Quels matériels pour nos sites ?

Journées LCG-France 2018

CCIN2P3

cnrs

Agenda

- Cette présentation contient, un rapide bilan sur 10 années d'utilisation de serveurs de calcul et de stockage au CC-IN2P3. Elle contient également un point non exhaustif sur les évolutions et les nouveautés à venir.

- Cette présentation ne contient pas une étude du marché ni une projection de ce que pourrait être le futur.

« Les prévisions sont difficiles, surtout lorsqu'elles concernent l'avenir. »

Citation de Pierre Dac, souvent attribuée à Jacques Chirac.


Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

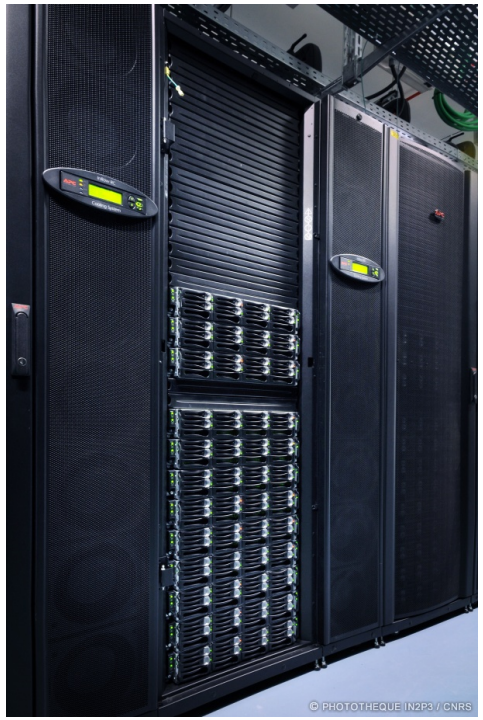
Calcul HTC

2009 – 2018 : 10 ans d'évolution de la plateforme de calcul HTC

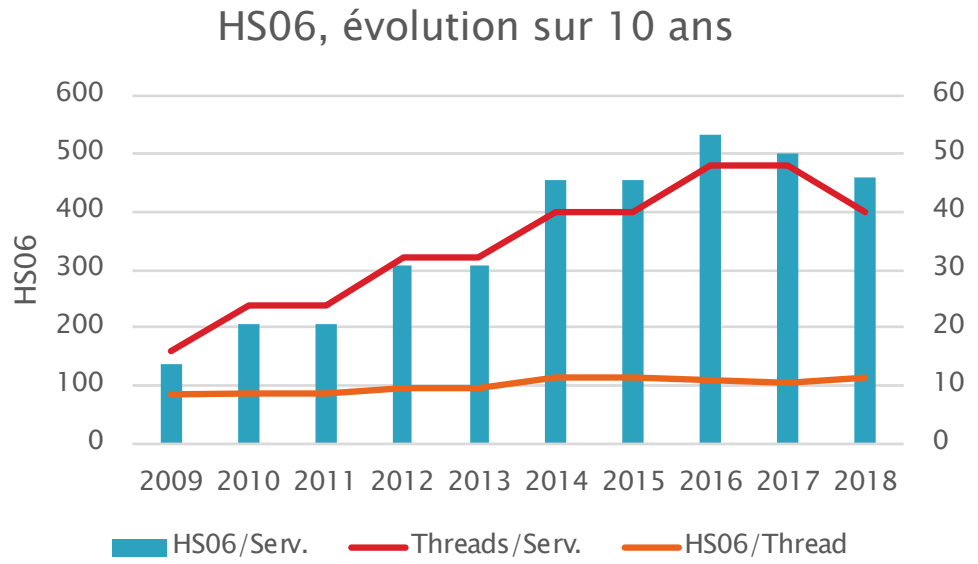
- Plusieurs générations de serveurs Dell PowerEdge se sont succédées.
  - M610, C6100, C6220, C6320, C6420
  - MATINFO/Dell profite bien !

