Université IBM i

19 et 20 novembre 2024

IBM Innovation Studio Paris

uii2024
#ibmi
#uii2024

S31 – Stockage interne sur NVMe : compréhension, performance et retour d'expérience client

20 novembre 11:15 - 12:15

Laurent Mermet - Laurent.Mermet@ibm.com
Jean-Luc Bonhommet – jeanluc_bonhommet@fr.ibm.com
Ludovic Menard - ludovic_menard@fr.ibm.com
IBM France

IBM

common FRANCE

Université IBM i

19 et 20 novembre 2024

IBM i
continuous innovation
continuous integration

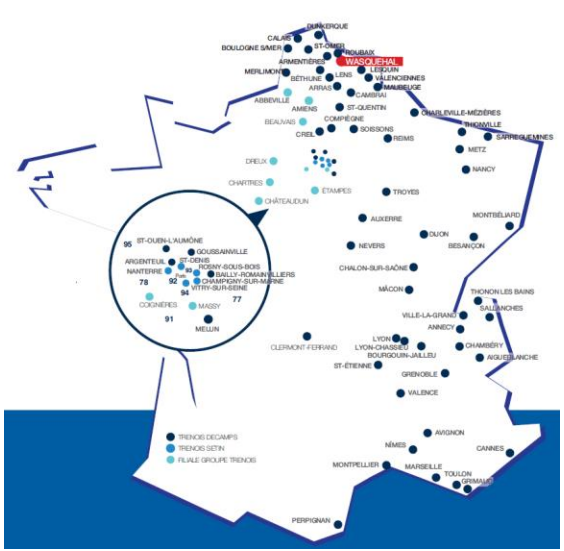
Une relation Client-Partenaire- Constructeur

Trenois Decamps

Le groupe Trenois, quincaillerie professionnelle :

- **fondé en 1878 à Lille (59). Entreprise familiale depuis cinq générations**
- **expert dans la distribution de fournitures techniques et équipements professionnels pour les secteurs du bâtiment et de l'industrie.**

64 agences	**800** grandes marques référencées	**+ de 600** commerciaux itinérants et sédentaires
800 collaborateurs	**247** millions de CA	**50 000** références
106 927€ donnés à des associations en 2022	**25 000 m²** de plateforme logistique	**7 000** colis envoyés par jour



BÂTIMENT	INDUSTRIE
BTP Génie Civil	Maintenance
Second œuvre du BTP	Construction métallique
Energie	Usinage
Bois	Manufacture
Collectivité et administration	Soudage

Infitex

Depuis 1988, Infitex accompagne les PME et ETI du secteur de la négoce/distribution en leur offrant des solutions digitales innovantes et adaptées à leurs besoins spécifiques.

Infitex met aujourd'hui à disposition de ses 50 clients son expertise à travers **ses 3 métiers** :

Développement des progiciels Olymp ERP et Olymp WMS sur IBMi
Des solutions conçues pour gérer les fonctions vitales de l'entreprise

Installation et maintenance d'infrastructures réseaux
Une équipe dédiée pour accompagner les clients dans la gestion et la maintenance de leur matériel informatique

Développement d'applications web et de site e-commerce
En y intégrant directement Olymp ERP via API pour faciliter la digitalisation des clients

Université IBM i

19 et 20 novembre 2024

IBM i
continuous innovation
continuous integration

NVMe

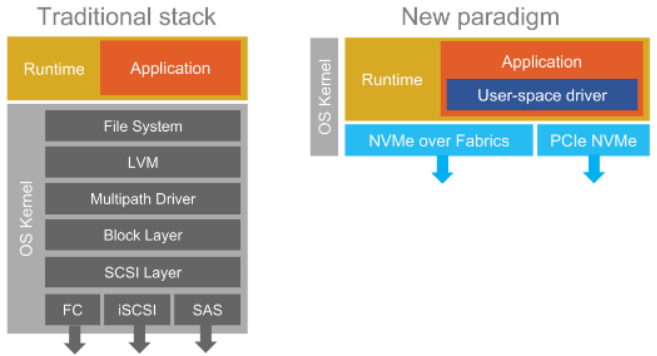
C’est quoi ?

Qu’est ce que NVMe? (Non-Volatile Memory Express)

- **NVMe est une interface ET un protocole de communication**
  - NVMe est conçu dès le départ pour offrir une bande passante élevée et un accès au stockage à faible latence
  - NVM Express définit une interface **efficace** permettant au logiciel hôte de communiquer avec un sous-système de mémoire non volatile via PCI Express (NVMe sur PCIe)
  - Fonctionnellement analogue à SAS et SATA, mais conçu pour réduire les traitements supplémentaires des pilotes, du système d'exploitation et des applications
  - Exploite les complétions d'E / S basées sur l'interrogation par opposition aux complétions basées sur les interruptions

- **NVMe utilise la structure PCIe**
  - Plusieurs unités aujourd'hui sur le marché
  - Plusieurs formats, y compris des disques 2,5 pouces

- **NVMe a été conçu pour des performances élevées**
  - Augmentation des E/S, bande passante et latence plus faible
  - Exploite Flash et les mémoires non volatiles de nouvelle génération
  - Tire parti des environnements multi-cœurs, parallélisme d'E / S élevé



Composant NVMe

NVMe Controller

PCIe Attached
Parallel Architecture (Multi Q, Q pairs ...)
Low Latency Design
Fabrics Attach Friendly
Self Encryption & Sanitize
Virtualization (Multiple Namespace, SR-IOV)
IO Determinism
Zoned Namespace
Management Interface Support (in & out of band)
Computational Storage
... More Innovations

Media

NAND TLC (most widely used)
Optane 3DXP
LL NAND
NAND QLC
DRAM – Flash backed
MRAM
.. More to come

Formats: M.2, U.2 (2.5” mince & épais), EDSFF (E1.S, E1.L, E3.S, E3.L mince & épais), Add in card

Infrastructure logicielle: investissement protégé par la réutilisation sur plusieurs appareils de fournisseurs

Caractéristiques NVMe et IBM i

- NVMe est capable de fournir des performances plus élevées que les SSD. La technologie NVMe peut fournir beaucoup plus d'E/S en lecture ou en écriture et un débit nettement plus élevé (Go / s) par rapport aux SSD SAS / SATA. Les différences de performances réelles du système ou des applications varient en fonction du client et de la charge de travail.
- NVMe fournit des capacités de virtualisation supplémentaires puisque chaque périphérique est un point de terminaison PCIe qui peut être dédié à une partition / LPAR
- Au moins une paire d'adaptateurs NVMe identiques est requise; les paires d'adaptateurs NVMe suivantes peuvent être différentes de la première paire. Après la commande d'une paire identique, un adaptateur NVMe de capacité différente est autorisé.
- Les périphériques NVMe nécessitent la mise en miroir du système d'exploitation IBM i car il n'y a pas de prise en charge du RAID matériel. Les paires en miroir doivent se trouver sur des unités physiques différentes. NVMe ne peut être mis en miroir que sur NVMe et les disques SAS ne peuvent être mis en miroir que sur des disques SAS.
- Le disque de secours à chaud n'est pas pris en charge, mais un NVMe supplémentaire peut être sur le système en tant que disque de secours à froid pour accélérer le processus de réparation, et ce n'est qu'un disque de rechange dans le fait qu'un client n'a pas à le commander / le brancher. Le développement IBM i est conscient du désir de quelque chose de plus performant qu'une pièce de rechange froide.

Unités NVMe supportées par IBM i en natif

FC	Description	CAPACITY TB	Read BW GB/s	Write BW GB/s	Read IOPS (K)	Write IOPS (K)	Latency Read (us)	Latency write (us)	DWPD (5 ans)	GA date
ES5B	Enterprise 800GB SSD PCIe4 NVMe U.2 for IBM i	800	6.4	1.3	800	140	80	15	2.4	Oct-23
ES5D	Enterprise 1.6 TB SSD PCIe4 NVMe U.2 module for IBM i	1.6	7.0	2.5	1300	260	80	15	3	Oct-23
ES5F	Enterprise 3.2 TB SSD PCIe4 NVMe U.2 module for IBM i	3.2	7.0	4.1	1600	280	80	15	3	Oct-23
ES5H	Enterprise 6.4 TB SSD PCIe4 NVMe U.2 module for IBM i	6.4	7.0	4.1	1600	280	80	15	3	Oct-23
EC5W	Enterprise 6.4 TB SSD PCIe4 NVMe U.2 module for IBM i	6.4	7.2	3.8	1500	250	80	20	3	Jul-20
ES4C	Enterprise 1.6 TB SSD PCIe4 NVMe U.2 module for IBM i	1.6	7.3	3.1	1740	320	138	215	3	Jul-24
ES4E	Enterprise 3.2 TB SSD PCIe4 NVMe U.2 module for IBM i	3.2	7.4	6	1700	280	134	114	3	Jul-24
ES4G	Enterprise 6.4 TB SSD PCIe4 NVMe U.2 module for IBM i	6.4	6.4	4.6	1380	390	152	107	3	Jul-24
ECTB	15.3 TB Mainstream NVMe U.2 SSD 4k for IBM i	15.3	2.2	5.6	1500	300	67	15	1	Jul-24

NVMe ECTB 15.3 TB Mainstream supporte jusqu'à 64 namespaces

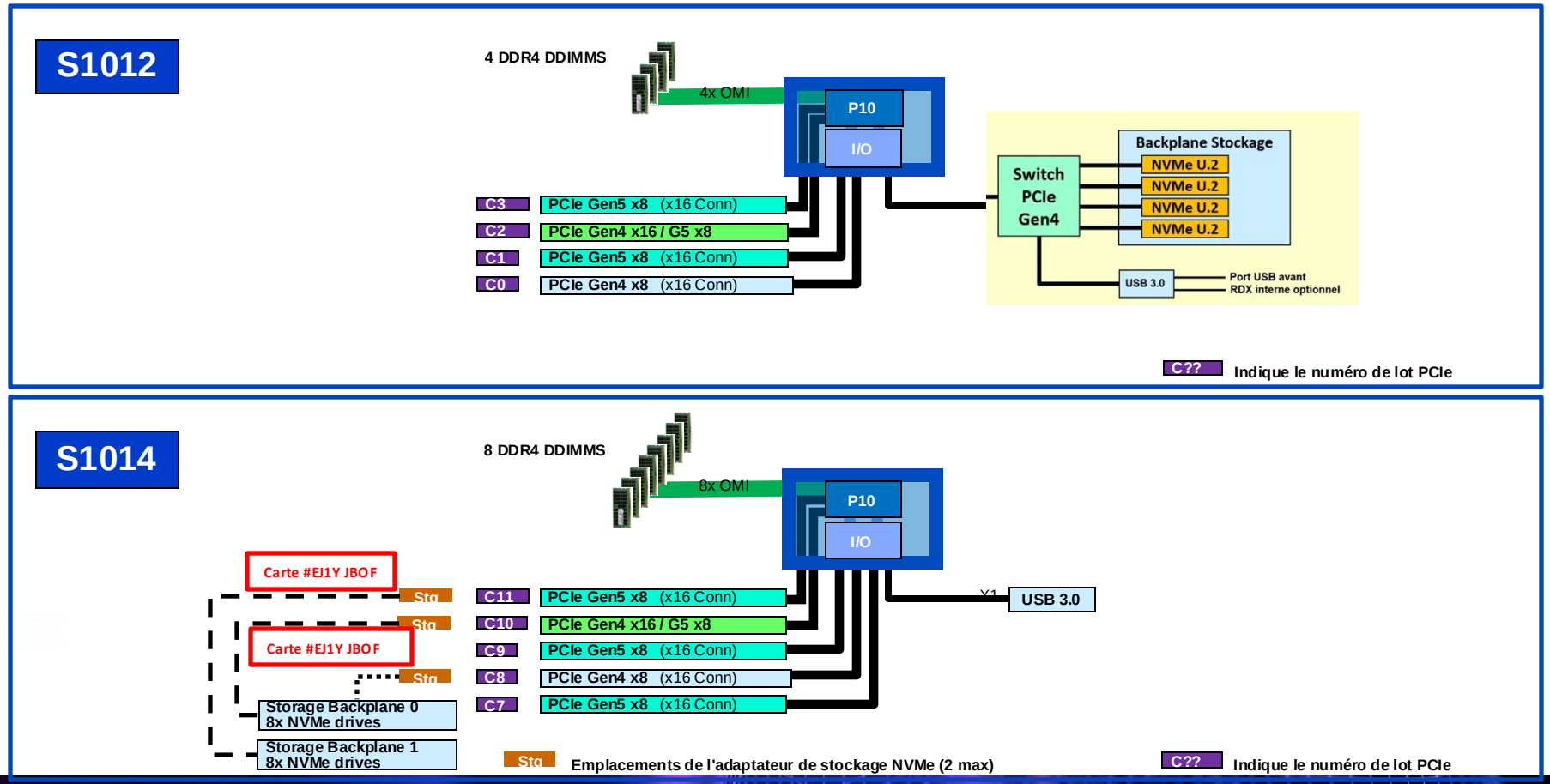
Versions IBM i et Technology Refresh supporté par NVMe et type de configuration

