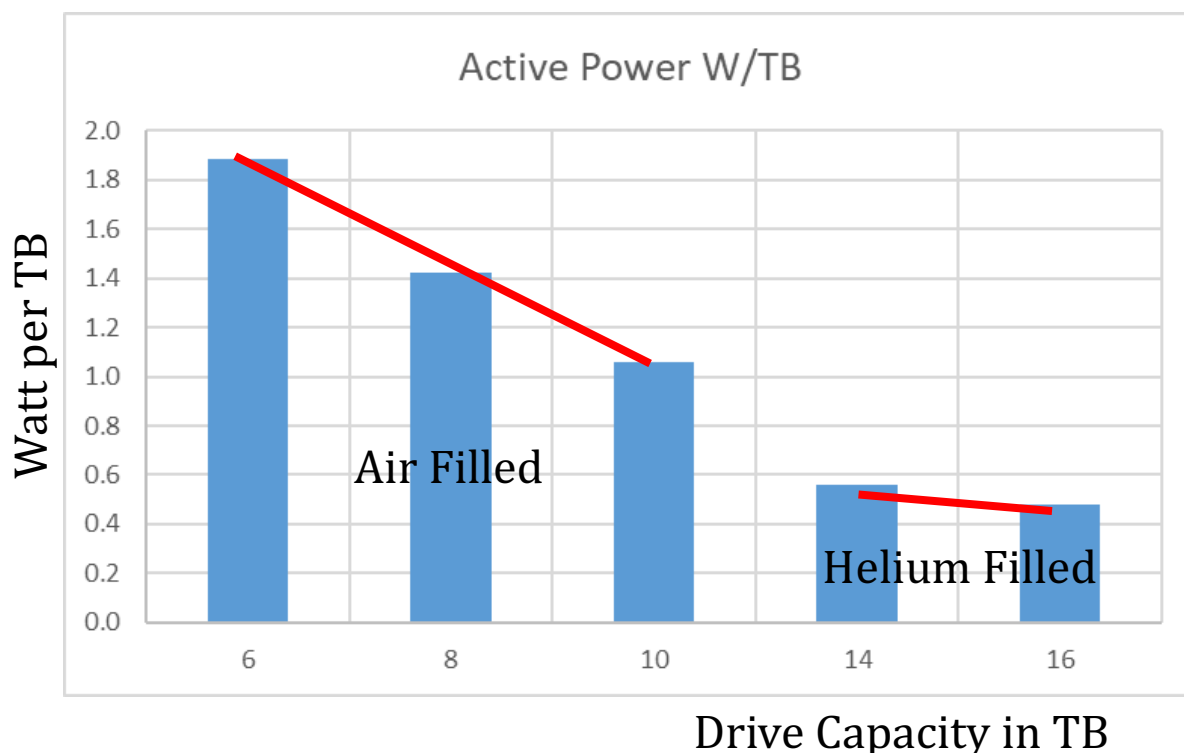




Figure 1: Consommation d'énergie par To pour différentes générations de disques durs

Les critères de consommation d'énergie totale et les besoins d'espace sont donc assez parlants en matière d'utilisation de disques durs de la plus grande capacité actuellement disponible, c'est à dire 16 To.

Les disques durs de 16 To correspondants à la série MG08 de Toshiba sont disponibles avec des interfaces SAS ou SATA. L'interface SAS possède deux canaux à 12 Go/s, ce qui la rend adaptée aux architectures au sein desquelles la vitesse et, surtout, la haute disponibilité sont importantes.

Cela se fait au détriment de la consommation d'énergie (les disques durs SAS consomment environ 1 à 2 W de plus que les disques durs SATA, en raison de la consommation d'énergie plus élevée de l'interface). Comme l'un des objectifs était d'optimiser la consommation d'énergie, le modèle MG08ACA16TE avec interface SATA a été choisi.