- Qu'est ce qui a orienté ces choix ?
  - Le besoin grandissant en puissance de calcul.
  - Les contraintes d'intégration et de fonctionnement.
  - Le porte monnaie.
  - L'évolution à marche forcée des constructeurs.

- Le calcul du TCO !



2009 – 2018 : Evolution de la puissance des serveurs

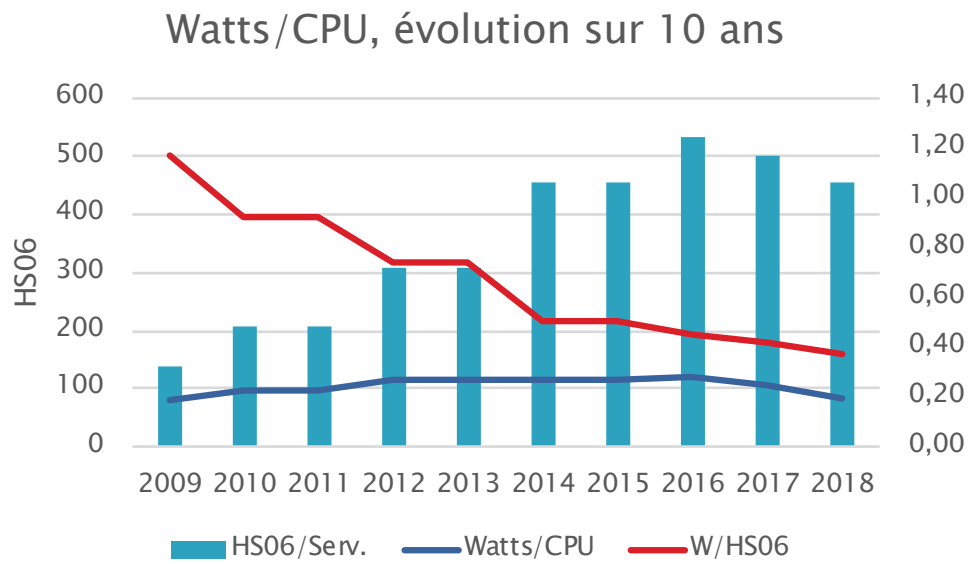


La puissance par thread augmente de 33,8% en 10 ans.

Le nombre de threads a été multiplié par 3, la puissance d'un serveur par 4.

Serveur	Année	HS06	Thread	HS06/Thread
DELL M610	2009	136,67	16	8,54
DELL C6100	2010	206,92	24	8,62
DELL C6100	2011	206,92	24	8,62
DELL C6220v1	2012	309,51	32	9,67
DELL C6220v1	2013	309,51	32	9,67
DELL C6220v2	2014	454,16	40	11,35
DELL C6220v2	2015	454,16	40	11,35
DELL C6320v1	2016	532,80	48	11,10
DELL C6320v2	2017	499,37	48	10,40
DELL C6420	2018	457,11	40	11,43