- iVirt = IBM i dans IBM i
- Native = dédié IBM i
- VIOS = disque virtuel via vSCSI
- All = tout type d'architecture

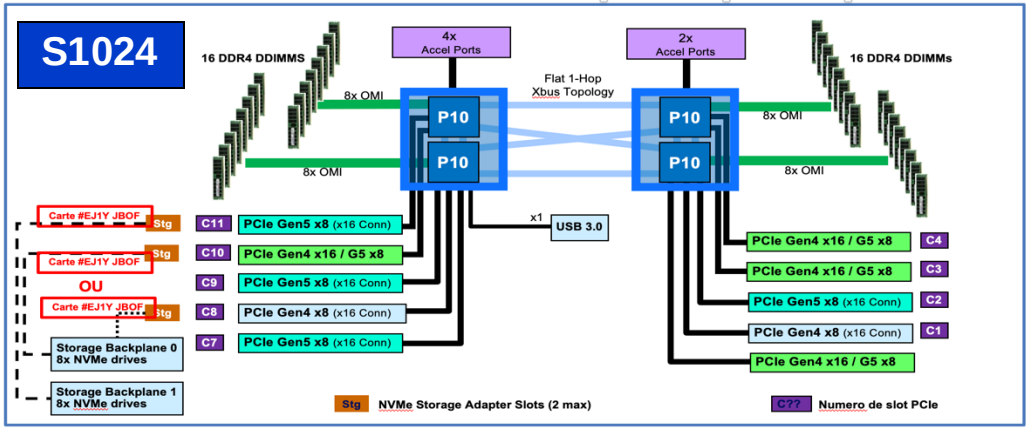
IBM i I/O Support	Type of Configuration (Native, VIOS, iVirt, All)	IBM i 7.5	IBM i 7.4	IBM i 7.3
Enhancements from July 2024				
#ESR0 NVMe Expansion Drawer adds Multipath capability (See also May 2023)	All	Tech Refresh 4	Tech Refresh 10	-
#EN24, #EN26 - PCIe 4-Port 25/10/1 GbE RoCE SFP28 Adapter adds SR-IOV support	All (see also Nov 2023)	Tech Refresh 4	Tech Refresh10	Base (iVirt, VIOS only)
#ES4C - Enterprise 1.6 TB SSD PCIe4 NVMe U.2 module for IBM i	Native, iVirt	Tech Refresh 4	Tech Refresh 10	-
#ES4E - Enterprise 3.2 TB SSD PCIe4 NVMe U.2 module for IBM i	Native, iVirt	Tech Refresh 4	Tech Refresh 10	-
#ES4G - Enterprise 6.4 TB SSD PCIe4 NVMe U.2 module for IBM i	Native, iVirt	Tech Refresh 4	Tech Refresh 10	-
#ES4B - Enterprise 1.6 TB SSD PCIe4 NVMe U.2 module for AIX/Linux	VIOS	Base	Base	-
#ES4D - Enterprise 3.2 TB SSD PCIe4 NVMe U.2 module for AIX/Linux	VIOS	Base	Base	-
#ES4F - Enterprise 6.4 TB SSD PCIe4 NVMe U.2 module for AIX/Linux	VIOS	Base	Base	-
#ECTB - New 15.3 TB Mainstream PCIe4 NVMe U.2 module for IBM i	Native, iVirt	Tech Refresh 4	Tech Refresh 10	-
#ECT9 - New 15.3 TB Mainstream PCIe4 NVMe U.2 module for AIX/Linux	VIOS	Base	Base	-

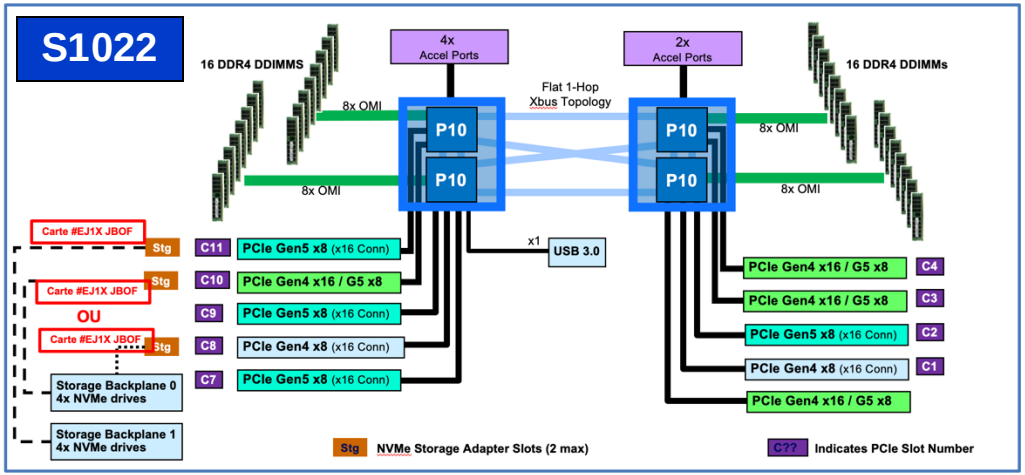
https://www.ibm.com/support/pages/ibm-i-io-support-summary

Architecture NVMe sur Power10

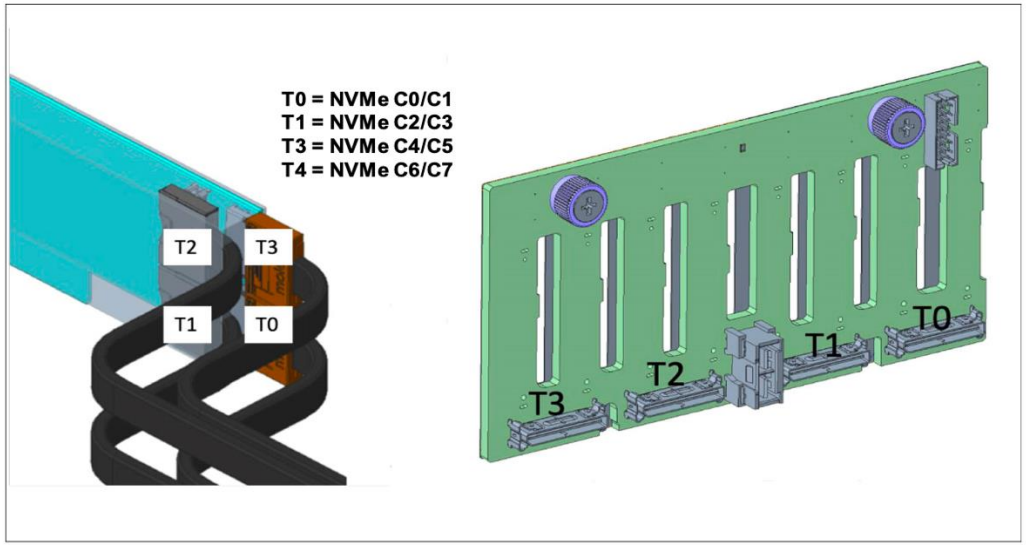


Architecture NVMe sur Power10



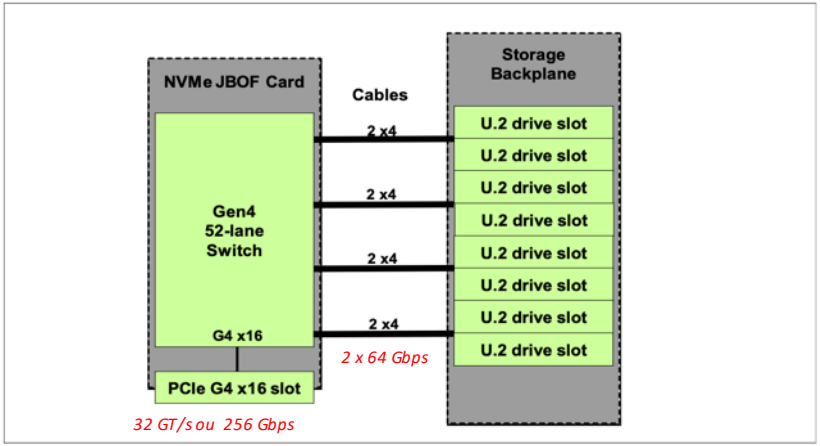


Carte PCIe4 4-port NVMe JBOF adapter #EJ1Y ou #EJ1X



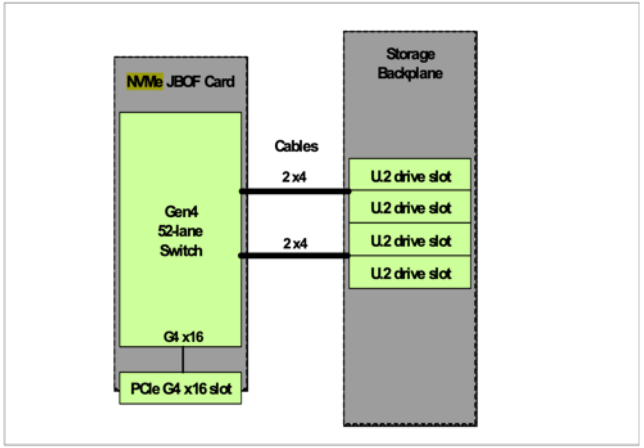
Architecture Cage de disque NVMe et carte JBOF*

