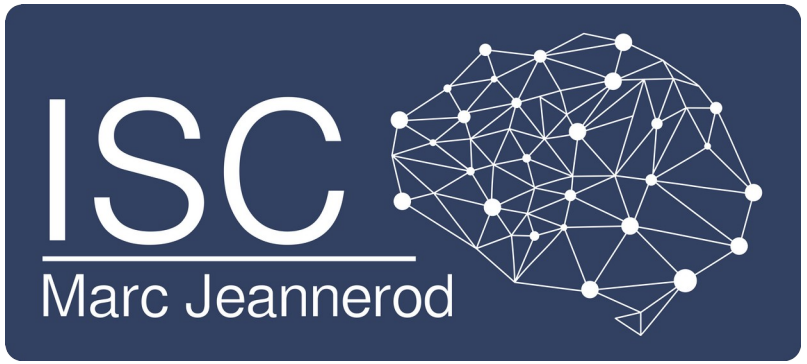
Stockage

Plus de post-scriptum dans mes inscriptions aux ANF ?

[…]
Ayant vécu sur la liste ASR des moments intenses de surprise à la lecture des choix de collègues sur des RAID ou des médias de stockage HDD/SSD, je ferai volontiers une intervention
[...]

Déjà en 2016…

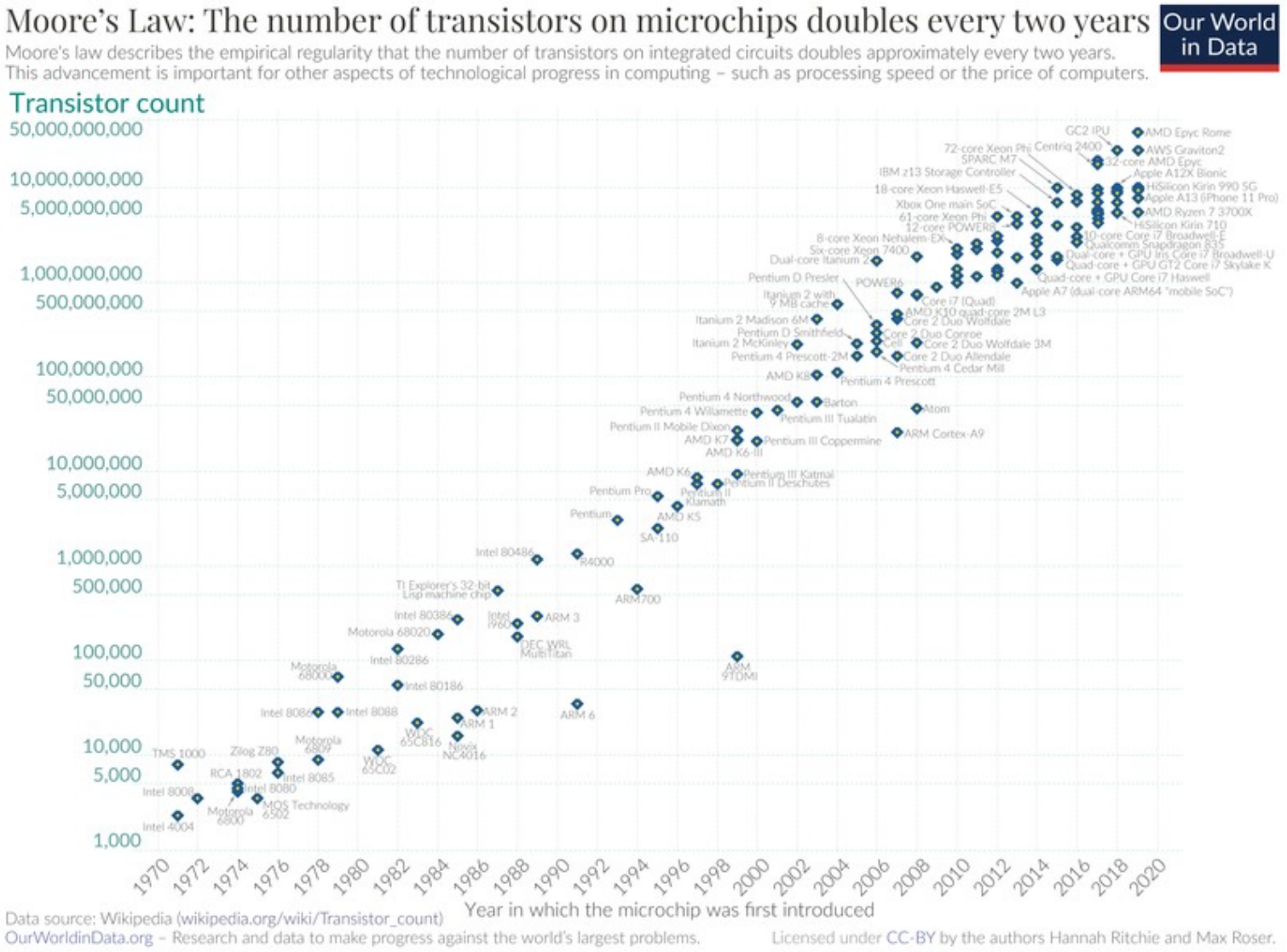
Qui suis-je ?



Histoire d'un labo

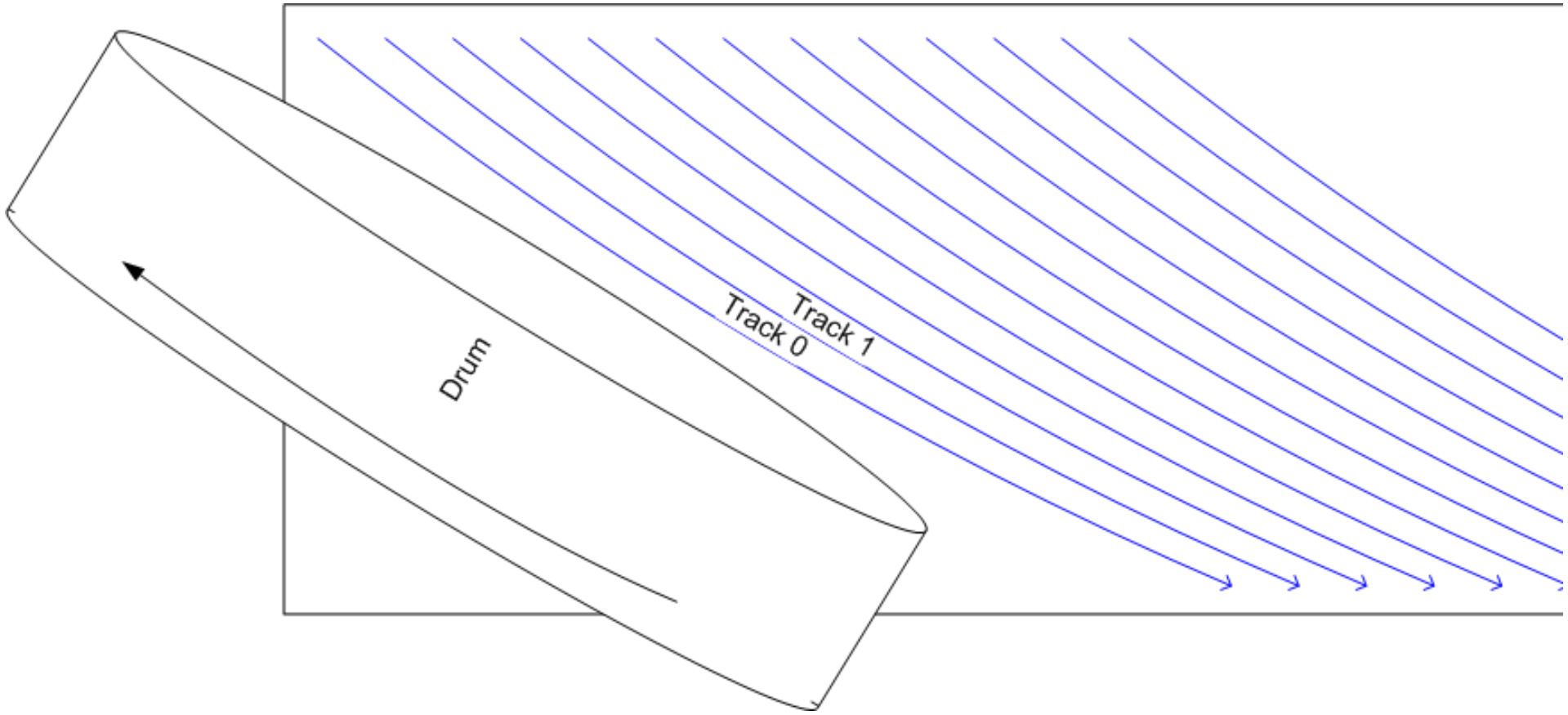
– 2004	100Go	3 k€/an	IRM,EEG 50Hz
– 2008	1To	5 k€/an	MEG
– 2012	10To	10 k€/an	EEG 40kHz
– 2016	100To	20 k€/an	IRM HD
– 2020	1Po	40 k€/an	Microscopie 3D
– Q1.22	4Po	65 k€/an	EEG+Vidéo FreeMotion

Stockages en laboratoire ?