2009 – 2018 : Evolution de la consommation électrique

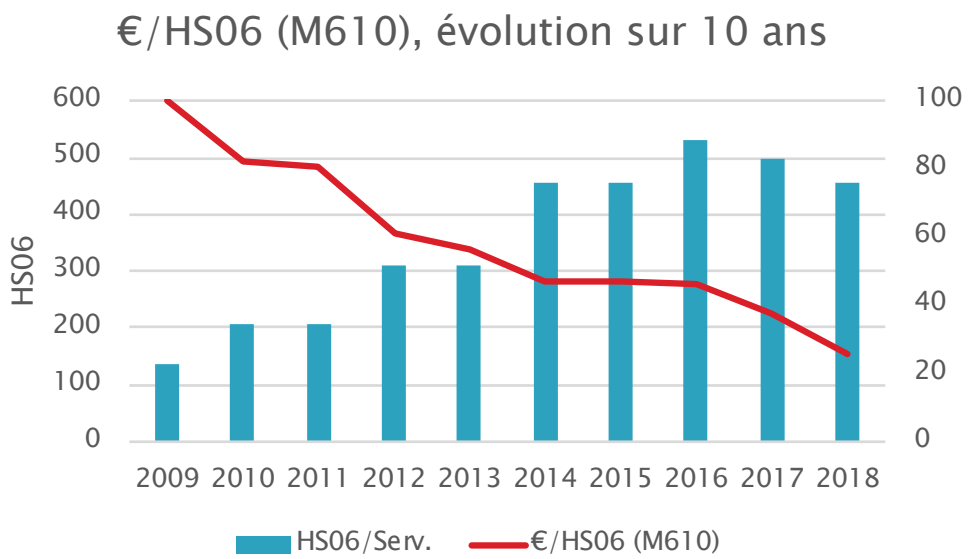


La consommation par HS06 a été divisée par 3,1.

Le calcul du TCO a orienté le choix sur des configurations dont la consommation reste « faible ».

Serveur	Année	HS06	W/CPU	W/HS06
DELL M610	2009	136,67	80	1,17
DELL C6100	2010	206,92	95	0,92
DELL C6100	2011	206,92	95	0,92
DELL C6220v1	2012	309,51	115	0,74
[illegible]	2013	309,51	115	0,74
[illegible]	[illegible]	454,16	115	0,51
DELL C6220v2	2015	[illegible]	115	0,51
DELL C6320v1	2016	532,80	120	0,45
DELL C6320v2	2017	499,37	105	0,42
DELL C6420	2018	457,11	85	0,37

2009 – 2018 : Evolution du coût d'investissement



Le coût par HS06 a été divisé par 4 en 10 ans.

Serveur	Année	HS06	Coût HS06 normalisé %(M610)
DELL M610	2009	136,67	100,00
DELL C6100	2010	206,92	82,49
DELL C6100	2011	206,92	80,62
DELL C6220v1	2012	309,51	61,25
DELL C6220v1	2013	309,51	56,04
DELL C6220v2	2014	454,16	46,95
DELL C6220v2	2015	454,16	46,95
DELL C6320v1	2016	532,80	45,82
DELL C6320v2	2017	499,37	37,51
DELL C6420	2018	457,11	25,72

2009 – 2018 : Configuration des serveurs de calcul

- Depuis 10 ans, nous utilisons pour le calcul HTC des serveurs dont l'architecture change peu.
  - 2 x CPU, n x cores/CPU, 2n x threads/CPU
  - 3 Go/thread
  - 1Gbps de bande passante réseau
- A base de processeurs Intel exclusivement !

Modèle	Première mise en service	Processeur	Fréq. (GHz)	Cache (MB)	Cons. (W)	Canaux mém.	Lithographie	CPU par serveur	Cores		Threads		Mém. (GB)	HS06/serv.	HS06/thread	W/HS06
									par CPU	par serv.	par CPU	par serv.				
DELL M610	2008	Intel E5540	2,53	8	80	3	45 nm	2	4	8	8	16	48,00	136,67	8,54	1,17
DELL C6100	2010	Intel X5650	2,66	12	95	3	32nm	2	6	12	12	24	72,00	206,92	8,62	0,92
DELL C6220v1	2012	Intel E5-2670	2,60	20	115	4	32 nm	2	8	16	16	32	96,00	309,51	9,67	0,74
DELL C6220v2	2014	Intel E5-2680v2	2,80	25	115	4	22 nm	2	10	20	20	40	128,00	454,16	11,35	0,51
DELL C6320v1	2015	Intel E5-2680v3	2,50	30	120	4	22 nm	2	12	24	24	48	144,00	532,80	11,10	0,45
DELL C6320v2	2017	Intel E5-2650v4	2,20	30	105	4	14 nm	2	12	24	24	48	144,00	499,37	10,40	0,42
DELL C6420	2017	Intel Xeon Silver 4114	2,20	14	85	6	14 nm	2	10	20	20	40	128,00	457,11	11,43	0,37