S1024 / S1014



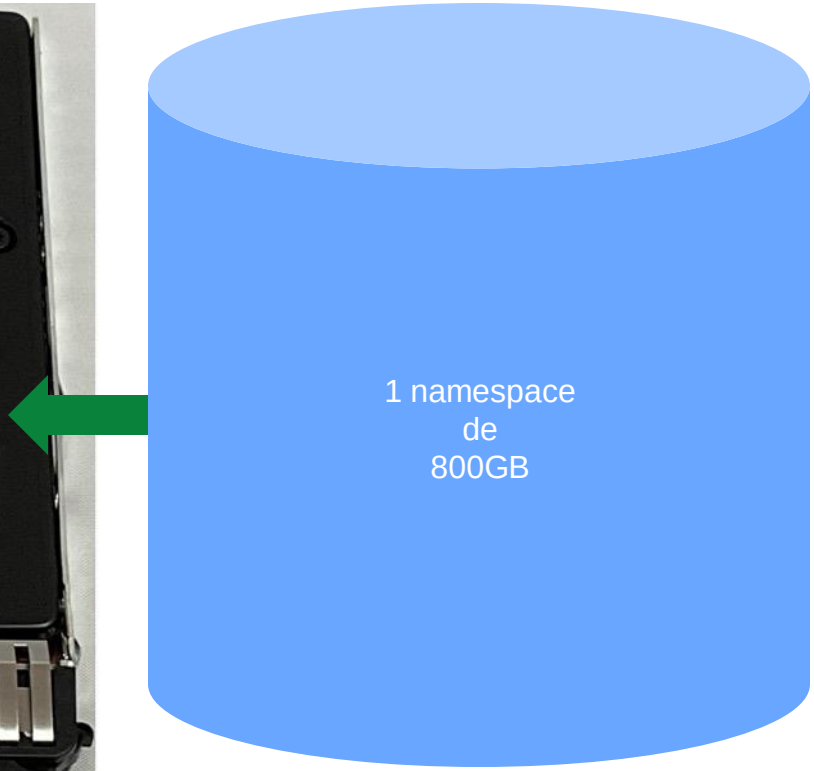
Cage de disque EJ1Y - Jusqu'à 8 NVMe
NVMe attaché au processeur par la carte JBOF
Chaque connecteur de la carte JBOF gère 2 NVMe
Carte sur slot Pcie Gen4 x16 = 16 lignes
Debit de 16 lignes x 16 Gbps/ ligne = **256 Gbps de bande passante**
Chaque NVMe a un debit theorique de 4 lignes * 16Gbps = **64 Gbps**
Le slot PCIe G4 x16 supportant la carte JBOF supporte *une bande passante de 32 GT/s ou 256 Gbps*

S1022 / S1022s



Cage de disque EJ1X
Jusqu'à 4 NVMe
NVMe attaché au processeur par la carte JBOF
Chaque connecteur de la carte JBOF gère 2 NVMe

NVMe Namespace – exemple NVMe ES3A - 800GB



NVMes sont livrés avec 1 namespace configuré prenant tout l'espace de stockage par défaut

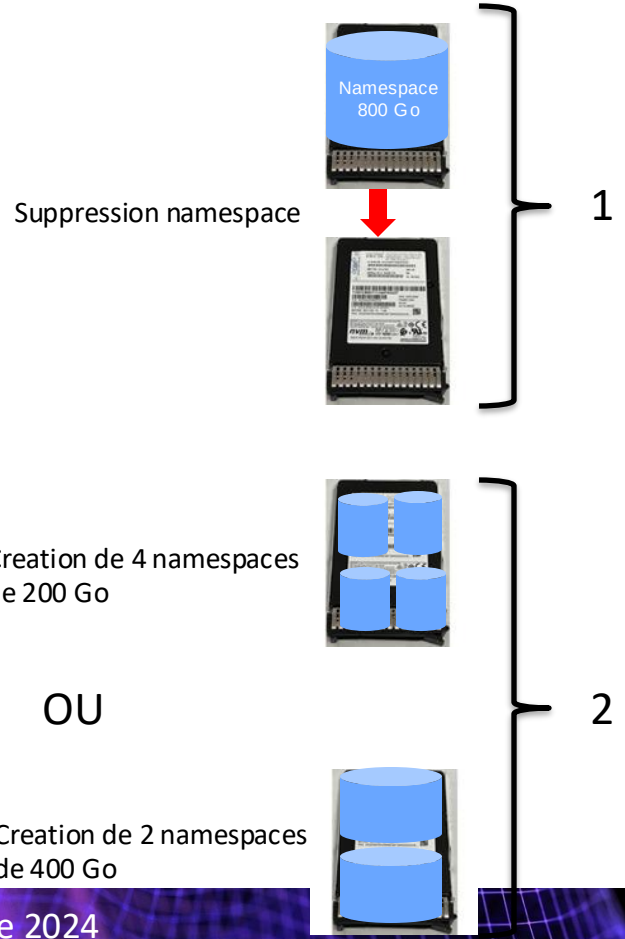
Configuration des disques NVMe pour IBM i

1. Suppression du namespace par défaut de l'unité
2. Création des namespaces de 200 Go ou 400 Go sur l'unité NVMe

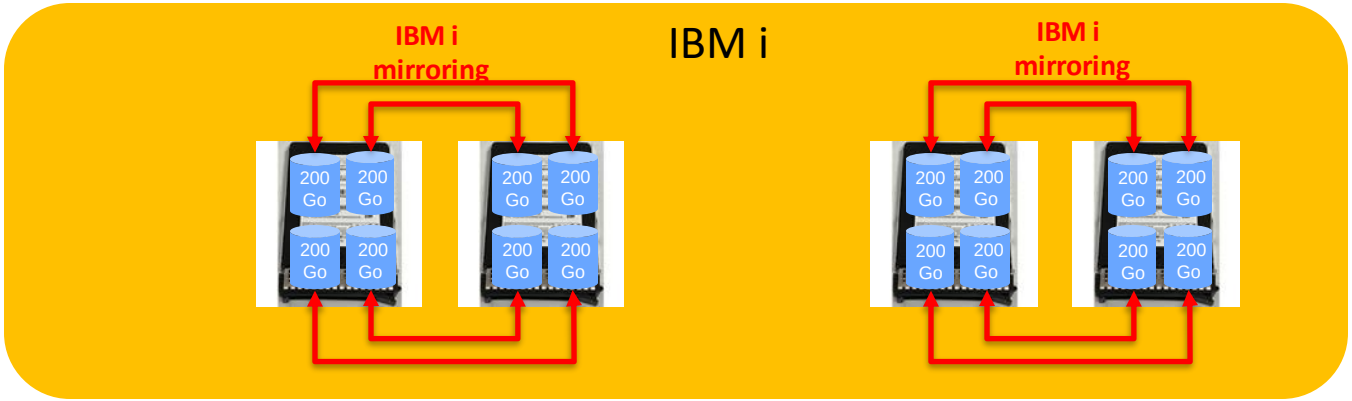
1 namespace = 1 disque IBM i = 1 bras disque

Possibilité de garder de l'espace non affécté à un namespace pour utilisation ultérieure

Même procédure pour les NVMe gérés par les VIOS



Namespace sur IBM i et protection (Mirroring)



- Exemple : 4 unités NVMe FC ES3A de 800 Go
  - création de 4 namespace de 200 Go par NVMe
  - IBM i voit donc 16 disques logiques
- Protection des namespaces par mirroring IBM i – seule protection disponible
  - Mirroring entre namespace d'unité NVMe différentes et de même taille
- Possibilité de créer des namespaces de tailles différentes sur les unités NVMe.
- Dans exemple : IBM i a 16 disques de 200 Go avec un espace de stockage de 1600 Go disponible

Université IBM i

19 et 20 novembre 2024

IBM i
continuous innovation
continuous integration

Performance NVMe

NVMe vs SATA / SAS

	NVMe	SATA	SAS
Bande passante	64 Gbps*	6 Gps	12 Gbps
Nombre E/S par seconde (IOPS)	> 1 000 000	100 000	400 000
Nbre de commande par File d'attente	64 000	32	256
Nbre file d'attente commande	64 000	1	1
Latence lecture	80 us	1,8 ms**	150 us
Latence écriture	15 us	3,6 ms**	60 us

* Bande passante pour 1 unité NVMe U.2 . Bande passant de 256 Gbps par carte JBOF

** Latence en ms dû aux contraintes physiques de deplacement bras disque et vitesse des plateaux disques. D'où la nécéssité d'avoir des cartes controleurs disques avec cache pour masquer cela

Performance NVMe

	32 namespaces 200Gb 2 NVMe	32 namespaces 200Gb 8 NVMe	64 namespaces 100Gb 8 NVMe	16 namespaces 400Gb 2 NVMe	1 namespace 3,2Tb 2 NVMe	12 namespaces 200Gb 3 NVMe
Temps de réponse AVG Write	0,0357 ms	0,0230 ms	0,0232 ms	0,0395 ms	0,0275 ms	0,0269 ms
Temps de réponse AVG Read	0.1629 ms	0,1230 ms	0,1221 ms	0.1631 ms	0.1530 ms	0,1521 ms
Disque busy % max Write	24,86 %	20,51 %	10,83 %	42,74 %	99,35 %	41,17 %
Disque busy % max Read	18,69 %	13,86 %	7,035 %	34,37 %	100 %	43,85 %
E/S max écriture	504 544	643 437	655 964	501 399	566 204	618 092
E/S max lecture	76 021	77 741	78 235	77 747	78 232	76 640

Conclusion :

Meilleures performances sur plusieurs NVMe

+ Namespaces = + bras disque = meilleures performances

IBM i
continuous innovation
continuous integration

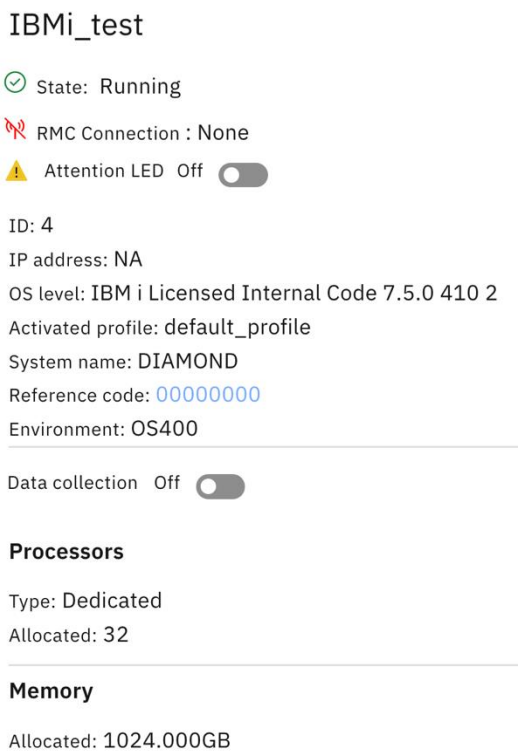
NVMe
Stress Tests

NVMe

Tests NVMe sur un S1024 48 cores 2Tb de mémoire

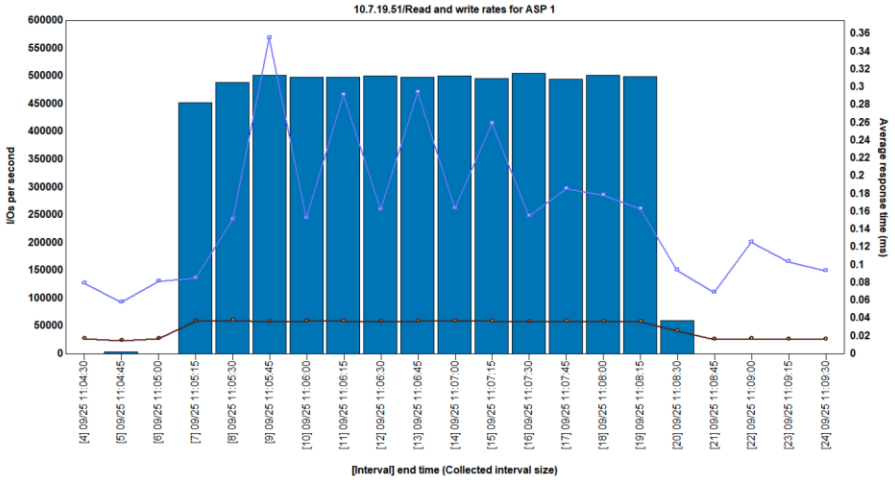
Fichier d'écriture (WF) : Programmes RPG qui écrivent dans 24 fichiers en même temps, 70 millions d'enregistrements dans chaque fichier en 24 tâches. L'enregistrement dans chaque fichier est de 75 caractères décimaux condensés. OVRDBF avec FRCRATIO(1000) est utilisé pour chaque fichier.