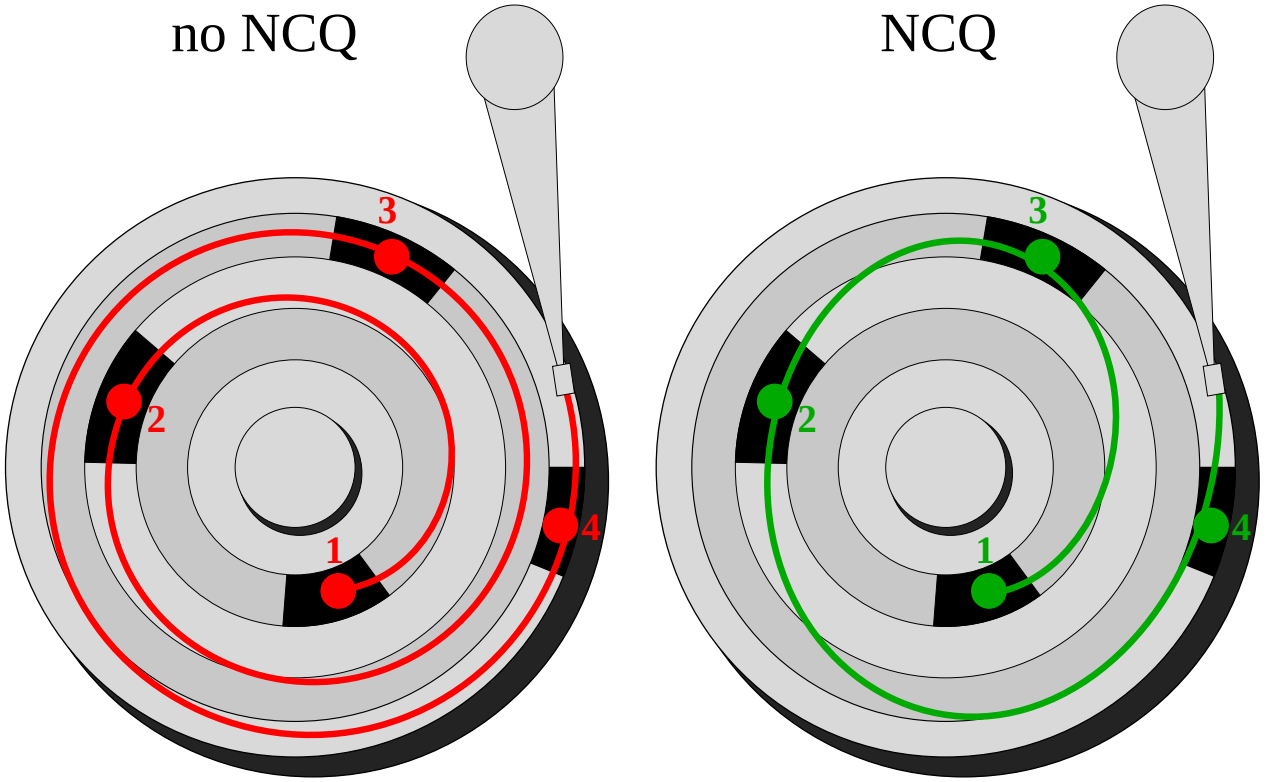


Le support magnétique

Camembert rotatif

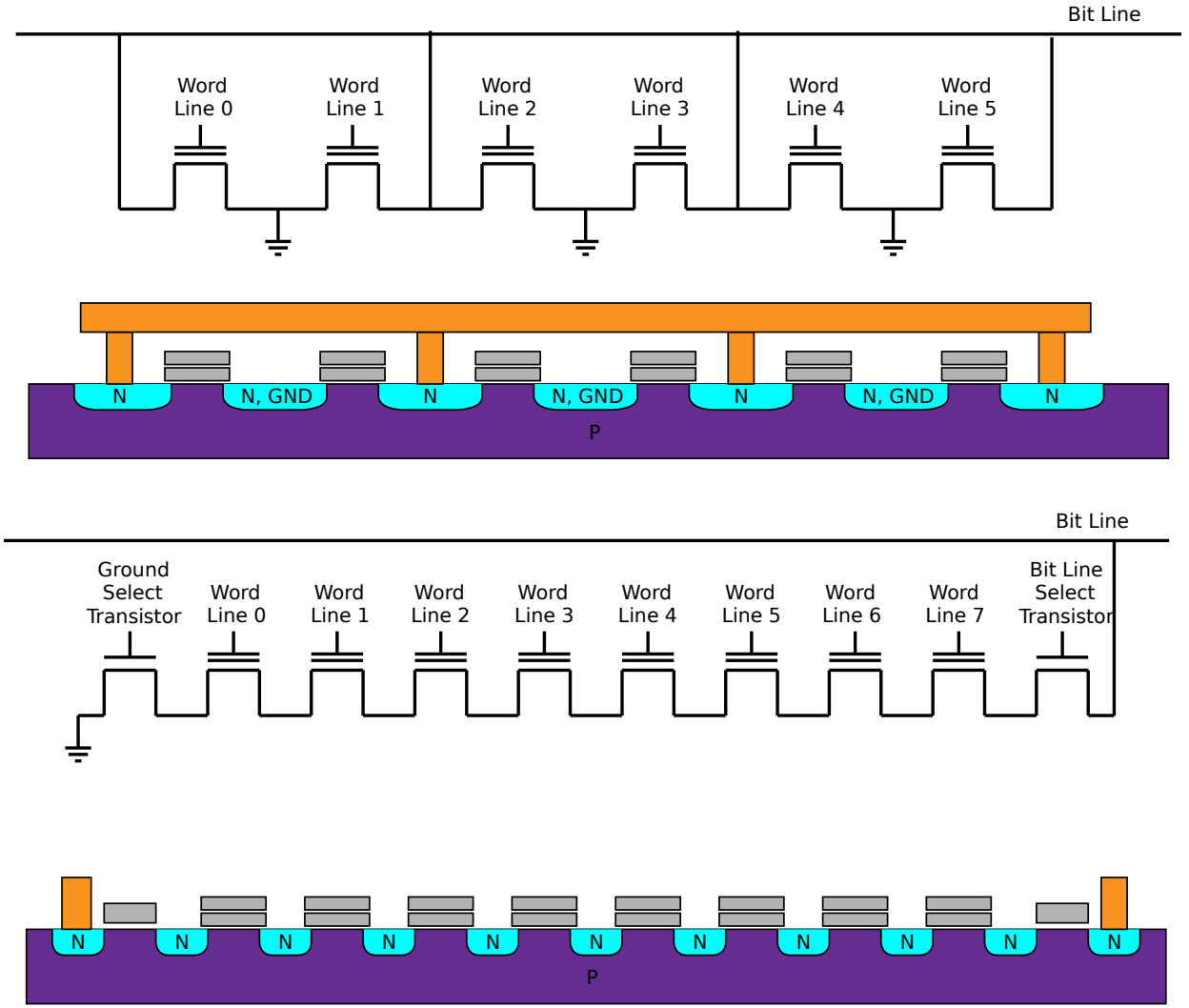


La gigue du bras



Non-volatile random-access memory (NVRAM) is random-access memory that retains its information when power is turned off (non-volatile)

Le match des années 2000



10000 écritures et moins ?