2009 – 2018 : Un premier bilan

- Pour notre discipline, tant que le modèle de calcul n'évolue pas, nous resterons sur des architectures « classiques ».
  - L'utilisation de GPGPU nécessite d'adapter le modèle de calcul.
  - Quelques rares tentatives ont démontré un gain mais…
- Si les évolutions majeures, il y a bien longtemps, se situaient sur le terrain de la vitesse des processeurs, l'augmentation de la puissance est liée depuis une décennie aux nombres de cœurs.
- Nous allons continuer à empiler les processeurs et les cœurs dans nos serveurs au rythme des nouveautés présentant le meilleur rapport puissance, conso, coûts.
  - Le choix d'utiliser des processeurs à 10 cœurs en 2018 n'est pas technologique mais financier.
  - Intel Xeon Platinium 8176 2.1GHz, 38M Cache, HT,28C/56T (165W).

Page de publicité

- Cascade Lake annoncé le 5 novembre 2018 serait un processeur 48 cœurs (96 threads ?), disponible début 2019.

https://www.zdnet.fr/actualites/intel-annonce-de-nouvelles-puces-xeon-le-cascade-lake-ap-et-le-xeon-e-2100-39876009.htm

Le retour de la concurrence ?

- Le fondeur AMD revient sur le marché des serveurs avec ses processeurs EPYC.
  - https://www.amd.com/system/files/2017-06/AMD-EPYC-Data-Sheet.pdf
  - AMD EPYC 7251 (2,1GHz/2,9GHz, 8C/16T, cache 32Mo, 120W)
  - AMD EPYC 7281 2.1GHz/2.7GHz, 16C/32T, 32M Cache (155W/170W)
  - AMD EPYC 7451 2.3GHz/2.9GHz, 24C/48T, 64M Cache (180W)
  - AMD EPYC 7551 2.00GHz/2.55GHz, 32C/64T, 64M Cache 180W)
  - AMD EPYC 7601 2.20GHz/2.7GHz, 32C/64T, 64M Cache (180W)
- Uniquement disponible sur Dell PowerEdge R7425, mais bien présent chez la concurrence.
  - HPE Apollo 2000, Supermicro
- Evaluation des processeurs EPYC prévue au CC-IN2P3.

AMD chez DELL, le futur ?

- Source interne Dell

« Des solutions HPC de type C6420 avec des processeurs AMD ROME (successeur de EPYC) sont effectivement en cours de développement. Pas encore d'information sur les caractéristiques techniques de ces nouvelles plateformes, car les processeurs ROME sont prévus pour mi-juillet 2019 pour les 1 socket et octobre 2019 pour les 2 sockets. »