Figure 2: Disque dur MG08 16 To de Toshiba

TOSHIBA

La fiche technique répertorie les valeurs de consommation d'énergie suivantes pour ce disque dur :

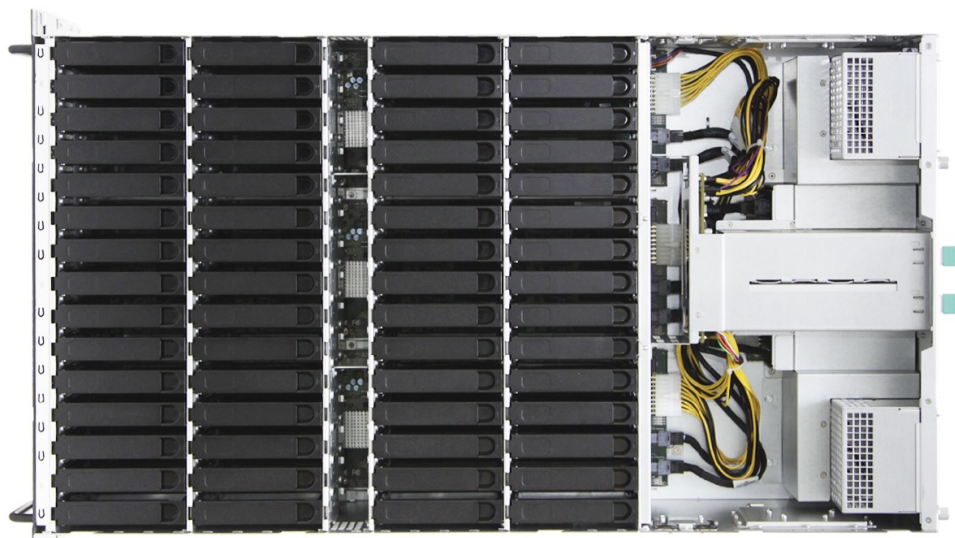
Blocs à lecture aléatoire de 4 Ko, QD = 16:	8,60 W
Blocs d'écriture aléatoire de 4 Ko, QD = 16:	5,83 W
Lecture séquentielle:	7,50 W
Écriture séquentielle:	6,83 W
Idle_A:	4.00W
Spin up Maximum en 500 ms:	16,85 W

Architecture de stockage – le choix du format de disque dur

Les modèles à charge de stockage supérieure de 45 à 100 baies avec 4 unités de hauteur (HE) offrent la meilleure utilisation d'espace pour les disques durs d'entreprise (" Nearline ") de 3,5". Ils sont disponibles en tant que serveur (avec server board) ou JBOD avec expandeurs SAS simples ou doubles.

Pour ce projet, un modèle commun de 60 baies AIC a été sélectionné et s'intègre dans n'importe quel rack de 1000 mm existant en raison de sa conception compacte. Il convient de noter que les modèles avec plus de 60 disques sont parfois très longs, ils ne peuvent donc pas être insérés dans des racks de 1000 mm et nécessitent des versions plus profondes. Cette variante JBOD a été choisie car elle permet de mesurer facilement la consommation d'énergie du disque dur et le câblage du signal (face arrière, expandeur). Enfin, un modèle avec un seul expandeur a été sélectionné. Cela réduit les coûts, la consommation d'énergie et correspond aux disques durs SATA choisis. Ces derniers ne comportent de toute façon qu'un seul canal de données sur l'interface. Le modèle AIC sélectionné est nommé AIC-J4060-02 (JBOD, 4 unités de hauteur, 60 baies, version 02 = simple expandeur).

TOSHIBA



AIC

Figure 3: JBOD J4060-02 d'AIC

Un tel JBOD à 60 baies, lorsqu'il est entièrement rempli de disques durs de 16 To, a une capacité de stockage brute de 960 To, ce qui correspond presque à un pétaoctet de stockage. Le JBOD est connecté à l'adaptateur de bus hôte (HBA) ou au contrôleur RAID du serveur par un câble mini-SAS-HD.

Configurations

La consommation électrique du JBOD 60 baies, complètement rempli, a été mesurée sur les bornes 220V. Toutes les mesures ont été effectuées à une température ambiante de 24 °C.