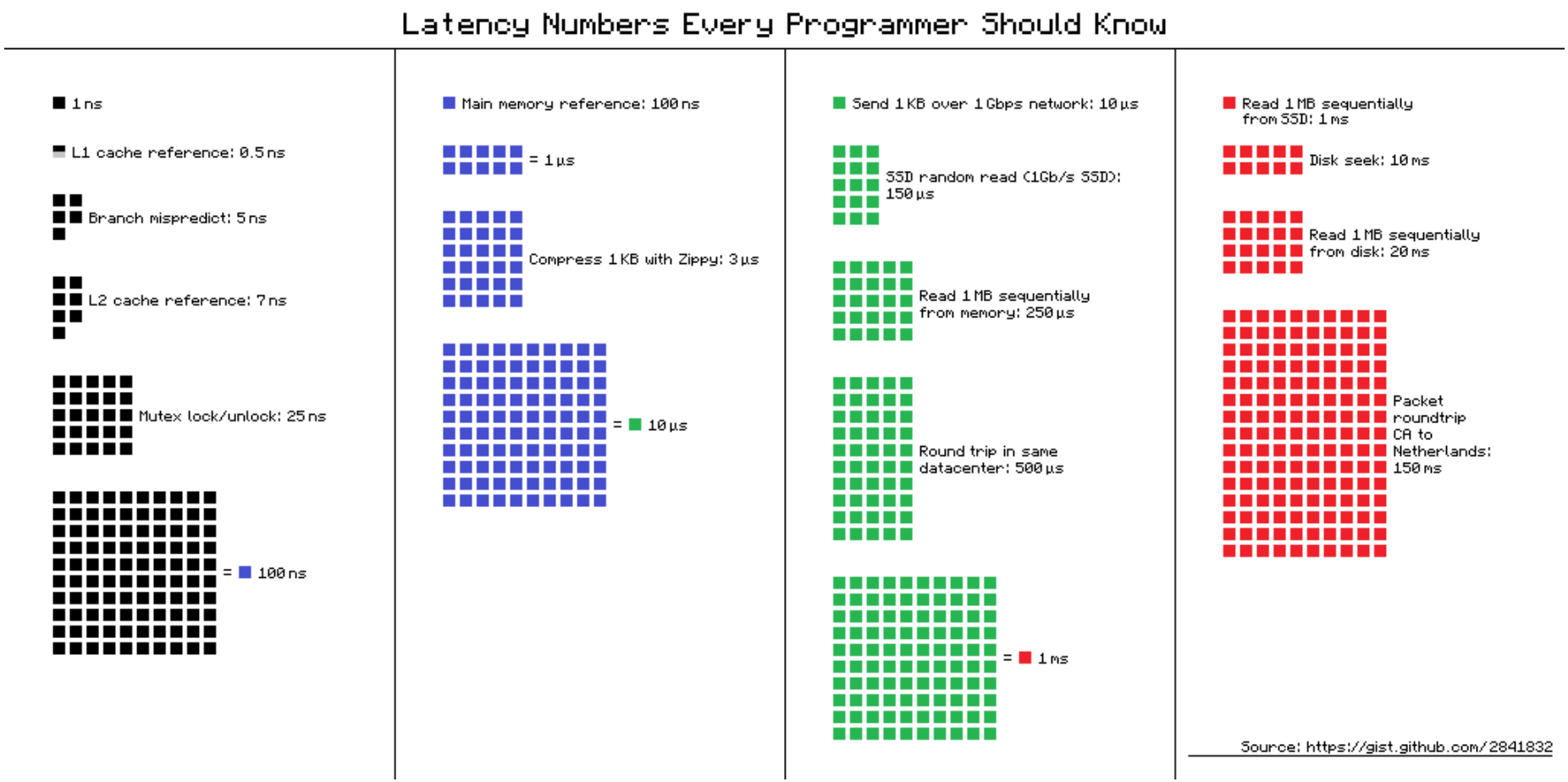
Pour modifier 1 bit, en écrire 512 ?

La répartition de l'usure.

Les caractéristiques

L'entrée/sortie

Et l'art d'attendre



https://people.eecs.berkeley.edu/~rcs/research/interactive_latency.html

Les caractéristiques

La bande passante

Les caractéristiques

La durée de vie

Les caractéristiques

Le taux d'erreur

Les caractéristiques

Le taux de panne

https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

Constructeur :

- AFR 0,35 %
- 1 secteur faux tous les 10^{15}bits lus
- Bande passante ~280Mo.s^{-1}
- Latence moyenne 4.16ms

https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

Publicités mensongères ?

Backblaze Hard Drive Failure Rates for 2021

Reporting period 1/1/2021 thru 12/31/2021 inclusive. For drives models in service as of 12/31/2021.

MFG	Model	Drive Size	Drive Count	Avg. Age (months)	Drive Days	Drive Failures	AFR
HGST	HMS5C4040ALE640	4TB	3,429	66.92	1,188,017	19	0.58%
HGST	HMS5C4040BLE640	4TB	12,703	62.37	4,647,157	39	0.31%
Seagate	ST4000DM000	4TB	18,611	74.37	6,856,981	339	1.80%
Toshiba	MD04ABA400V	4TB	97	79.32	35,781	2	2.04%
Seagate	ST6000DX000	6TB	886	80.85	323,390	1	0.11%
HGST	HUH728080ALE600	8TB	1,124	44.85	397,463	7	0.64%
Seagate	ST8000DM002	8TB	9,718	62.63	3,554,465	142	1.46%
Seagate	ST8000NM0055	8TB	14,334	52.82	5,253,943	214	1.49%
Seagate	ST10000NM0086	10TB	1,192	50.07	436,951	27	2.26%
HGST	HUH721212ALE600	12TB	2,600	27.04	946,710	7	0.27%
HGST	HUH721212ALE604	12TB	13,138	9.40	3,305,589	26	0.29%
HGST	HUH721212ALN604	12TB	10,818	32.95	3,951,844	52	0.48%
Seagate	ST12000NM0007	12TB	1,324	25.80	2,799,888	154	2.01%
Seagate	ST12000NM0008	12TB	20,201	21.13	7,340,502	218	1.08%
Seagate	ST12000NM001G	12TB	12,171	13.84	3,770,446	54	0.52%
Seagate	ST14000NM001G	14TB	10,738	11.10	3,054,188	86	1.03%
Seagate	ST14000NM0138	14TB	1,611	12.86	586,327	77	4.79%
Toshiba	MG07ACA14TA	14TB	38,214	14.28	11,617,844	245	0.77%
Toshiba	MG07ACA14TEY	14TB	462	11.81	153,659	7	1.66%
WDC	WUH721414ALE6L4	14TB	8,408	12.81	2,951,046	35	0.43%
Seagate	ST16000NM001G	16TB	10,861	7.74	1,606,863	49	1.11%
Toshiba	MG08ACA16TE	16TB	5,985	3.57	320,260	8	0.91%
Toshiba	MG08ACA16TEY	16TB	2,367	8.52	573,726	11	0.70%
WDC	WUH721816ALE6L0	16TB	1,767	5.06	256,533	1	0.14%
		Totals	**202,759**		**65,929,573**	**1,820**	**1.01%**

Backblaze

https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

HDD = 18To = 0.144×10^{15}

Un secteur en erreur 512o ? 4ko ?

=>1 HDD sur 7 porterait une erreur ?

Des gars ont étudié le problème…

Google Scholar :

hard drive disk uncorrectable error bit rate

https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

$P=\pi D$

Si ~280Mo.s^{-1} à Dmax 3.5”
alors 80Mo.s^{-1} à Dmin 1” ?