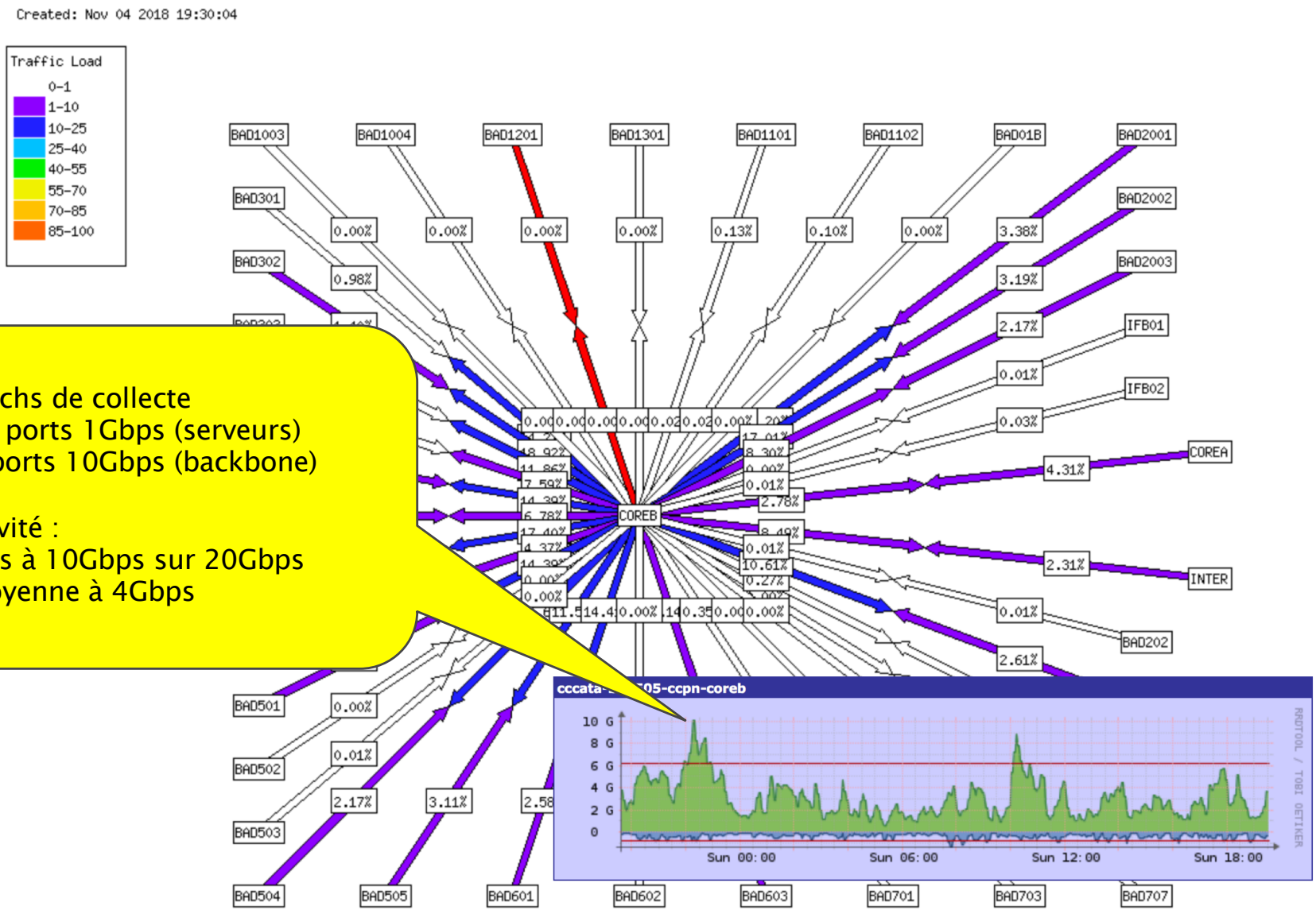
2009 – 2018 : Interconnexion réseau

- Pas de demande spécifique sur la bande passante à fournir aux serveurs de calcul. Dimensionnement découle de l'observation de la production.

- Switchs de collecte
  - M610 : 10Gbps pour 16 serveurs connectés en 1Gbps.
  - C6x00 : 10Gbps pour 48 serveurs connectés en 1Gbps.
  - 20Gbps / 48 serveurs introduit en 2016 et généralisé sur tous les switchs de collecte en 2018.
  - Tests en 2018 de serveurs à 10Gbps directement sur le backbone.

- En 2018,
  - 400Mbps de bande passante garantie par serveur de calcul.
  - Maximum possible à 1Gbps.
  - Dimensionnement des services de stockage ?