Readfile (RF) : Programmes RPG qui lisent séquentiellement les fichiers contenant 70 millions d'enregistrements qui ont été écrits avec WF en 24 tâches. Chaque fichier est lu 16 fois. OVRDBF avec SEQONLY(*YES 2000) est utilisé pour chaque fichier.



NVMe

32 namespaces de 200 GB sur 2 NVMe



Temps de réponse AVG Read
0,1629 ms

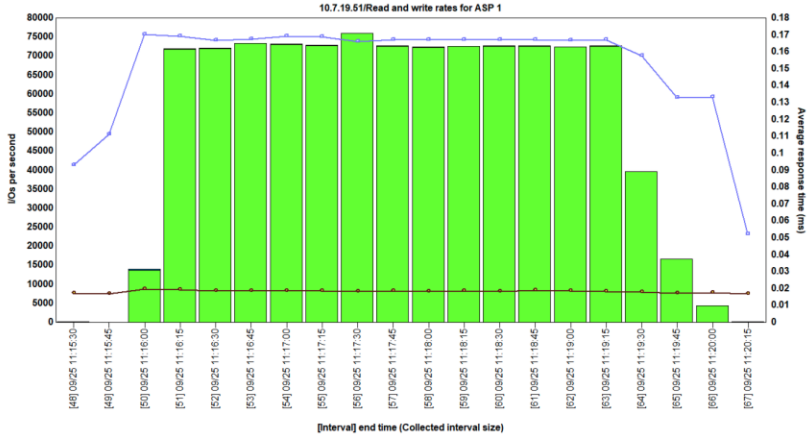
76021 I/Os max

Disk busy : 34,37%

Temps de réponse AVG Write
0,0357 ms

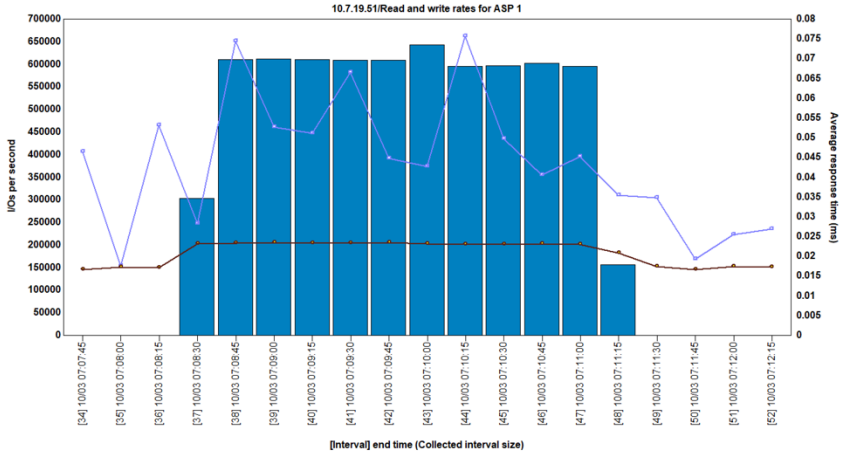
504544 I/Os max

Disk busy : 42,74%



NVMe

32 namespaces de 200 GB sur 8 NVMe



Temps de réponse AVG Read
0,1230 ms

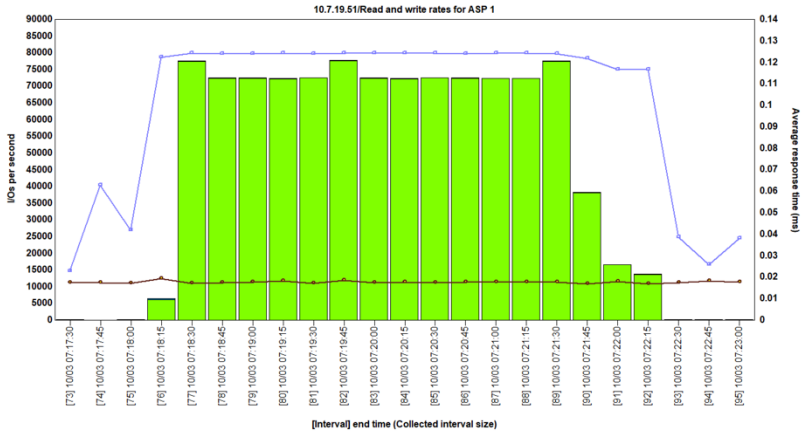
77741 I/Os max

Disk busy : 13,86%

Temps de réponse AVG Write
0,0230 ms

643437 I/Os max

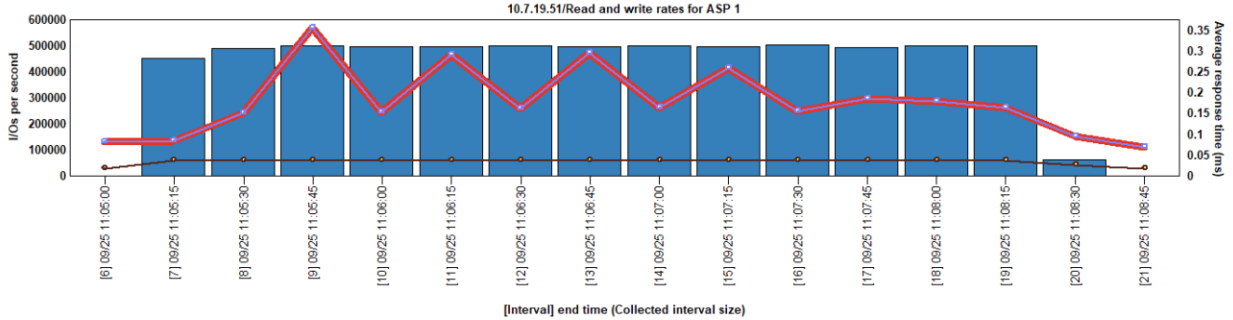
Disk busy : 20,51%



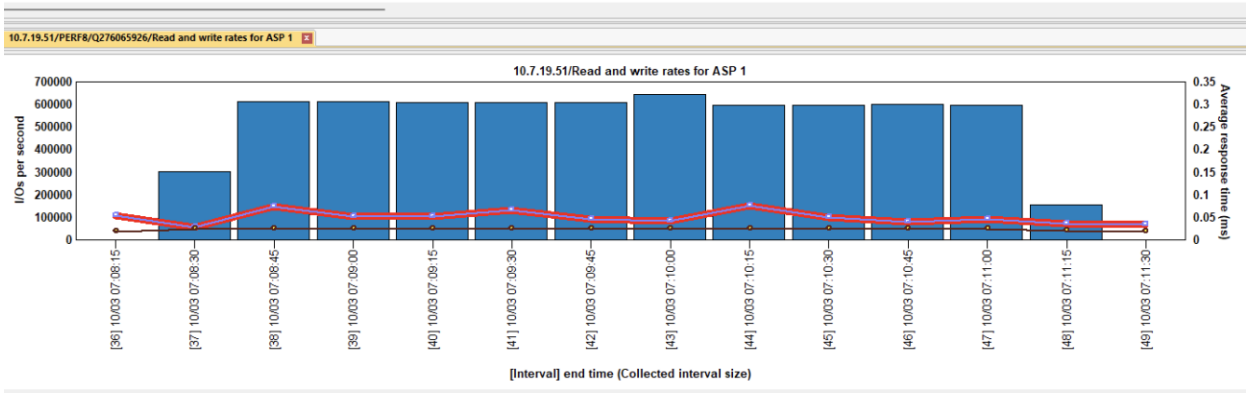
NVMe

32 namespaces de 200 GB 2 versus 8 NVMe

2 NVMe



8 NVMe

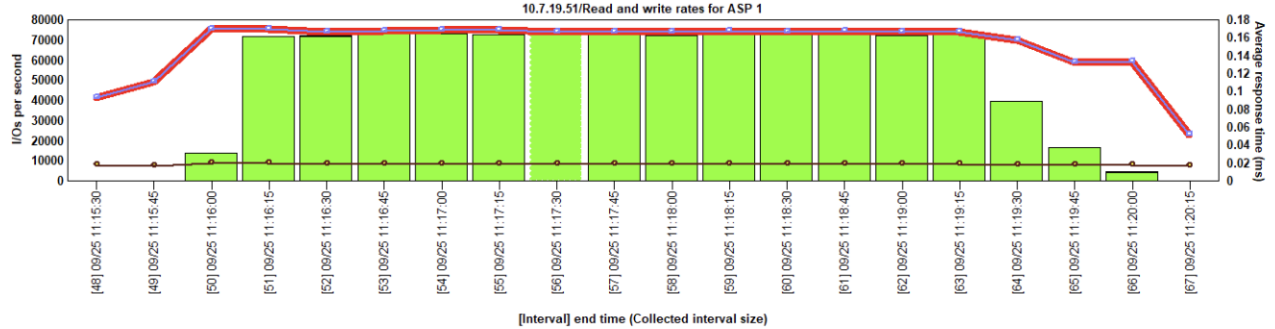


Ecriture

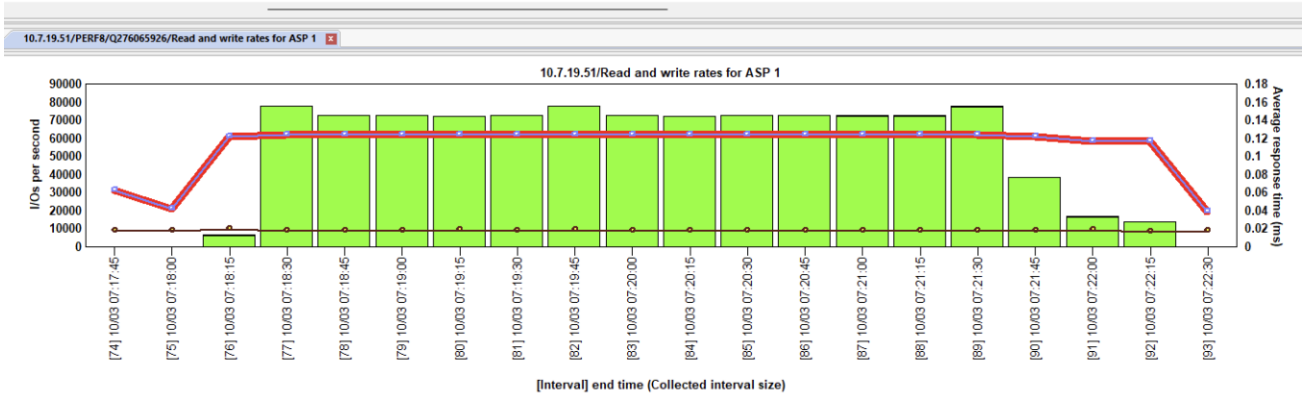
NVMe

32 namespaces de 200 GB 2 versus 8 NVMe

2 NVMe



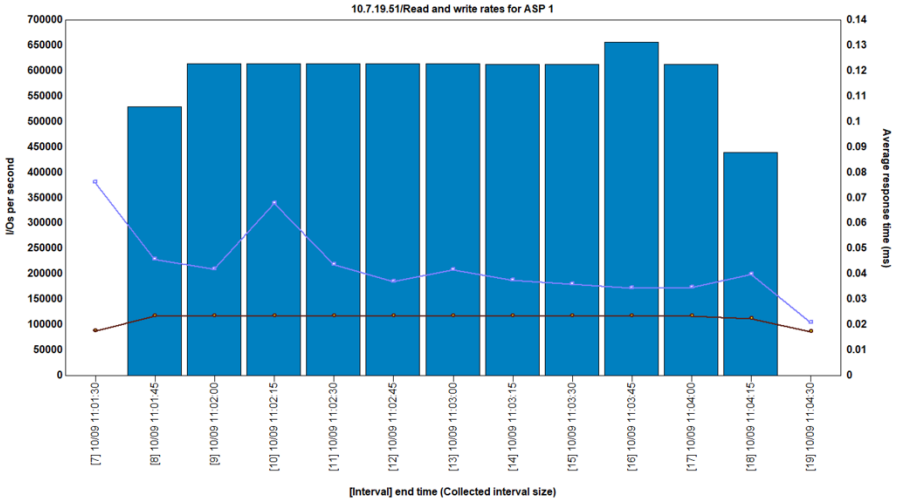
8 NVMe



Lecture

NVMe

64 namespaces de 100 GB sur 8 NVMe



Temps de réponse AVG Read
0,01221 ms

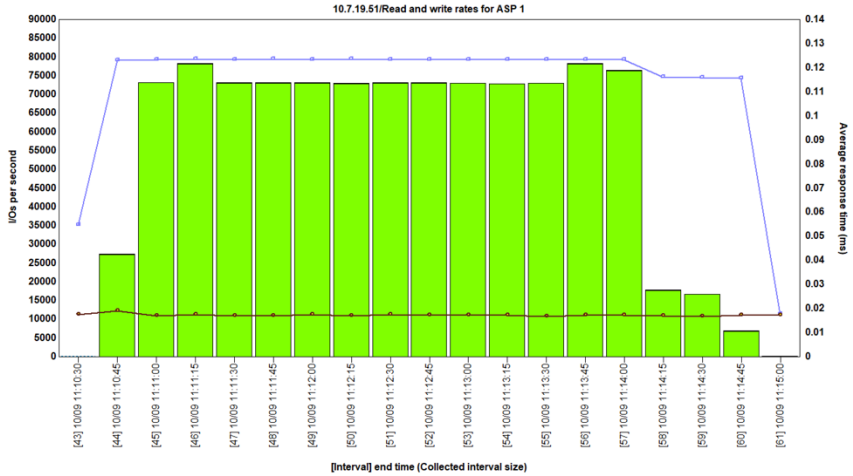
78235 I/Os max

Disk busy : 7,035%

Temps de réponse AVG Write
0,0232 ms

655964 I/Os max

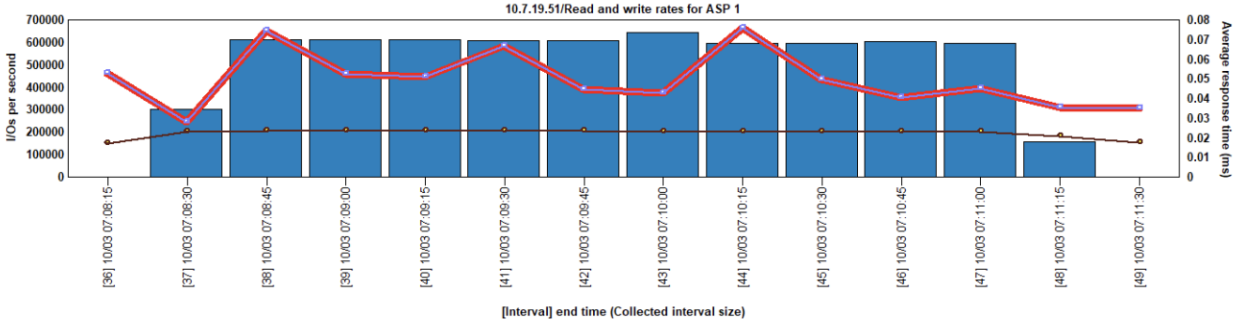
Disk busy : 10,83%



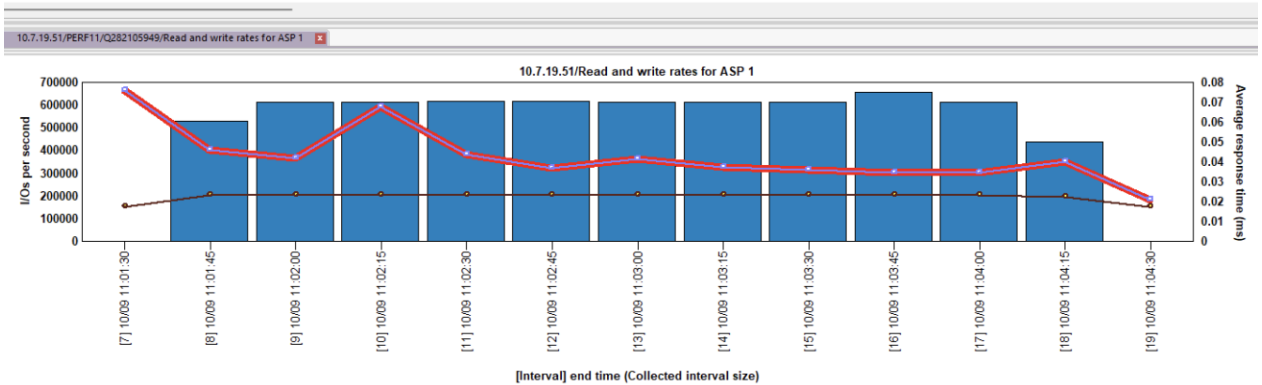
NVMe

8 NVMe 32 versus 64 namespaces

32 Namespaces



64 Namespaces

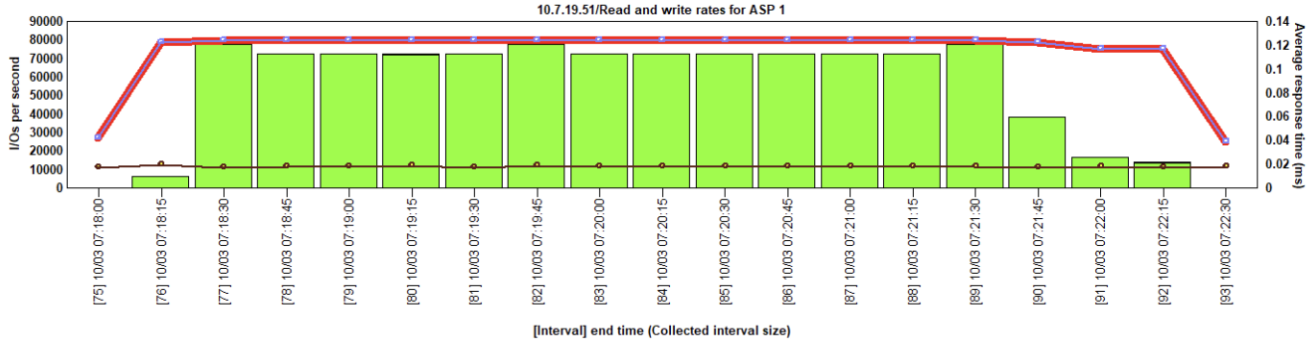


Ecriture

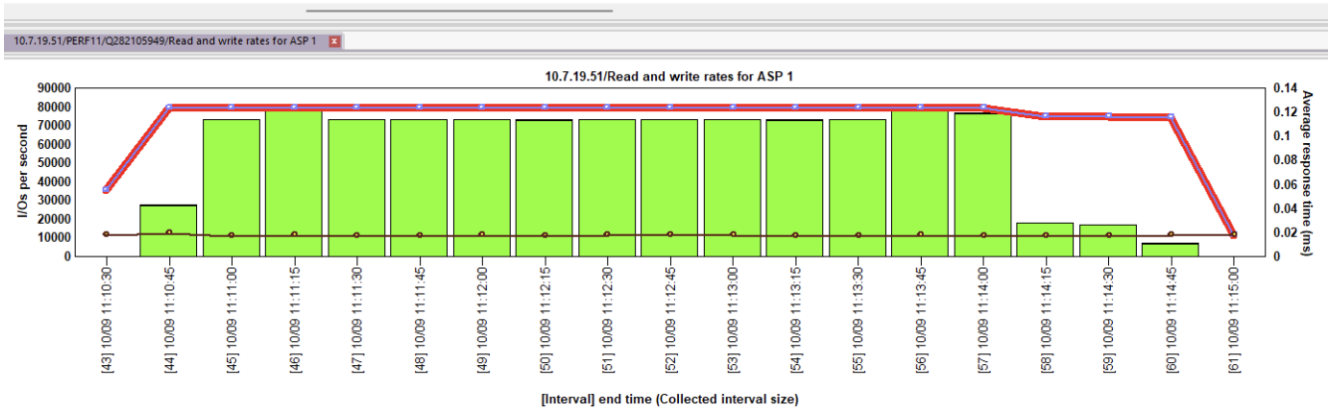
NVMe

8 NVMe 32 versus 64 namespaces

32 Namespaces



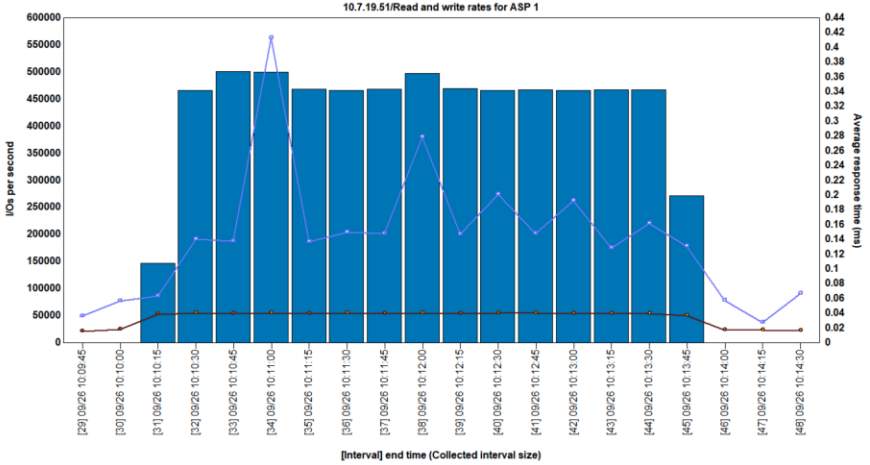
64 Namespaces



Lecture

NVMe

16 namespaces de 400 GB sur 2 NVMe



Temps de réponse AVG Read
0,1631 ms