Tout d'abord, la consommation d'énergie du JBOD alimenté a été mesurée mais sans les disques durs installés:

JBOD up, sans disques, SAS link up: 80W

L'étape suivante consistait à installer un seul lecteur dans le JBOD et à prendre des mesures dans différentes conditions de charge de travail. Des blocs séquentiels de 64 Ko ont été écrits (équivalant à la charge de travail d'archivage, d'enregistrement vidéo

TOSHIBA

et de sauvegarde) ainsi que des lectures séquentielles de blocs de 64 Ko (équivalant à la charge de travail de récupération de sauvegarde et de streaming multimédia). À titre de référence, la consommation d'énergie lors de la lecture/écriture aléatoire de blocs de 4 Ko a également été mesurée, ce qui correspondant à la charge de travail du stockage agile de «données sensibles» dans les bases de données. Naturellement, ce n'est pas l'application cible pour la configuration avec un ou plusieurs disques durs, donc ces valeurs sont à titre indicatif uniquement. Pour toutes les configurations de test, la consommation d'énergie ainsi que les performances qui en résultent (IOPS pour aléatoire, MB / s pour séquentiel) ont été mesurées.

En plus de ces cas précis, un test avec une charge de travail réelle approximative a été effectué. Un mélange de différentes tailles de blocs a été lu et écrit de manière aléatoire (4 Ko: 20%, 64 Ko: 50%, 256 Ko: 20%, 2 Mo: 10%). Afin d'atteindre les performances maximales possibles, toutes les charges synthétiques ont été exécutées avec une profondeur de file d'attente (QD) de 16. En plus de ces tests, un processus de copie standard a été démarré sur un lecteur logique sous Windows et la consommation d'énergie mesurée.

Charge de travail	Energie	IOPS	Bande passante
=====			
=====			
Ecriture séquentielle, 64kB blocks	85.0W	n/a	270MB/s
Lecture séquentielle, 64kB blocks	86.0W	n/a	270MB/s
Ecriture aléatoire, 4kB blocks	83.6W	350	n/a
Lecture aléatoire, 4kB blocks	84.0W	420	n/a
Charge de travail mixte	84.2W	200	70MB/s
Lecture/écriture			
Copie Windows	85.0W	n/a	110MB/s

Les valeurs pour le variateur individuel (différence par rapport aux 80W du JBOD non occupé) sont systématiquement inférieures aux valeurs de la fiche technique. Il est à noter que, contrairement à la fiche technique du variateur individuel, les valeurs pour les charges séquentielles sont plus élevées que pour les charges aléatoires. Cela est dû à une consommation d'énergie plus élevée des extenseurs SAS du JBOD transportant des bandes passantes de données plus grandes en fonctionnement séquentiel.

TOSHIBA

Avec tous les emplacements du JBOD occupés par les disques durs de 16 To, la consommation maximale d'énergie au démarrage a été enregistrée ainsi que la consommation d'énergie en mode veille sans activité de lecture/écriture sur les disques.

JBOD avec disques, rotation, max en 500ms: 720W
JBOD inactif: 420W

La consommation électrique maximale lors du démarrage du JBOD est inférieure à la valeur calculée ($80W + 60 \times 16.85W = 1100W$) car les disques durs démarrent avec un délai (rotation décalée). La valeur pour JBOD inactif est supérieure à la valeur calculée ($80W + 60 \times 4W = 320W$) car le contrôleur s'adresse occasionnellement aux disques durs même en mode inactif.

60 disques durs en mode JBOD, charges parallèles

Dans l'étape suivante, les 60 disques durs en mode JBOD ont été directement adressés en parallèle par le système d'exploitation avec des charges de travail synthétiques. Les charges de travail décrites ci-dessus ont été réalisées et la consommation d'énergie, ainsi que les performances, ont été mesurées comme référence.

Charge de travail	Energie	IOPS	Bande passante
=====			
=====			
Ecriture séquentielle, 64kB blocks	445W	n/a	1900MB/s
Lecture séquentielle, 64kB blocks	500W	n/a	2100MB/s
Ecriture aléatoire, 4kB blocks	445W	23000	n/a
Lecture aléatoire, 4kB blocks	470W	7600	n/a
Charge de travail mixte	475W	1800	550MB/s
Lecture/écriture			

La consommation d'énergie active reste constamment inférieure à 500 W.

Configuration RAID locale

Dans une autre étape, les 60 disques durs ont été combinés en un seul lecteur virtuel à l'aide d'un contrôleur RAID, spécifiquement en tant que RAID10 avec 5 sous-baies. Sur le stockage net résultant de 480 To, deux lecteurs logiques de 240 To ont chacun été formatés sous Windows Server 2016.