Bande passante moyenne ?

https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

Accès séquentiels ?

f(Hz)=1/t(s)

7200 Rotation Par Minutes
120 passages par seconde d'un secteur sous 1 tête de lecture

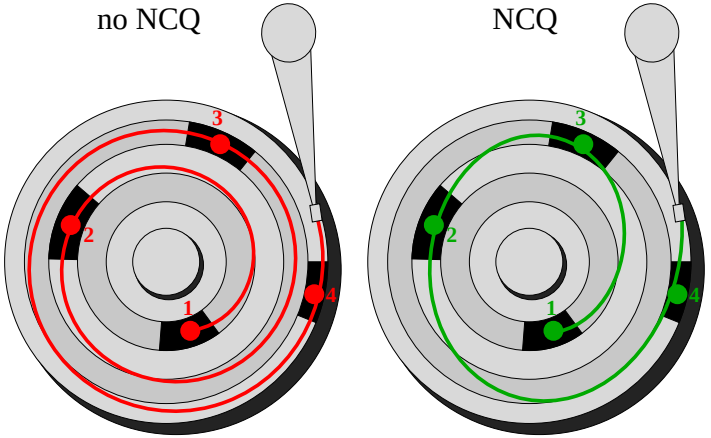
https://www.seagate.com/files/www-content/datasheets/pdfs/exos-x20-channel-DS2080-2111FR-fr_FR.pdf

Caractéristiques	SATA 6 Gbits/s
Capacité	20 To
Modèle FastFormat™ standard (512e/4Kn)[1]	ST20000NM007D
Modèle FastFormat avec autochiffrement (512e/4Kn)[1,2]	ST20000NM000D
Modèle FastFormat avec autochiffrement FIPS (512e/4Kn)[1,2]	—
Boîtier à l'hélium scellé	Oui
Conventional Magnetic Recording (CMR, enregistrement magnétique conventionnel)	Oui
Protection des informations (DIF T10)	—
Super parité	Oui
Faible teneur en halogène	Oui
Technologie PowerChoice™ d'optimisation de la consommation au repos	Oui
Technologie PowerBalance™ pour un meilleur équilibre entre performances et consommation énergétique	Oui
Hot-Plug Support[3]	Yes
Cache, multisegment (Mo)	256
Finition OSP (Organic Solderability Preservative)	Oui
Vérification du firmware RSA 3072 (SD&D)	Oui
Temps moyen entre deux pannes (MTBF, heures)	2 500 000
Taux de panne annualisé pour un fonctionnement 24h/24 et 7j/7 (AFR)	0,35 %
Erreurs de lecture irréparables par bit lu	1 secteur par 10E15
Nombre d'heures de fonctionnement par an (24h/24 et 7j/7)	8 760
Taille des secteurs 521e (octets par secteur)	512
Taille des secteurs 4Kn (octets par secteur)	4 096
Garantie limitée (années)	5
Vitesse de rotation (tr/min)	7 200 tr/min
Vitesse d'accès interface (Gbits/s)	6,0, 3,0
Capacité de transfert continu max. (diamètre extérieur) (Mo/s, Mio/s)	285/272
Lecture/écriture aléatoire 4K QD16 WCD (IOPS)	168/550
Latence moyenne (ms)	4,16
Ports d'interface	Simple
Vibrations rotationnelles à 20-1 500 Hz (rad/s²)	12,5

Accès séquentiels ?

1 secteur 512o/4ko

$120io.s^{-1} \times 0.5ko.io^{-1} = 60ko.s^{-1}$



Redondant
Array of
I nexpensive
Disks

A Case for Redundant Arrays of InexpensiveDisks (RAID)

D.A.Patterson,

G.Gibson,

et

R.H.Katz,

1988

3 A Solution: Arrays of Inexpensive Disks

Rapid improvements in capacity of large disks have not been the only target of disk designers, since personal computers have created a market for inexpensive magnetic disks These lower cost disks have lower performance as well as less capacity Table I below compares the top-of-the-line IBM 3380 model AK4 mainframe disk, Fujitsu M2361A "Super Eagle" minicomputer disk, and the Conner Peripherals CP 3100 personal computer disk

Characteristics	*IBM 3380*	*Fujitsu M2361A*	*Conners CP3100*	*3380 v 3100 (>1 means 3100 is better)*	*2361 v 3100*
Disk diameter (inches)	14	10 5	3 5	4	3
Formatted Data Capacity (MB)	7500	600	100	01	2
Price/MB(controller incl)	$18-$10	$20-$17	$10-$7	1-2 5	1 7-3
MTTF Rated (hours)	30,000	20,000	30,000	1	1 5
MTTF in practice (hours)	100,000	?	?	?	?
No Actuators	4	1	1	2	1
Maximum I/O's/second/Actuator	50	40	30	6	8
Typical I/O's/second/Actuator	30	24	20	7	8
Maximum I/O's/second/box	200	40	30	2	8
Typical I/O's/second/box	120	24	20	2	8
Transfer Rate (MB/sec)	3	2 5	1	3	4
Power/box (W)	6,600	640	10	660	64
Volume (cu ft)	24	3 4	03	800	110

Table I *Comparison of IBM 3380 disk model AK4 for mainframe computers, the Fujitsu M2361A "Super Eagle" disk for minicomputers, and the Conners Peripherals CP 3100 disk for personal computers By "Maximum I/O's/second" we mean the maximum number of average seeks and average rotates for a single sector access Cost and reliability information on the 3380 comes from widespread experience [IBM 87] [Gawlick87] and the information on the Fujitsu from the manual [Fujitsu 87], while some numbers on the new CP3100 are based on speculation The price per megabyte is given as a range to allow for different prices for volume discount and different mark-up practices of the vendors (The 8 watt maximum power of the CP3100 was increased to 10 watts to allow for the inefficiency of an external power supply, since the other drives contain their own power supplies)*

Synchronisme des écritures atomiques

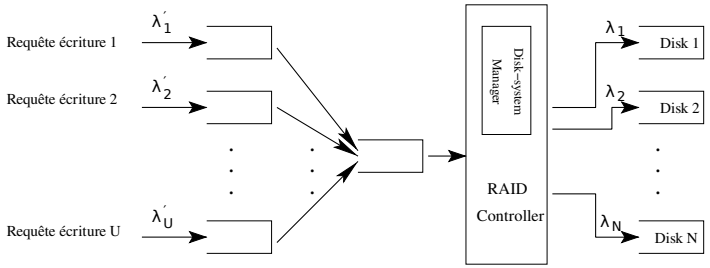


Fig. 1. Requests flow in a RAID storage system

Synchronisme des écritures atomiques

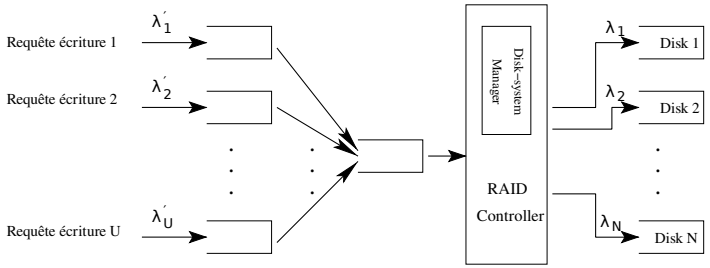
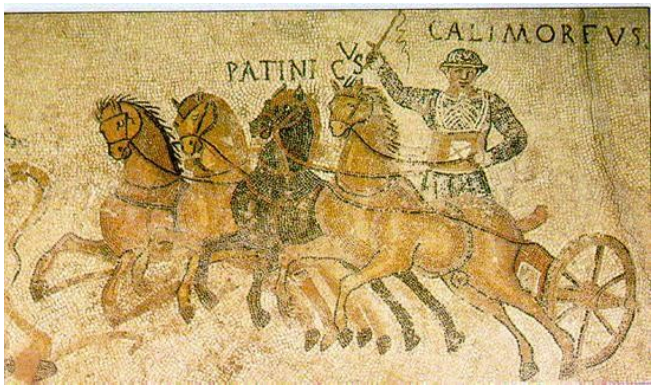