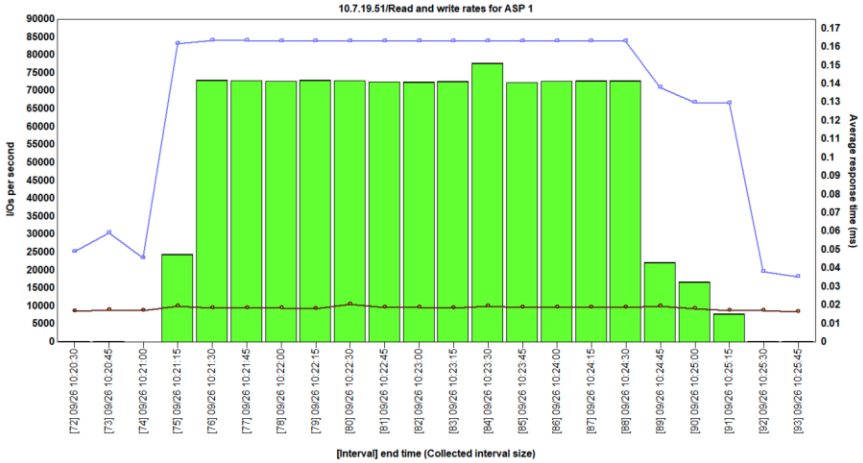
77747 I/Os max

Disk busy : 34,37%

Temps de réponse AVG Write
0,0395 ms

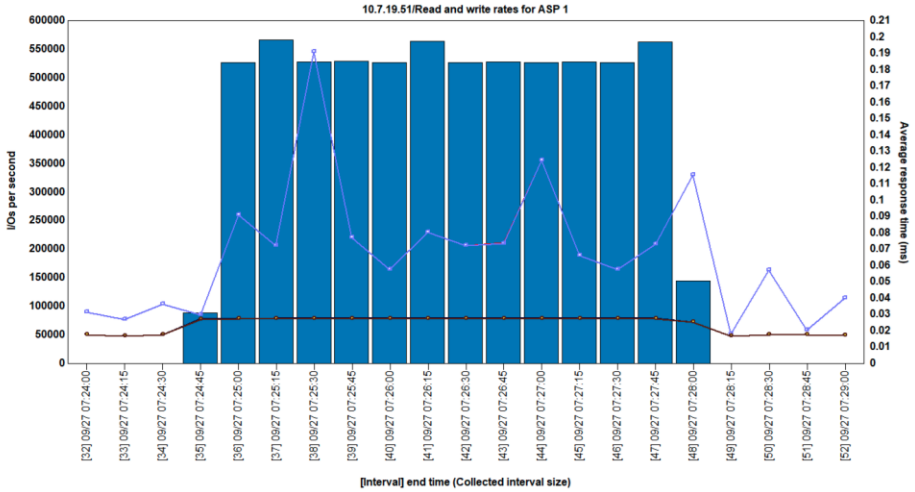
501399 I/Os max

Disk busy : 42,74%



NVMe

1 namespace de 3,2 TB sur 2 NVMe



Temps de réponse AVG Write
0,0275 ms

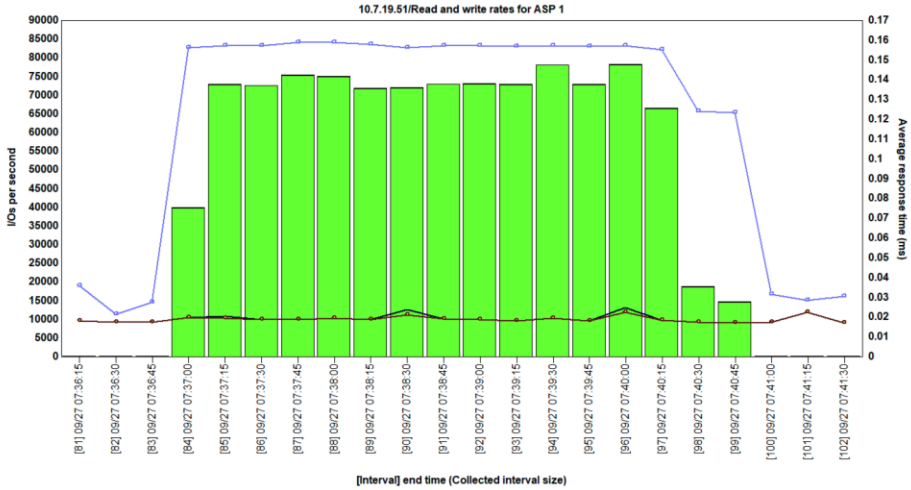
566204 I/Os max

Disk busy : 99,35%

Temps de réponse AVG Read
0,1530 ms

78232 I/Os max

Disk busy : 100%



3 NVMe

Protection Planar ET Bus

```
                    Confirm Start Mirrored Protection
_
 Press Enter to confirm your choice to start mirrored
 protection.  During this process the partition will be IPLed.
 You will return to the DST main menu after the IPL is
 complete.  The ASP will have the displayed protection.

 Press F12 to return to change your choice.

                Serial                          Resource                    Hot Spare
ASP Unit        Number          Type Model      Name        Protection      Protection
  1                                                         Mirrored
        1  Y4D8WZR25XQ4         6B7D  205       DD001       Planar          N
        1  YAXR2VUSAEQ7         6B7D  205       DD017       Planar          N
        2  YXMBMY79DAHW         6B7D  205       DD002       Planar          N
        2  YZDCA5UV6E55         6B7D  205       DD018       Planar          N
        3  YR3DD789TBLG         6B7D  205       DD003       Planar          N
        3  Y6ETHC974336         6B7D  205       DD019       Planar          N
        4  YC6TJLGKQEQL         6B7D  205       DD004       Planar          N
        4  Y7D3T3XBAE4K         6B7D  205       DD020       Planar          N
                                                                          More..

 F12=Cancel
```

```