2018 : Switchs de collecte et bande passante réseau



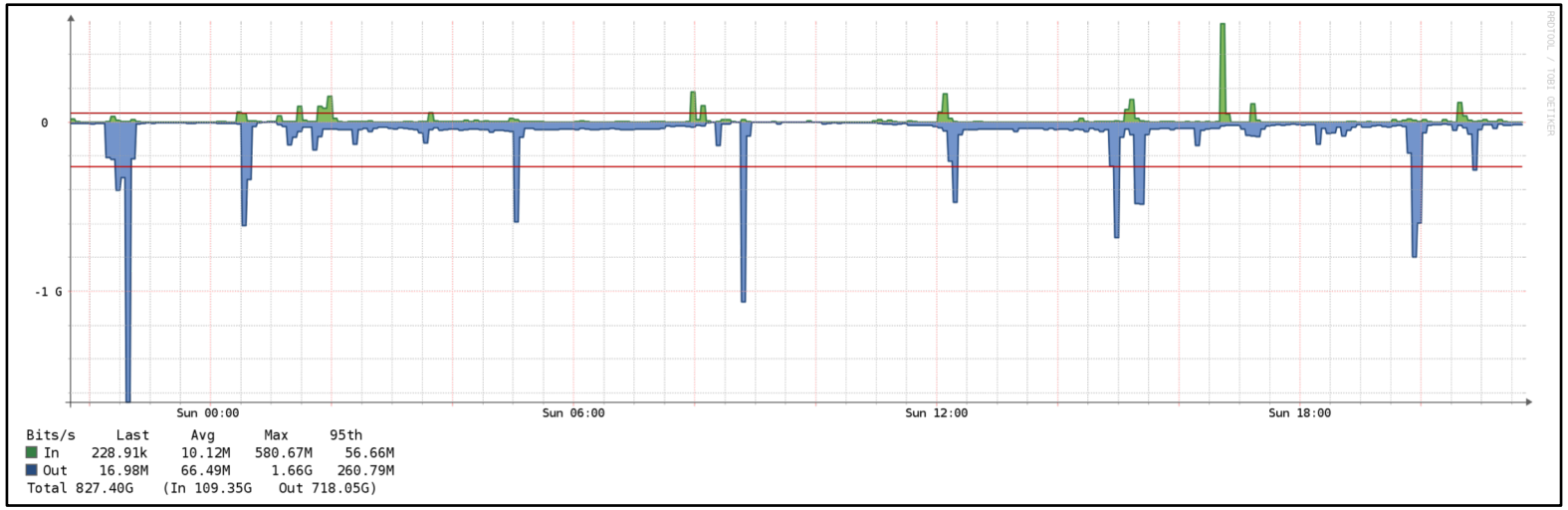
Switchs de collecte
. 48 ports 1Gbps (serveurs)
. 2 ports 10Gbps (backbone)

Activité :
. Pics à 10Gbps sur 20Gbps
. Moyenne à 4Gbps

2018 : Serveurs 10Gbps et utilisation de la bande passante

- 10 serveurs de calcul en 10Gbps depuis le mois de mai 2018 pour étude.

- Cette expérimentation ne fait pas apparaître à ce jour une contention réseau et donc un besoin d'augmenter la bande passante par serveur.



2018 : Plateforme de calcul, quelques règles

- Attention au dimensionnement des I/O sur les disques locaux.
  - 1 disque SATA 7200tpm pour 20 threads.
  - Actuellement 2 disques par serveur 40 threads.
  - Jusqu' 3 disques sur les serveurs 48 threads.

- Attention au peuplement des canaux mémoires pour ne pas dégrader les performances.
  - En pratique, des tests réalisés (bench HS06), sur un peuplement asymétrique des canaux, n'a pas permis de mettre en évidence une dégradation.

Stockage sur disques (DAS)