Fig. 1. Requests flow in a RAID storage system



Synchronisme des écritures atomiques

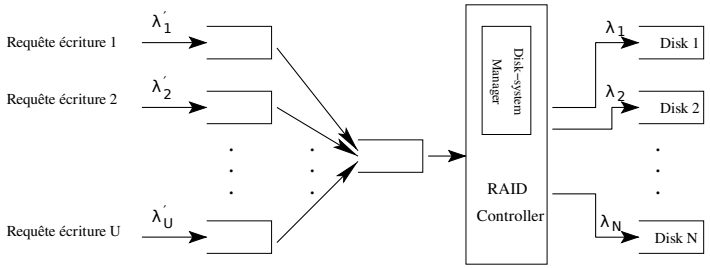
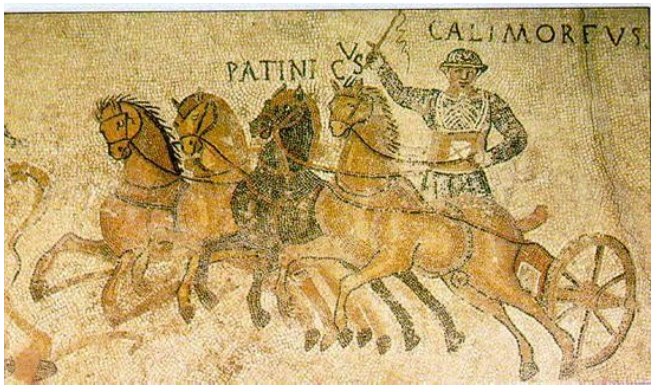
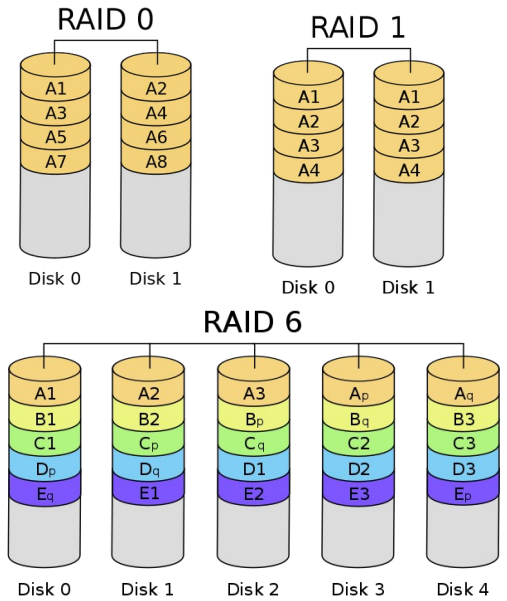


Fig. 1. Requests flow in a RAID storage system





Synchronisme des écritures atomiques

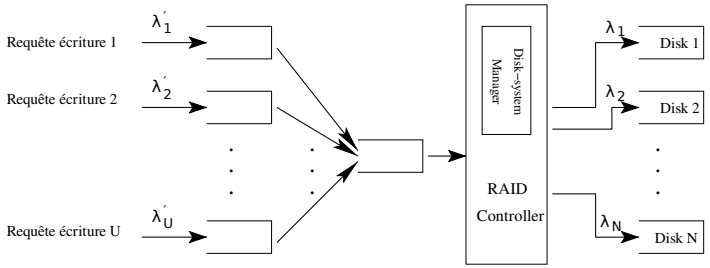
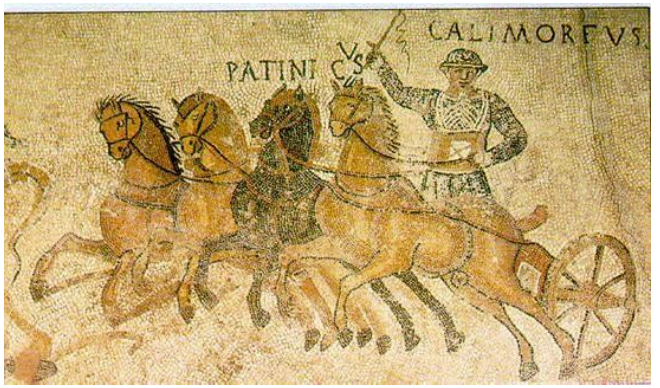
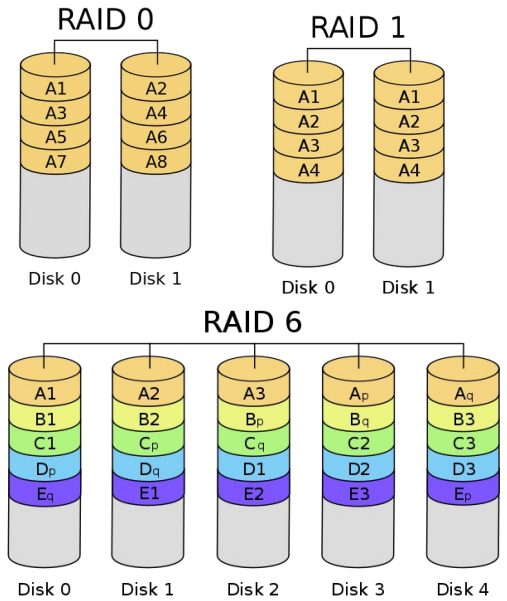


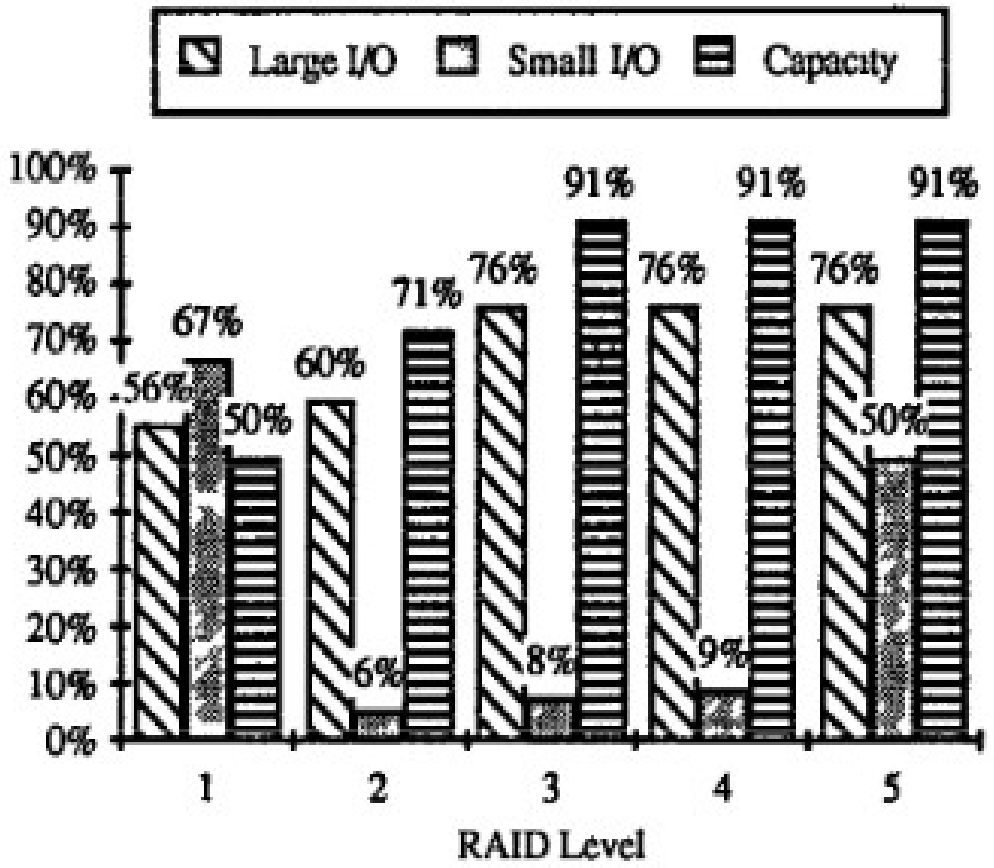
Fig. 1. Requests flow in a RAID storage system





Synchronisme des écritures atomiques

Dans un relais, c’est le dernier porteur du bâton qui franchit la ligne d’arrivée.