                Serial                          Resource                    Hot Spare
ASP Unit        Number          Type Model      Name        Protection      Protection
        5  YHT4PTK79UGA         6B7D  205       DD005       Bus             N
        5  YFPF3ME4QZCP         6B7D  205       DD013       Bus             N
        6  Y5ARQXGV89M5         6B7D  205       DD006       Bus             N
        6  YM86ZQWYBWYT         6B7D  205       DD014       Bus             N
        7  YQ972CSGJRH6         6B7D  205       DD007       Bus             N
        7  Y5WJNMU52PU6         6B7D  205       DD015       Bus             N
        8  YGRJCCNJWYDQ         6B7D  205       DD008       Bus             N
        8  YKK2AGTUXYWV         6B7D  205       DD016       Bus             N
        9  YBVY2UJWDUG9         6B7D  205       DD009       Planar          N
                                                                          More...
```

ASP Unit	Serial Number	Type	Model	Resource Name	Protection	Hot Spare Protection
9	YV7DUVAU8EDW	6B7D	205	DD021	Planar	N
10	YUKS2ZEAR4KT	6B7D	205	DD010	Planar	N
10	YK6CXSRZPHEA	6B7D	205	DD022	Planar	N
11	YFF4USALHQC8	6B7D	205	DD011	Planar	N
11	Y5Z46F6BEAY7	6B7D	205	DD023	Planar	N
12	YAUG86GJWAGQ	6B7D	205	DD012	Planar	N
12	YKCETPH9FGDV	6B7D	205	DD024	Planar	N

NVMe

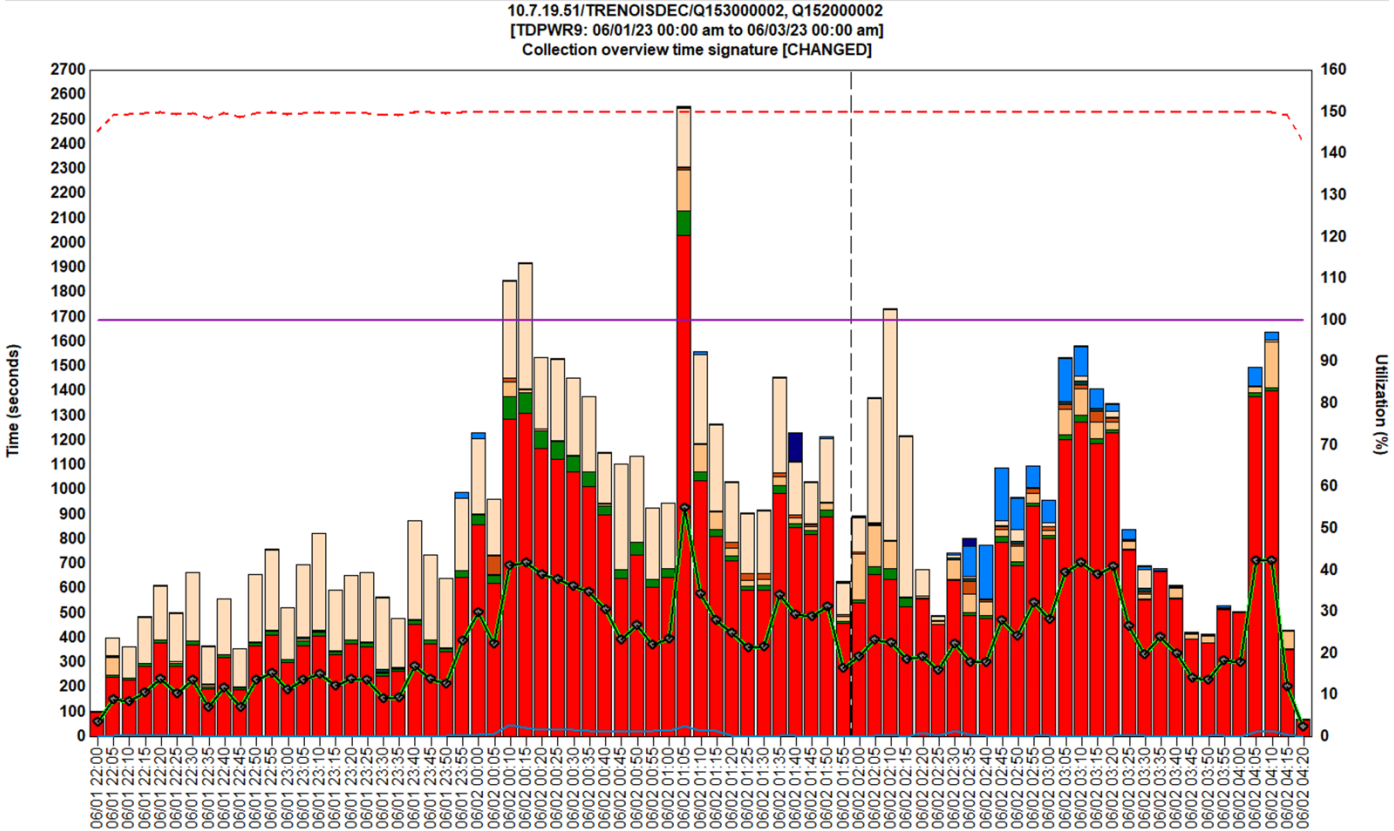
	32 namespaces 200Gb 2 NVMe	32 namespaces 200Gb 8 NVMe	64 namespaces 100Gb 8 NVMe	16 namespaces 400Gb 2 NVMe	1 namespace 3,2Tb 2 NVMe	12 namespaces 200Gb 3 NVMe
Temps de réponse AVG Write	0,0357 ms	0,0230 ms	0,0232 ms	0,0395 ms	0,0275 ms	0,0269 ms
Temps de réponse AVG Read	0.1629 ms	0,1230 ms	0,1221 ms	0.1631 ms	0.1530 ms	0,1521 ms
Disque busy % max Write	24,86 %	20,51 %	10,83 %	42,74 %	99,35 %	41,17 %
Disque busy % max Read	18,69 %	13,86 %	7,035 %	34,37 %	100 %	43,85 %