TOSHIBA

Charge de travail	Energie	IOPS	Bande passante
=====			
=====			
Ecriture séquentielle, 64kB blocks	425W	n/a	3900MB/s
Lecture séquentielle, 64kB blocks	460W	n/a	6200MB/s
Ecriture aléatoire, 4kB blocks	445W	9800	n/a
Lecture aléatoire, 4kB blocks	480W	12000	n/a
Charge de travail mixte	465W	2700	790MB/s
Lecture/écriture			
Copie Windows	430W	n/a	320MB/s

Stockage défini par le logiciel

Enfin, les 60 disques durs ont été configurés pour former un pool de stockage dans un environnement de stockage défini par logiciel. Un ZFS (zettabyte file system), géré par le logiciel JovianDSS d'Open-E.



Figure 4: logiciel JovianDSS d'Open-E

La redondance est mise en œuvre en mettant en miroir les données, avec un pool composé de 5 sous-baies et équipé d'un SSD d'entreprise de 800 Go comme cache de lecture, avec un autre SSD de 800 Go comme tampon de journal d'écriture. La capacité de stockage du pool est mise à la disposition du serveur via le protocole iSCSI où, à leur tour, des disques logiques de 240 To sont installés. Des tests comme pour le lecteur logique sur un ensemble RAID local ont été utilisés (lecture et écriture aléatoires, écriture et lecture séquentielles, charge de travail et copie mixtes). Les

TOSHIBA

performances d'un lecteur logique fourni par ZFS via iSCSI dépendent fortement de la bande passante du réseau et, surtout, de la configuration avec des caches de lecture SSD et des journaux d'écriture SSD. Par conséquent, les valeurs des charges de travail uniquement synthétiques sont données ici à titre de référence uniquement.

Charge de travail	Energie	IOPS	Bande passante
=====	=====	=====	=====
Inactif (avec tâches en arrière-plan ZFS)	430W		
Ecriture séquentielle, 64kB blocks	445W		(250MB/s)
Lecture séquentielle, 64kB blocks	440W		(550MB/s)
Ecriture aléatoire, 4kB blocks	470W	(2700)	
Lecture aléatoire, 4kB blocks	455W	(7000)	
Charge de travail mixte	480W	1100	330MB/s
Lecture/écriture			
Copie Windows	450W		230MB/s

Conclusion

Un pétaoctet (1000 To) de stockage de disque dur en ligne peut désormais être obtenu avec les derniers disques durs d'entreprise de 16 To dans un 4U-JBOD avec une consommation électrique inférieure à 500W. Cette consommation d'énergie varie entre 420W (veille, pas d'activité de lecture / écriture) et 480W (lecture / écriture continue de blocs de tailles différentes).

Dans des configurations de stockage typiques, telles que la mise en miroir ou RAID, des capacités de stockage nettes comprises entre 480 To (RAID 10 / miroir rayé) et 800 To (RAID 60 / double parité rayée) sont disponibles en utilisant 60 x 16 To. Dans le système global, cela se traduit par une consommation d'énergie d'environ 1 W par To de capacité nette (mise en miroir) jusqu'à 0,5 W par To (RAID de parité).

Développement futur

Toshiba Electronics Europe GmbH estime la capacité totale des disques durs de d'entreprise (Nearline) expédiés en 2019 à environ 500 exaoctets (500000 pétaoctets).

TOSHIBA

Si tous ces disques durs étaient exploités en tant que modèles de 16 To dans des JBOD à 60 baies, cela entraînerait une consommation électrique continue de 225MW (équivalent à une centrale électrique au charbon moyenne). Cependant, comme la majorité des disques durs livrés en 2019 avaient des capacités encore plus faibles, on peut supposer que la consommation d'énergie était encore plus élevée. Et, comme il est prévu que la quantité de données augmentera encore plus à l'avenir, la consommation d'énergie requise pour stocker ces données jouera un rôle de plus en plus important. Par conséquent, il est de la responsabilité de notre industrie, de développer des disques durs avec des capacités toujours plus élevées tout en optimisant leur consommation d'énergie.