Large I/O
Small I/O
Capacity
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
56%
67%
50%
60%
6%
71%
76%
8%
91%
76%
9%
91%
76%
50%
91%
1
2
3
4
5
RAID Level

2/3/4 Mort-nés

	Configuration #1	Configuration #2
RAID Type:	RAID 10	RAID 6
Drive Capacity (GB):	8000	8000
Single drive performance: (●IO/s ○MB/s)	120	120
Single drive cost:	500	500
Number of drives per RAID group:	4	8
Number of RAID groups:	2	1
Read operations (%):	100	100

Calculate

Results:	**Configuration #1**	**Configuration #2**
Total performance (IO/s):	**960**	**960**
Total usable storage capacity (TB)	**32.00**	**48.00**
RAID type:	RAID 10	RAID 6
Reads / Writes (%):	100 / 0	100 / 0
Number of RAID groups:	2	1
Number of drives per RAID group:	4	8
Total number of drives:	8	8
Single RAID group performance (IO/s):	480	960
Capacity of a single RAID group (GB):	16000	48000
Single drive cost:	500	500
Cost per TB usable:	125.00	83.33
Total cost:	**4000.00**	**4000.00**

	Configuration #1	Configuration #2
RAID Type:	RAID 10	RAID 6
Drive Capacity (GB):	8000	8000
Single drive performance: IO/s MB/s	120	120
Single drive cost:	500	500
Number of drives per RAID group:	4	8
Number of RAID groups:	2	1
Read operations (%):	0	0

Calculate

Results:	**Configuration #1**	**Configuration #2**
Total performance (IO/s):	**480**	**160**
Total usable storage capacity (TB)	**32.00**	**48.00**
RAID type:	RAID 10	RAID 6
Reads / Writes (%):	0 / 100	0 / 100
Number of RAID groups:	2	1
Number of drives per RAID group:	4	8
Total number of drives:	8	8
Single RAID group performance (IO/s):	240	160
Capacity of a single RAID group (GB):	16000	48000
Single drive cost:	500	500
Cost per TB usable:	125.00	83.33
Total cost:	**4000.00**	**4000.00**

https://wintelguy.com/raidperf2.pl

RAID Performance Calculator

Compare two RAID configurations:

	Configuration #1	Configuration #2
RAID Type:	RAID 6	RAID 6
Drive Capacity (GB):	8000	8000
Single drive performance: ◉ **IO/s** ○ **MB/s**	170	170
Single drive cost:	410	410
Number of drives per RAID group:	5	5
Number of RAID groups:	12	12
Read operations (%):	0	100

Calculate

Results:	**Configuration #1**	**Configuration #2**
Total performance (IO/s):	**1700**	**10200**
Total usable storage capacity (TB)	**288.00**	**288.00**
RAID type:	RAID 6	RAID 6
Reads / Writes (%):	0 / 100	100 / 0
Number of RAID groups:	12	12
Number of drives per RAID group:	5	5
Total number of drives:	60	60
Single RAID group performance (IO/s):	141.67	850
Capacity of a single RAID group (GB):	24000	24000
Single drive cost:	410	410
Cost per TB usable:	85.42	85.42
Total cost:	**24600.00**	**24600.00**

With advantages in cost-performance, reliability, power consumption, and modular growth, we expect RAIDs to replace SLEDs in future I/O systems There are, however, several open issues that may bare on the practicality of RAIDs

- *What is the impact of a RAID on latency?*
- *What is the impact on MTTF calculations of non-exponential failure assumptions for individual disks?*
- *What will be the real lifetime of a RAID vs calculated MTTF using the independent failure model?*
- *How would synchronized disks affect level 4 and 5 RAID performance?*
- *How does "slowdown" S actually behave? [Livny 87]*
- *How do defective sectors affect RAID?*
- *How do you schedule I/O to level 5 RAIDs to maximize write parallelism?*
- *Is there locality of reference of disk accesses in transaction processing?*
- *Can information be automatically redistributed over 100 to 1000 disks to reduce contention?*
- *Will disk controller design limit RAID performance?*
- *How should 100 to 1000 disks be constructed and physically connected to the processor?*
- *What is the impact of cabling on cost, performance, and reliability?*
- *Where should a RAID be connected to a CPU so as not to limit performance? Memory bus? I/O bus? Cache?*
- *Can a file system allow differ striping policies for different files?*
- *What is the role of solid state disks and WORMs in a RAID?*
- *What is the impact on RAID of "parallel access" disks (access to every surface under the read/write head in parallel)?*

A Case for Redundant Arrays of Inexpensive Disks (RAID)
Patterson 88

Mean time to meaningless:

MTTDL, Markov models, and storage system reliability

Kevin M. Greenan

James S. Plank

Jay J. Wylie

6 Conclusions

We have argued that MTTDL is essentially a meaningless reliability metric for storage systems and that Markov models, the normal method of calculating MTTDL, is flawed. We are not the first to make this argument (see [2] and [8]) but hope to be the last. We believe NOMDL_t has the desirable features of a good reliability metric, namely that it is calculable, meaningful, understandable, and comparable, and we exhort researchers to exploit it for their future reliability measurements. Currently, we believe that Monte Carlo simulation is the best way to calculate NOMDL_t.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under grants CNS-0615221.

8 HFRS Availability

The High-Fidelity Reliability (HFR) Simulator is a command line tool written in Python and is available at

`http://users.soe.ucsc.edu/~kmgreen/.`

For example, AFR is used to characterize the reliability of hard disk drives.

The relationship between AFR and MTBF (in hours) is:[1]

$$AFR = 1 - exp(-8766/MTBF)$$

This equation assumes that the device or component is powered on for the full 8766 hours of a year, and gives the estimated fraction of an original sample of devices or components that will fail in one year, or, equivalently, 1 − AFR is the fraction of devices or components that will show no failures over a year. It is based on an exponential failure distribution (see failure rate for a full derivation). Note: Some manufacturers count a year as 8760 hours.[2]

This ratio can be approximated by, assuming a small AFR,

$$AFR = \frac{8766}{MTBF}$$

For example, a common specification for PATA and SATA drives may be an MTBF of 300,000 hours, giving an approximate theoretical 2.92% annualized failure rate i.e. a 2.92% chance that a given drive will fail during a year of use.

The AFR for a drive is derived from time-to-fail data from a reliability-demonstration test (RDT).[3]

AFR will increase towards and beyond the end of the service life of a device or component. Google's 2007 study found, based on a large field sample of drives, that actual AFRs for individual drives ranged from 1.7% for first year drives to over 8.6% for three-year-old drives.[4] A CMU 2007 study showed an estimated 3% mean AFR over 1–5 years based on replacement logs for a large sample of drives.[5]

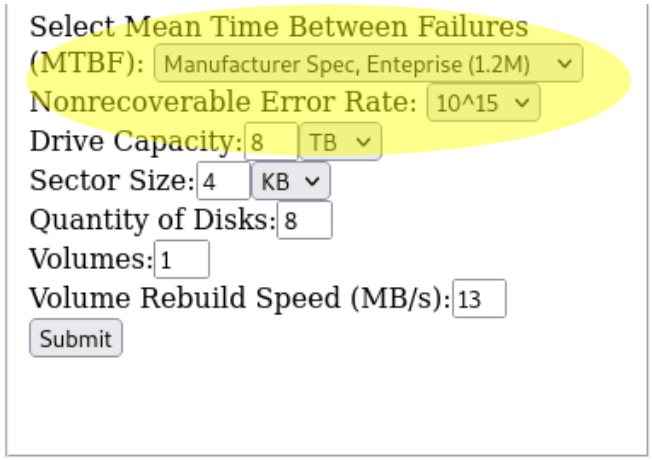
https://en.wikipedia.org/wiki/Annualized_failure_rate

Vitesse de reconstruction d’un volume :

Vitesse réelle de la reconstruction + le temps de remplacement du disque

Si le système est occupé à 90 % nous aurons une bande passante
Séquentielle d’au mieux 10 % de la bande passante du disque, avec
le Seagate de 20To à 180Mo.s^{-1} : 18Mo.s^{-1} !

Soit (20*10^12)/(18*10^6*3600) = 308heures
Si panne le vendredi matin et changement du disque mardi après midi :
100+308 heures pour reconstruire, soit 13Mo de bande passante effective
Pour la reconstruction.