	32 namespaces 200Gb 2 NVMe	32 namespaces 200Gb 8 NVMe	64 namespaces 100Gb 8 NVMe	16 namespaces 400Gb 2 NVMe	1 namespace 3,2Tb 2 NVMe	12 namespaces 200Gb 3 NVMe
I/Os max en Write	504544	643437	655964	501399	566204	618092
I/Os max en Read	76021	77741	78235	77747	78232	76640

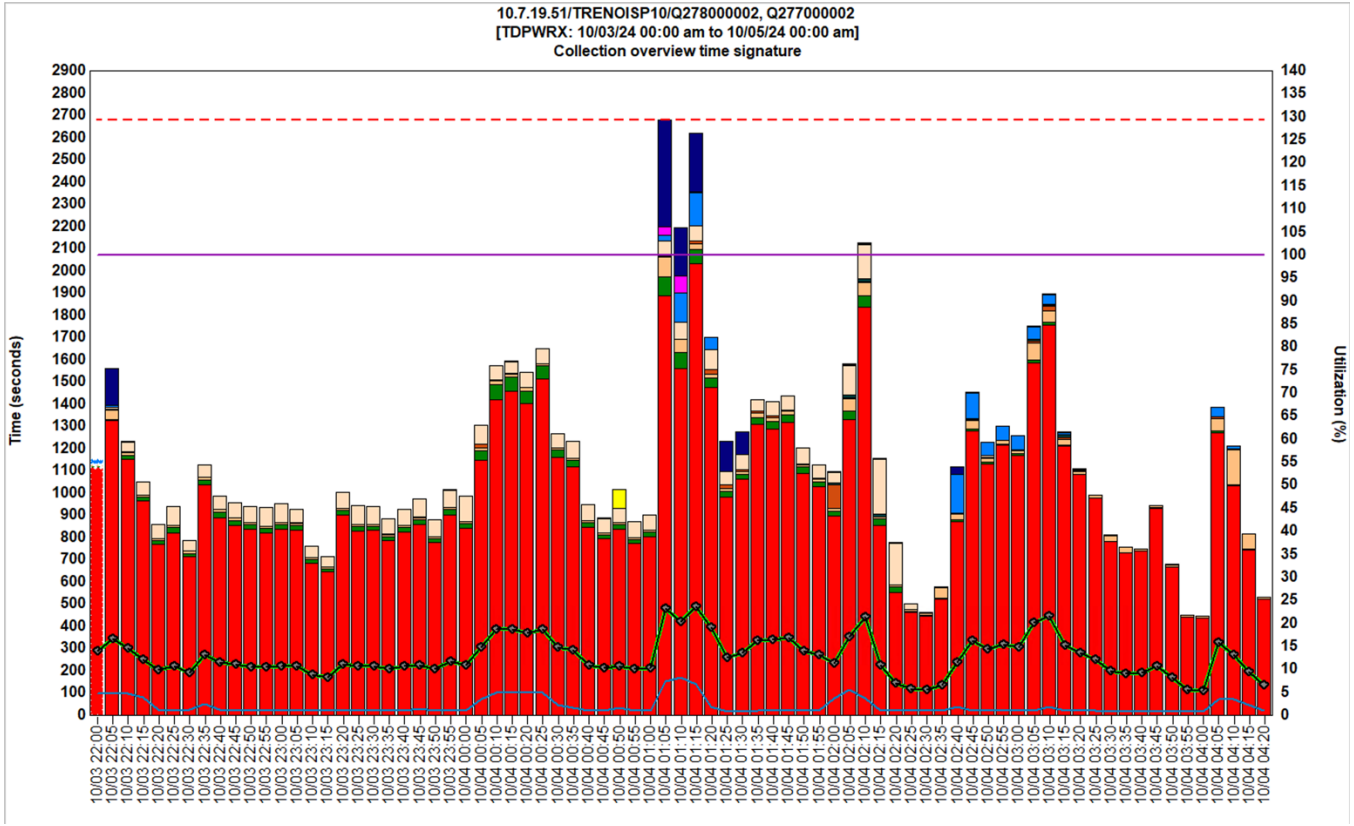
IBM i
continuous innovation
continuous integration