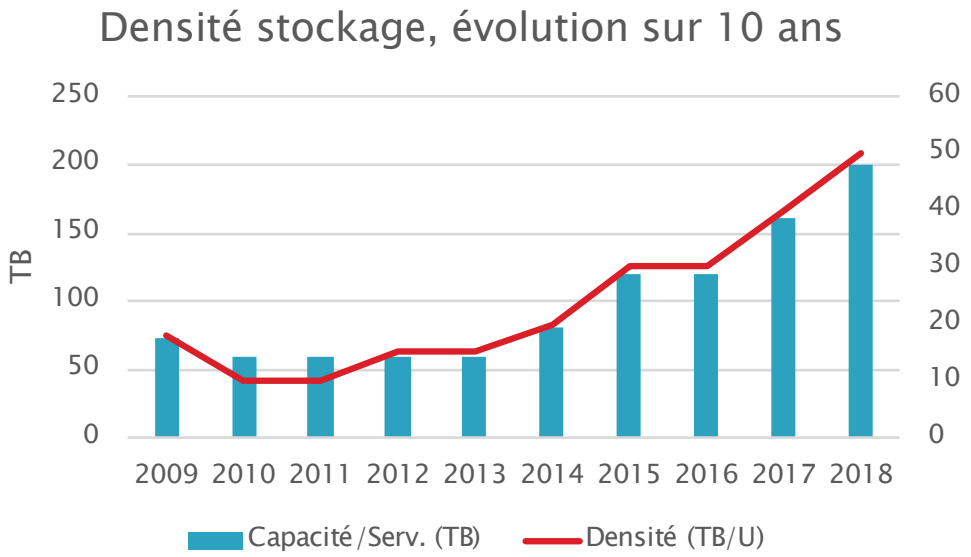
2009 – 2018 : 10 d'évolution des moyens de stockage sur disques

- Serveurs de stockage (DAS).
  - Serveur disposant d'une capacité de stockage sécurisée, performante et à faible coût.
  - Utilisés principalement pour EOS, DCACHE, DPM et XROOTD.

- Plusieurs générations de serveurs depuis 10 ans.
  - SUN Fire x4540, Dell PowerEdge R510, R720xd, R730xd, R740xd.
  - Effet MATINFO encore !

- Qu'est ce qui a orienté ces choix ? (bis)
  - L'augmentation des besoins de stockage.
  - Les contraintes d'intégration et de fonctionnement.
  - Le porte monnaie.

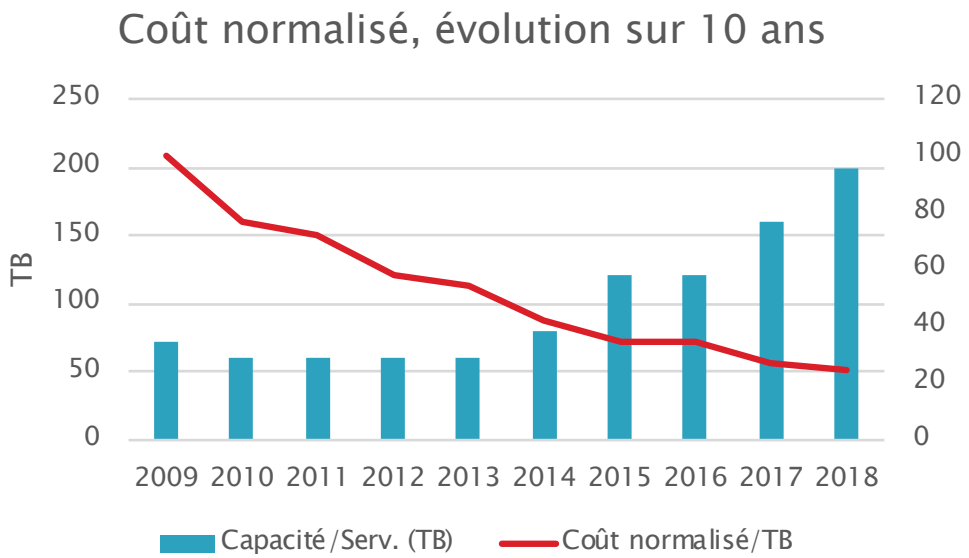
2009 – 2018 : Evolution de la densité du stockage



La capacité des disques a été multipliée par 5 en 10 ans, la densité (capacité) des serveurs par 3,1.

Année	Serveur	Extension		Disques		Réseau (Gbps)	Capacité Serv. (TB)	Encombrement (U)	Densité (TB/U)
		Modèle	Nombre	Capacité (TB)	Nombre				
2009	x4540	-	-	2	48	2	64	4	16
2010	R510	MD1200	2	2	36	2	60	6	10
2011	R510	MD1200	2	2	36	2	60	6	10
2012	R510	MD1200	1	3	24	2	60	4	15
2013	R720xd	MD1200	1	3	24	2	60	4	15
2014	R720xd	MD1200	1	4	24	2	80	4	20
2015	R730xd	MD1400	1	6	24	10	120	4	30
2016	R730xd	MD1400	1	6	24	10	120	4	30
2017	R730xd	MD1400	1	8	24	10	160	4	40
2018	R740xd	MD1400	1	10	24	10	200	4	50

2009 – 2018 : Evolution du coût d'investissement



Le coût par TB a été divisé par 4 en 10 ans.