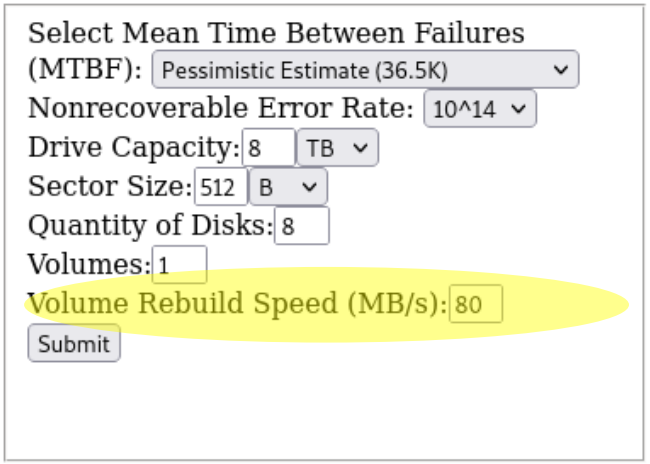


Valeurs du constructeur

RAID Level	Formatted Capacity (GB)	Mean Time To Data Failure (MTTDF) in hours	Bit Error Rate MTTDL	Mean Time To Data Loss (MTTDL) in hours	MTTDL (Years)
RAID 0	60,293.12	150,000.00	< .01	75,000.00	8.56
RAID 1	30,146.56	2,183,084,902.56	< .01	1,091,542,451.28	124,605.30
RAID 10	30,146.56	2,183,084,902.56	< .01	1,091,542,451.28	124,605.30
RAID 5	52,756.48	22,276,376.56	400,999.43	11,338,687.99	1,294.37
RAID 6	45,219.84	5,253,372,676.11	78,579,171.85	2,665,975,923.98	304,335.15
RAID-Z3	37,683.20	2,201,963,246,730.04	21,930,874,073,457.96	12,066,418,660,094.00	1,377,445,052.52
RAID 50	45,219.84	242,564,989.17	3,298,233.57	122,931,611.37	14,033.29
RAID 60	30,146.56	330,962,478,595.01	2,904,104,623.20	166,933,291,609.11	19,056,311.83

RAID Level	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years
RAID 0	11.02568 %	20.83571 %	29.56411 %	37.33015 %	44.23993 %	50.38786 %	55.85794 %	60.72490 %	65.05525 %	68.90815 %
RAID 1	0.00080 %	0.00161 %	0.00241 %	0.00321 %	0.00401 %	0.00482 %	0.00562 %	0.00642 %	0.00722 %	0.00803 %
RAID 10	0.00080 %	0.00161 %	0.00241 %	0.00321 %	0.00401 %	0.00482 %	0.00562 %	0.00642 %	0.00722 %	0.00803 %
RAID 5	0.07723 %	0.15440 %	0.23150 %	0.30855 %	0.38554 %	0.46247 %	0.53934 %	0.61616 %	0.69291 %	0.76960 %
RAID 6	0.00033 %	0.00066 %	0.00099 %	0.00131 %	0.00164 %	0.00197 %	0.00230 %	0.00263 %	0.00296 %	0.00329 %
RAID-Z3	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %	0.00000 %
RAID 50	0.00713 %	0.01425 %	0.02138 %	0.02850 %	0.03562 %	0.04275 %	0.04987 %	0.05699 %	0.06411 %	0.07123 %
RAID 60	0.00001 %	0.00001 %	0.00002 %	0.00002 %	0.00003 %	0.00003 %	0.00004 %	0.00004 %	0.00005 %	0.00005 %

https://www.servethehome.com/RAID/simpleMTTDL.php

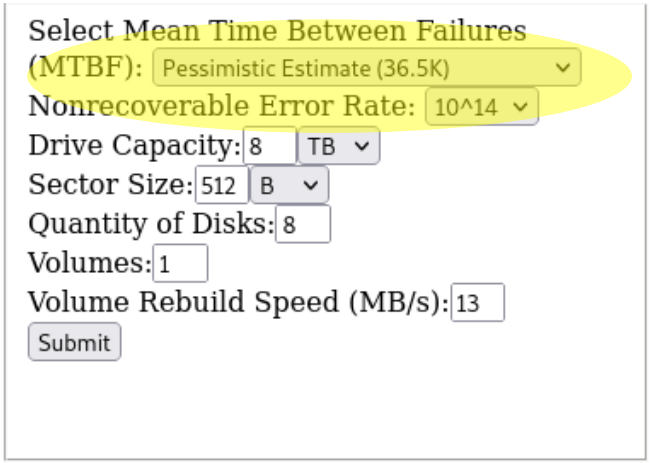


Avec un disque de secours

RAID Level	Formatted Capacity (GB)	Mean Time To Data Failure (MTTDF) in hours	Bit Error Rate MTTDL	Mean Time To Data Loss (MTTDL) in hours	MTTDL (Years)
RAID 0	60,293.12	4,562.50	< .01	2,281.25	0.26
RAID 1	30,146.56	12,429,123.34	< .01	6,214,561.67	709.42
RAID 10	30,146.56	12,429,123.34	< .01	6,214,561.67	709.42
RAID 5	52,756.48	126,827.79	4,565.59	65,696.69	7.50
RAID 6	45,219.84	5,598,431.07	148,250.69	2,873,340.88	328.01
RAID-Z3	37,683.20	439,234,498.74	1,086,069,924.60	762,652,211.67	87,060.75
RAID 50	45,219.84	1,381,013.70	19,086.77	700,050.24	79.91
RAID 60	30,146.56	352,701,157.50	2,365,646.30	177,533,401.90	20,266.37

RAID Level	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years
RAID 0	97.86383 %	99.95437 %	99.99903 %	99.99998 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %
RAID 1	0.14086 %	0.28152 %	0.42199 %	0.56225 %	0.70232 %	0.84219 %	0.98187 %	1.12135 %	1.26063 %	1.39971 %
RAID 10	0.14086 %	0.28152 %	0.42199 %	0.56225 %	0.70232 %	0.84219 %	0.98187 %	1.12135 %	1.26063 %	1.39971 %
RAID 5	12.48267 %	23.40717 %	32.96800 %	41.33538 %	48.65829 %	55.06710 %	60.67593 %	65.58462 %	69.88058 %	73.64029 %
RAID 6	0.30440 %	0.60788 %	0.91044 %	1.21207 %	1.51278 %	1.81258 %	2.11147 %	2.40945 %	2.70652 %	3.00268 %
RAID-Z3	0.00115 %	0.00230 %	0.00345 %	0.00459 %	0.00574 %	0.00689 %	0.00804 %	0.00919 %	0.01034 %	0.01149 %
RAID 50	1.24361 %	2.47175 %	3.68463 %	4.88241 %	6.06531 %	7.23349 %	8.38714 %	9.52645 %	10.65159 %	11.76273 %
RAID 60	0.00493 %	0.00987 %	0.01480 %	0.01974 %	0.02467 %	0.02960 %	0.03453 %	0.03947 %	0.04440 %	0.04933 %

https://www.servethehome.com/RAID/simpleMTTDL.php



Valeurs au pire

RAID Level	Formatted Capacity (GB)	Mean Time To Data Failure (MTTDF) in hours	Bit Error Rate MTTDL	Mean Time To Data Loss (MTTDL) in hours	MTTDL (Years)
RAID 0	60,293.12	4,562.50	< .01	2,281.25	0.26
RAID 1	30,146.56	2,019,732.54	< .01	1,009,866.27	115.28
RAID 10	30,146.56	2,019,732.54	< .01	1,009,866.27	115.28
RAID 5	52,756.48	20,609.52	4,565.59	12,587.55	1.44
RAID 6	45,219.84	147,833.57	24,090.74	85,962.15	9.81
RAID-Z3	37,683.20	1,884,762.10	176,486,362.75	89,185,562.42	10,181.00
RAID 50	45,219.84	224,414.73	19,086.77	121,750.75	13.90
RAID 60	30,146.56	9,313,514.94	384,417.52	4,848,966.23	553.53

RAID Level	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years
RAID 0	97.86383 %	99.95437 %	99.99903 %	99.99998 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %
RAID 1	0.86370 %	1.71994 %	2.56879 %	3.41030 %	4.24455 %	5.07159 %	5.89149 %	6.70431 %	7.51010 %	8.30894 %
RAID 10	0.86370 %	1.71994 %	2.56879 %	3.41030 %	4.24455 %	5.07159 %	5.89149 %	6.70431 %	7.51010 %	8.30894 %
RAID 5	50.06482 %	75.06478 %	87.54855 %	93.78235 %	96.89520 %	98.44961 %	99.22581 %	99.61341 %	99.80695 %	99.90360 %
RAID 6	9.69134 %	18.44345 %	26.34737 %	33.48530 %	39.93146 %	45.75291 %	51.01018 %	55.75795 %	60.04559 %	63.91771 %
RAID-Z3	0.00982 %	0.01964 %	0.02946 %	0.03928 %	0.04910 %	0.05892 %	0.06873 %	0.07855 %	0.08836 %	0.09817 %
RAID 50	6.94155 %	13.40126 %	19.41256 %	25.00658 %	30.21229 %	35.05664 %	39.56472 %	43.75987 %	47.66381 %	51.29675 %
RAID 60	0.18050 %	0.36067 %	0.54051 %	0.72003 %	0.89923 %	1.07810 %	1.25665 %	1.43488 %	1.61278 %	1.79037 %

https://www.servethehome.com/RAID/simpleMTTDL.php

Plusieurs volumes P=1-(1-PpanneVol)^NbVolume

Select Mean Time Between Failures (MTBF): Pessimistic Estimate (36.5K)
Nonrecoverable Error Rate: 10^14
Drive Capacity: 8 TB
Sector Size: 4 KB
Quantity of Disks: 8
Volumes: 7
Volume Rebuild Speed (MB/s): 13
Submit