NVMe
Retour d’expérience client

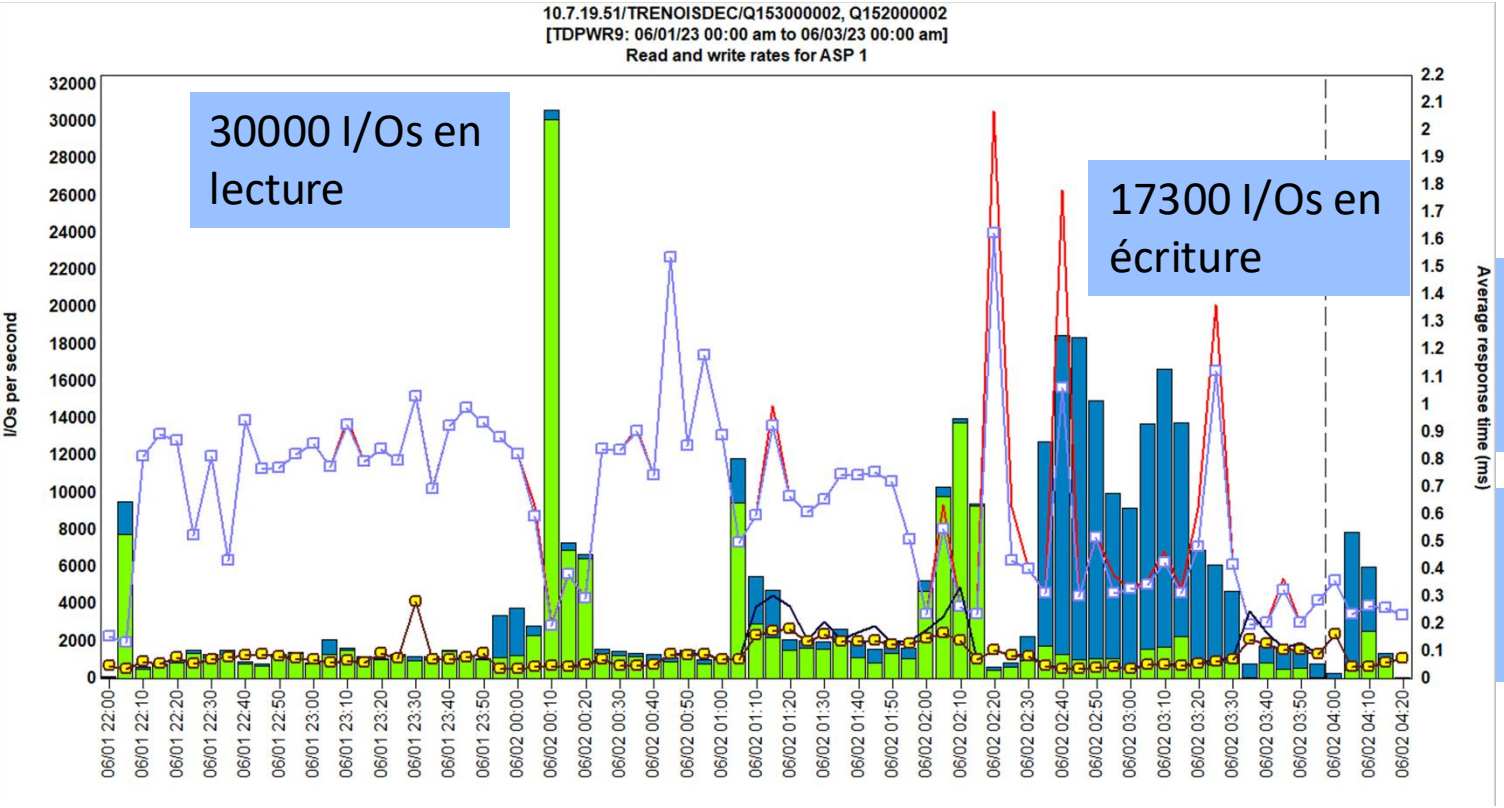
Résumé de l'activité sur P9



Résumé de l'activité sur P10



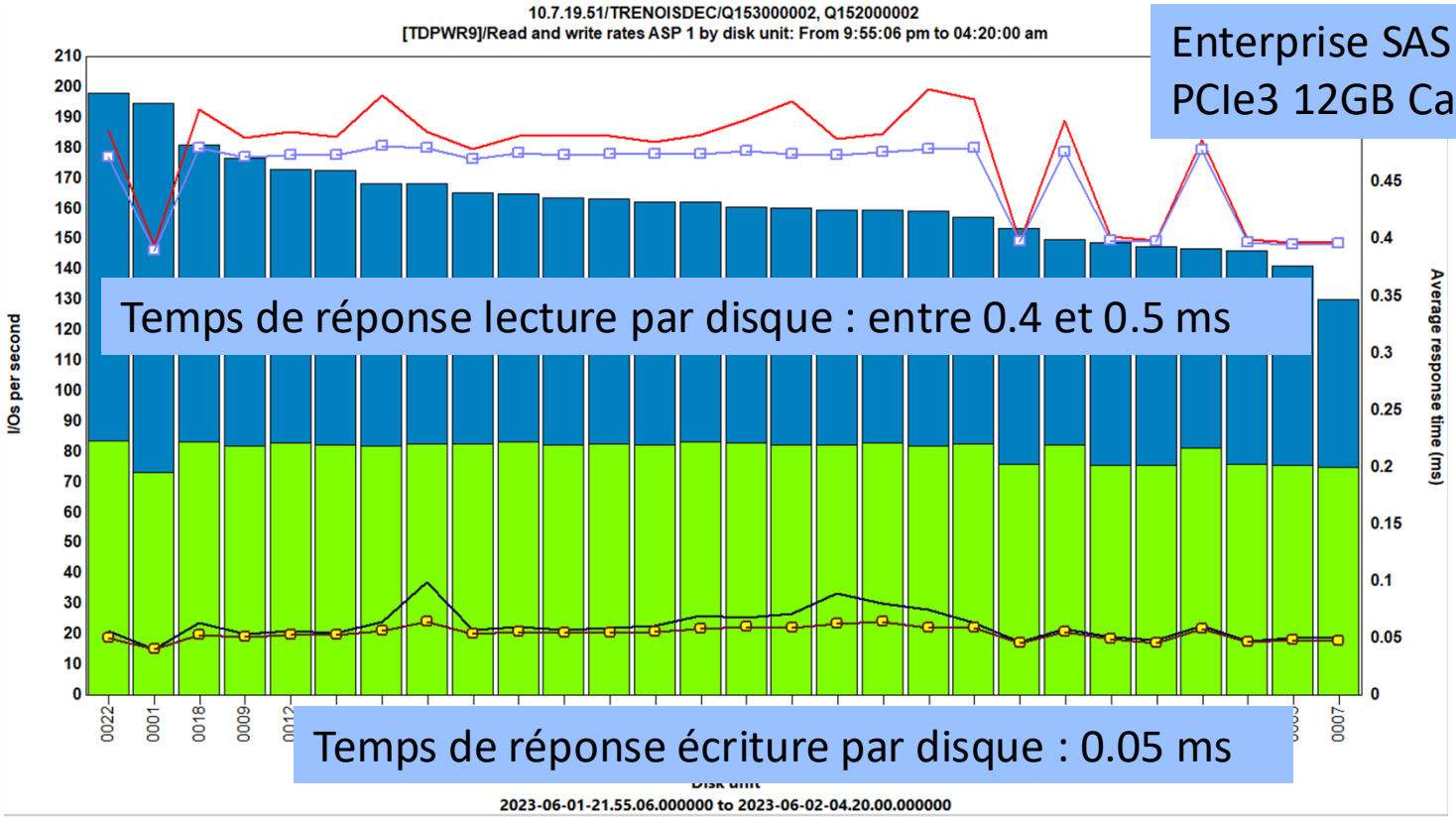
Activité disque sur P9



Temps de réponse écriture : entre 0.1 et 0.2 ms

Temps de réponse lecture : entre 0.5 et 1.6 ms

Activité disque sur P9



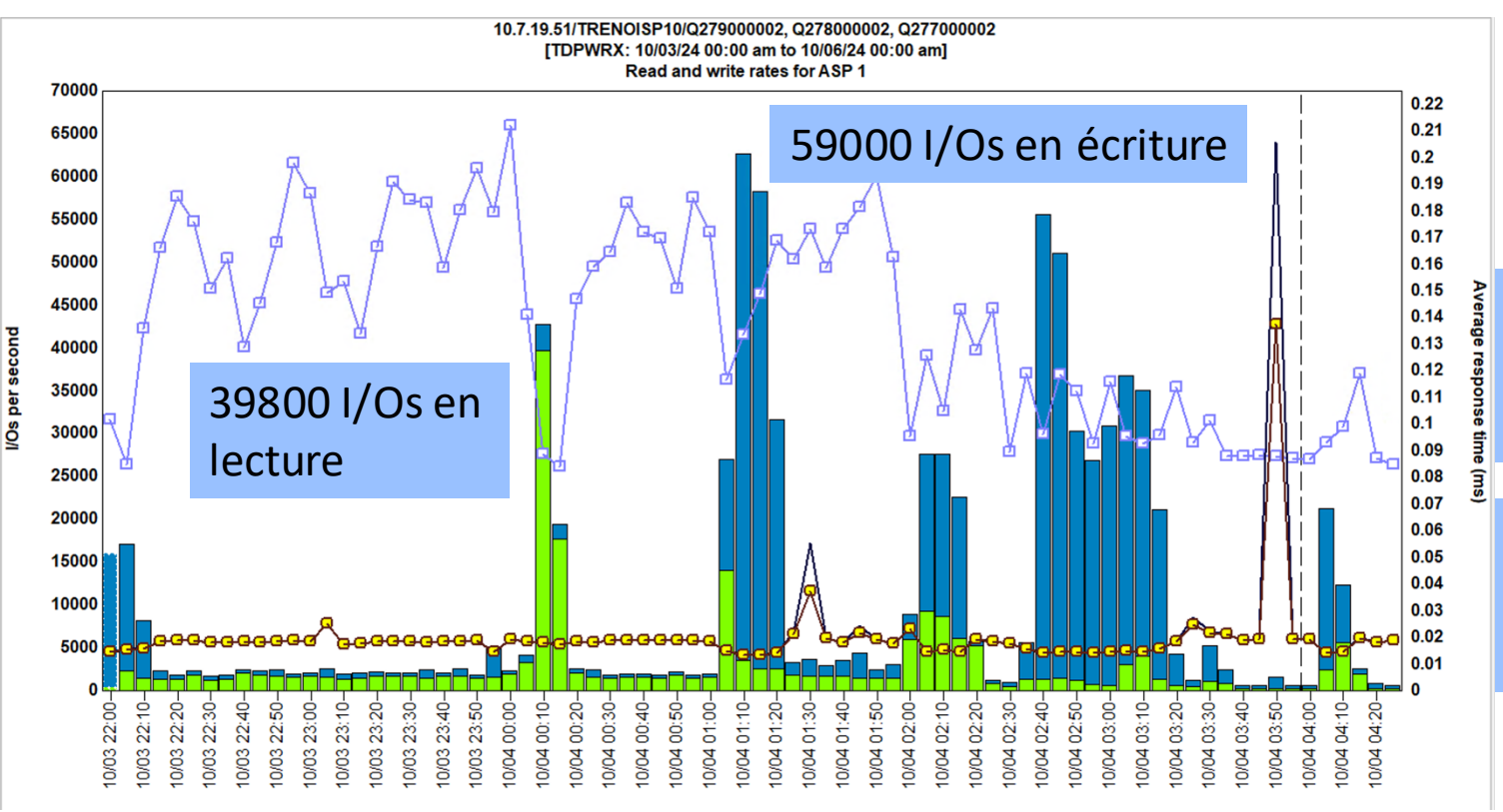
Enterprise SAS 4k SFF-3 et SFF-2
PCIe3 12GB Cache RAID PLUS

Temps de réponse lecture par disque : entre 0.4 et 0.5 ms

28 disques SSD avec controleur RAID

Temps de réponse écriture par disque : 0.05 ms

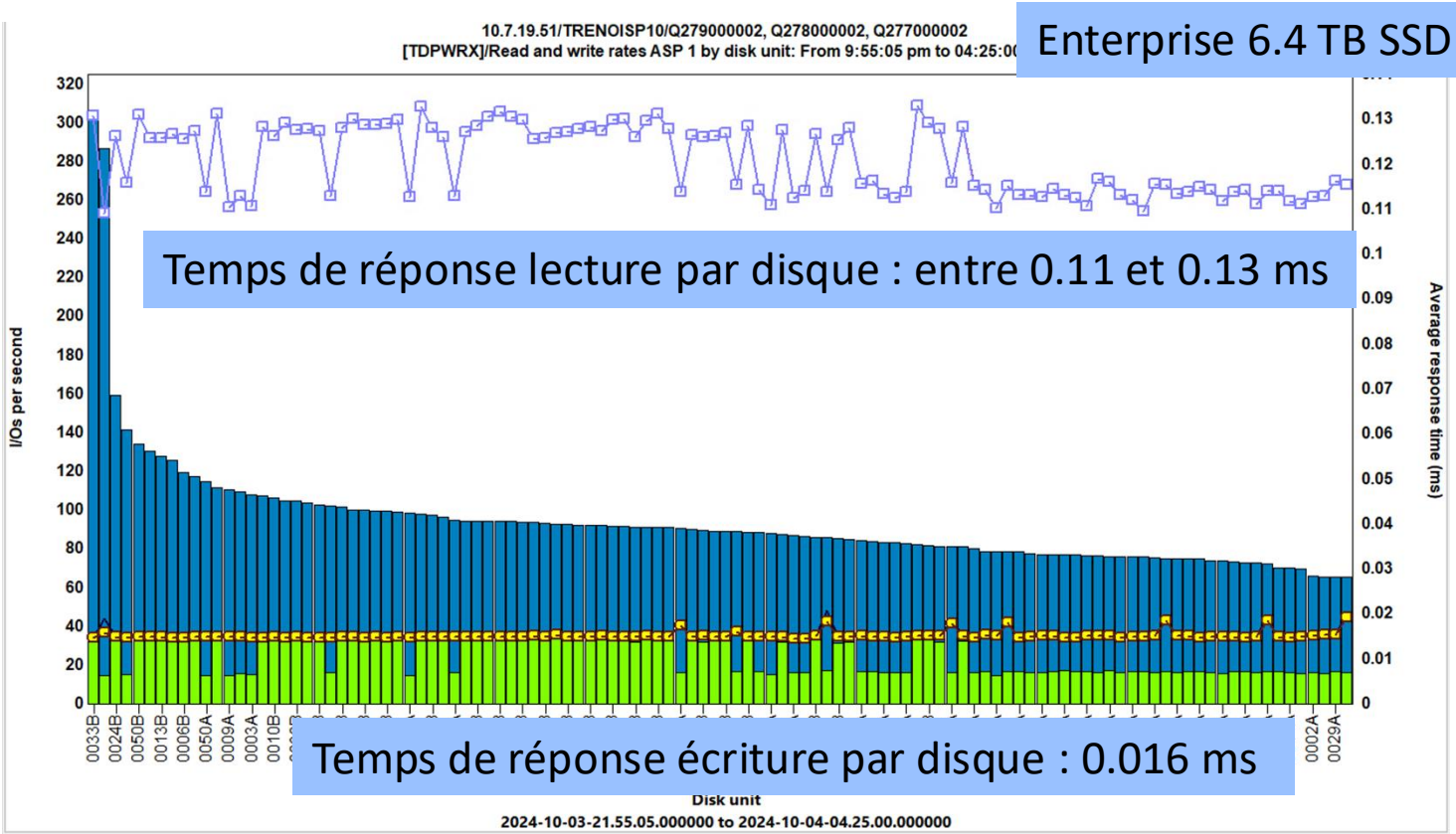
Activité disque sur P10



Temps de réponse écriture : entre 0.02 ms et 0.03 ms

Temps de réponse lecture : entre 0.08 et 0.21 ms

Activité disque sur P10



Enterprise 6.4 TB SSD PCIe4 NVMe U.2

Temps de réponse lecture par disque : entre 0.11 et 0.13 ms

112 namespaces sur 8 NVMe

Temps de réponse écriture par disque : 0.016 ms

Comparaison

	P9	P10	Résultat
Nombre d'IOs max lecture	30000	39800	X 1.32 ⬆
Nombre d'IOs max écriture	17300	59000	X 3.41 ⬆
Temps de réponse lecture	0.5 ms	0.08 ms	X 6.25 ⬇
Temps de réponse écriture	0.1	0.02	X 5 ⬇

Conclusion

+ de jobs

+ d'I/Os

- de temps de batch

IBM i
continuous innovation
continuous integration

IBM

IBM i
ANSIBLE

IBM i

Soyez proactif et non réactif


MERCI