Année	Serveur	Coût TB normalisé %(x4540)
2009	x4540	100,00
2010	R510	76,37
2011	R510	72,17
2012	R510	58,16
2013	R720xd	53,93
2014	R720xd	41,64
2015	R730xd	34,24
2016	R730xd	34,35
2017	R730xd	26,61
2018	R740xd	24,64

2009 – 2018 : Densification du stockage

- Recherche de la densification des configurations au cours de années.
  - Essentiellement par évolution de la capacité des disques.
  - Par augmentation du nombre de disques par serveur.
  - Par augmentation du nombre de disques par serveur dans un volume équivalent.

- Impact de la densification des serveurs.
  - Sur la bande passante par TB utile (25 à 80Mbps/TB).
  - Sur la capacité des cartes contrôleurs RAID (aujourd'hui non limitant).
  - Sur la quantité de mémoire par serveur (pas vraiment de règle).
  - Sur la disponibilité des données.

- 2 incidents avec perte de données en 2018 au CC-IN2P3.
  - Cartes contrôleurs RAID et disques défectueux.
  - Mauvaise série ?
  - Matériel inadapté à nos exigences, trop « cheap » ?

2018 : Questionnement

- Jusqu'à quel niveau de densification devons aller ?
  - A ce jour 10 serveurs, 2PB dans un rack pour une faible consommation électrique.
- Réflexion à avoir sur la sécurisation des données et la disponibilité des accès.
- Problématique globale de migration/transfert des données.
- Comment maintenir un niveau de coût faible, mais en augmentant le niveau de disponibilité des serveurs ?
  - Configuration à remettre en cause ? RAID matériel ? RAID logiciel ? Sécurisation « out of the box » ?

Page de publicité #1 : HPE Apollo 4200

- Serveur HPE Apollo 4200
  - 12 + 12 disques 3,5 pouces en façade avant.
  - 2 disques en façade arrière.
  - RAID matériel, contrôleur HPE.
  - Densité double d'un serveur Dell PowerEdge R740XD.

https://h20195.www2.hpe.com/v2/GetPDF.aspx/c04616497.pdf

Page de publicité #2

- Serveur HPE Apollo 4510
  - 30 + 30 disques 3,5 pouces en façade avant.
  - 2 disques 2,5 pouces.
  - RAID matériel, contrôleur HPE.
  - Densité x 2.5 d'un serveur Dell PowerEdge R740XD.

- Configuration actuellement en test au CC-IN2P3.

https://www.hpe.com/fr/fr/product-catalog/servers/proliant-servers/pip.hpe-apollo-4510-system.1010193037.html

Page de publicité #3 : Dell PowerEdge R740XD2

- Serveur Dell PowerEdge R740XD2
  - Plusieurs options de châssis seront disponibles :
    - 24x 3.5” : 12 + 12 single PERC
    - 26x 3.5” : 12 + 12 + 2 (flexBay) single PERC
    - 26x 3.5” : PERC1 : 12 + 12 + PERC2 : 2 (flexBay)
  - Densité x 2.5 d’un serveur Dell PowerEdge R740XD.

- Devis possible à partir du 21 novembre, livraison à partir du 5 décembre.

- Evaluation du matériel à venir.

Stockage sur bandes magnétiques

- Voir la présentation de Pierre-Emmanuel BRINETTE aux journées informatique 2018.
  - https://indico.in2p3.fr/event/17206/contributions/64147/attachments/50000/63764/20181002_-_JI_2018_-_Tape_Saves.pdf

Merci