RAID Level	Formatted Capacity (GB)	Mean Time To Data Failure (MTTDF) in hours	Bit Error Rate MTTDL	Mean Time To Data Loss (MTTDL) in hours	MTTDL (Years)
RAID 0	60,293.12	4,562.50	< .01	2,281.25	0.26
RAID 1	30,146.56	14,138,127.80	< .01	7,069,063.90	806.97
RAID 10	30,146.56	14,138,127.80	< .01	7,069,063.90	806.97
RAID 5	52,756.48	144,266.61	4,565.59	74,416.10	8.49
RAID 6	45,219.84	7,243,844.95	168,635.16	3,706,240.06	423.09
RAID-Z3	37,683.20	646,473,399.32	1,235,404,539.23	940,938,969.27	107,413.12
RAID 50	45,219.84	1,570,903.09	19,086.77	794,994.93	90.75
RAID 60	30,146.56	456,362,232.08	2,690,922.66	229,526,577.37	26,201.66

RAID Level	1 Year	2 Years	3 Years	4 Years	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years
RAID 0	97.86383 %	99.95437 %	99.99903 %	99.99998 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %	100.00000 %
RAID 1	0.12384 %	0.24753 %	0.37107 %	0.49445 %	0.61769 %	0.74076 %	0.86369 %	0.98646 %	1.10909 %	1.23156 %
RAID 10	0.12384 %	0.24753 %	0.37107 %	0.49445 %	0.61769 %	0.74076 %	0.86369 %	0.98646 %	1.10909 %	1.23156 %
RAID 5	11.11134 %	20.98806 %	29.76735 %	37.57114 %	44.50782 %	50.67375 %	56.15456 %	61.02637 %	65.35687 %	69.20618 %
RAID 6	0.23608 %	0.47160 %	0.70656 %	0.94097 %	1.17483 %	1.40813 %	1.64088 %	1.87309 %	2.10474 %	2.33585 %
RAID-Z3	0.00093 %	0.00186 %	0.00279 %	0.00372 %	0.00465 %	0.00559 %	0.00652 %	0.00745 %	0.00838 %	0.00931 %
RAID 50	1.09588 %	2.17975 %	3.25174 %	4.31199 %	5.36061 %	6.39774 %	7.42351 %	8.43804 %	9.44145 %	10.43386 %
RAID 60	0.00382 %	0.00763 %	0.01145 %	0.01527 %	0.01908 %	0.02290 %	0.02671 %	0.03053 %	0.03434 %	0.03816 %

https://www.servethehome.com/RAID/simpleMTTDL.php

Proxmox

Package debian : Libpve-storage-perl

Matrice de compatibilité et de fonctionnalité

Man 8 pvesm - Proxmox VE Storage Manager

Base de la configuration : storage.cfg

API : /storage/

.deb : Libpve-storage-perl

Utilitaire de gestion

`/usr/sbin/pvesm`

Manuel

`/usr/share/man/man1/pvesm.1.gz`

Library Perl

`/usr/share/perl5/PVE/Storage.pm`

```
# dpkg -L libpve-storage-perl \
| perl -ape 's@^(/.*/)[^/]+@$1@ ; print "\n";' \
| sort -u

/.
/usr
/usr/
/usr/lib/
/usr/lib/udev/
/usr/lib/udev/rules.d/
/usr/libexec/
/usr/sbin/
/usr/share/
/usr/share/bash-completion/
/usr/share/bash-completion/completions/
/usr/share/doc/
/usr/share/doc/libpve-storage-perl/
/usr/share/man/
/usr/share/man/man1/
/usr/share/perl5/
/usr/share/perl5/PVE/
/usr/share/perl5/PVE/API2/
/usr/share/perl5/PVE/API2/Disks/
/usr/share/perl5/PVE/API2/Storage/
/usr/share/perl5/PVE/CLI/
/usr/share/perl5/PVE/Storage/
/usr/share/perl5/PVE/Storage/LunCmd/
/usr/share/zsh/
/usr/share/zsh/vendor-completions/
```

Laisser faire

- Wizard Ceph
- Assistance ZFS, avec réplication

… et caetera.

Matrice

compatibilité & fonctionnalité

Description	PVE type	Level	Shared	Snapshots	Stable
ZFS (local)	zfspool	file	no	yes	yes
Directory	dir	file	no	no	yes
BTRFS	btrfs	file	no	yes	technology preview
NFS	nfs	file	yes	no	yes
CIFS	cifs	file	yes	no	yes
Proxmox Backup	pbs	both	yes	n/a	yes
GlusterFS	glusterfs	file	yes	no	yes
CephFS	cephfs	file	yes	yes	yes
LVM	lvm	block	no	no	yes
LVM-thin	lvmthin	block	no	yes	yes
iSCSI/kernel	iscsi	block	yes	no	yes
iSCSI/libiscsi	iscsidirect	block	yes	no	yes
Ceph/RBD	rbd	block	yes	yes	yes
ZFS over iSCSI	zfs	block	yes	yes	yes

Sources : https://pve.proxmox.com/wiki/Storage
`man 1 pvesm`

Fonctionnalités d’un stockage

- Type : file vs block
- Thin provisioning ou allocation fine à l’usage
  - modalité block ou fichier avec format QCOW2
  - discard dans l’émulation du stockage VM
- Compression
- Déduplication

ACID :
Atomicité, Cohérence, Intégrité, Durabilité

Arborescence <file>

Content type	Subdir
VM images	images/<VMID>/
ISO images	template/iso/
Container templates	template/cache/
Backup files	dump/
Snippets	snippets/

Sources : https://pve.proxmox.com/wiki/Storage:_Directory

pmxcfs

Projection d’une base de donnée (SQLite) dans l’espace du système de fichier (FUSE).

- /var/lib/pve-cluster/config.db
- */etc/pve/…*

Synchronisation temps réel entre les nœuds

FS Posix incomplet

corosync

```
# man corosync_overview
```

<<Corosync is designed for applications to replicate their state to up to 16 processors.>>

Aime le multicast pour réduire la latence mais le mode unicast a été choisi pour proxmox, il impliquera des risques de /etc/pve mal synchronisé entre les nœuds qui utiliseront les timestamp pour s…

Vous êtes avertis.

/etc/pve/storage.cfg

```
<PVEtype>: <STORAGE_ID>

  <property>       [<value>]

# Common properties
  <disable>        # storage_check_enabled
  <shared>         # ! atomic lock by vmid
  <nodes>          <cluster_nodes_id> #allowed
  <content>        <content_type> # images,rootdir,vztmpl,backup,iso,snippets
  <prune-backups>  <keep-(all|hourly|daily|weekly|monthly|yearly)=NN>
  <bandwith>       <NN> # ratelimit_bps

# File storage backend common properties
  <format>         <file_format> # raw|vmdk|qcow2
  <preallocation>  <mode> # off|metadata|falloc|full


# Plugin properties
# /usr/share/perl5/PVE/Storage/[Custom/]<STORAGE_ID>Plugin.pm
#
#      sub properties # backend plugin configuration properties
#      sub option
```

<*VMID*>.conf

/etc/pve/nodes/<node_name>/<lxc|qemu-server>

<vm_storage_bus_ID>:
    <storage_backend>:
        vm-<vmid>-disk-<NN>
            [options storage bus]

Exemple :

```
scsi0: cephinfo:vm-500-disk-0,discard=on,size=3G,ssd=1
rootfs: cephinfo:vm-104-disk-0,acl=0,size=10G
```

Configuration stockage

VM / LXC : VMID unique pour Proxmox

storage.cfg → <vmid>.conf

En mode « file » :

Utile pour les migrations de

storage (changer de backend)

VM (passer un disque à une autre)

Cluster (passer une VM à un autre)

Rappel sur les FS (dans les VM)

- Ecriture atomique
- Journalisation

Comment feriez-vous avec un SGBD ?
Pourquoi pas voir la VM de la même manière ?

Oups !

Des questions ?