Evolution et tendances pour les technologies et solutions de stockage

Marc Mendez-Bermond – Expert Solutions HPC

marc_mendez_bermond@dell.com

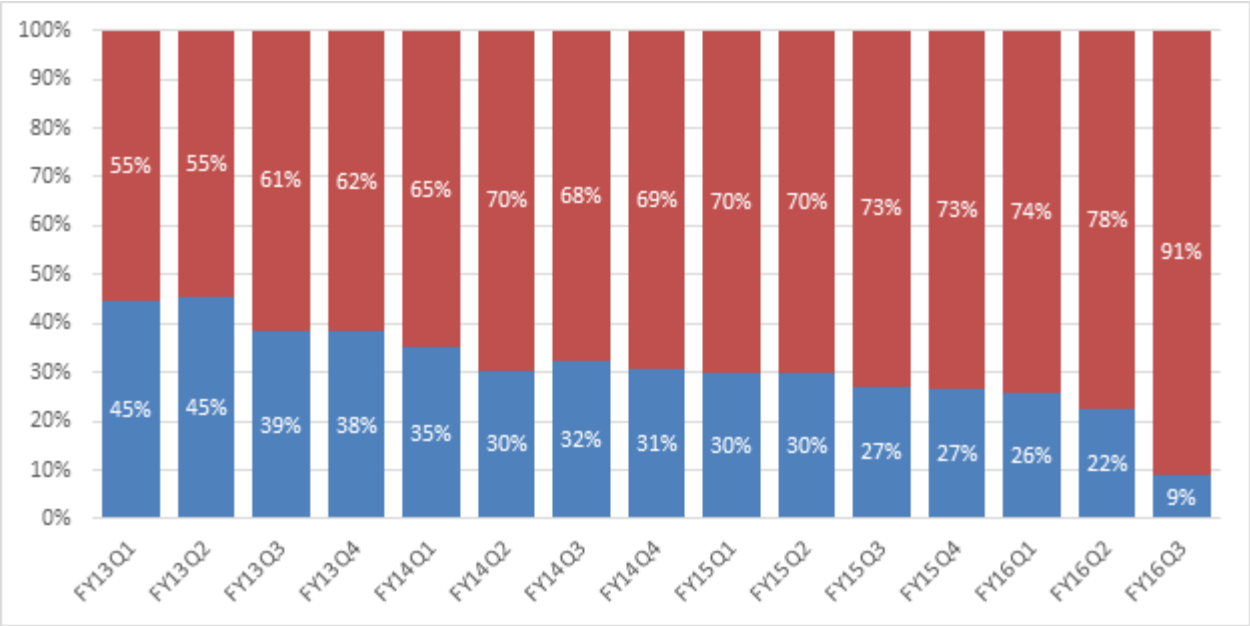
Agenda

- Technologies des media de stockage, tendances
- Solutions et directions
- Conclusions

Technologies

Paysage technologique

Paysage marché



Mix média : SAS 15k vs autres types.

Caractéristiques des différentes technologies

Technologie	Vitesse (relative)	Capacité (Go)	Prix/To (relatif)	W/To
DRAM	4000	64	5000	~160
SSD	12	4000	100	~0.8
HDD	4	8000	8	~1
TAPE	1 (?!?)	6250	1	~0 (?!?)

NOTE : les valeurs sont indicatives et varient en fonction des types de technologie.

Le **trilemne** de l'enregistrement magnétique

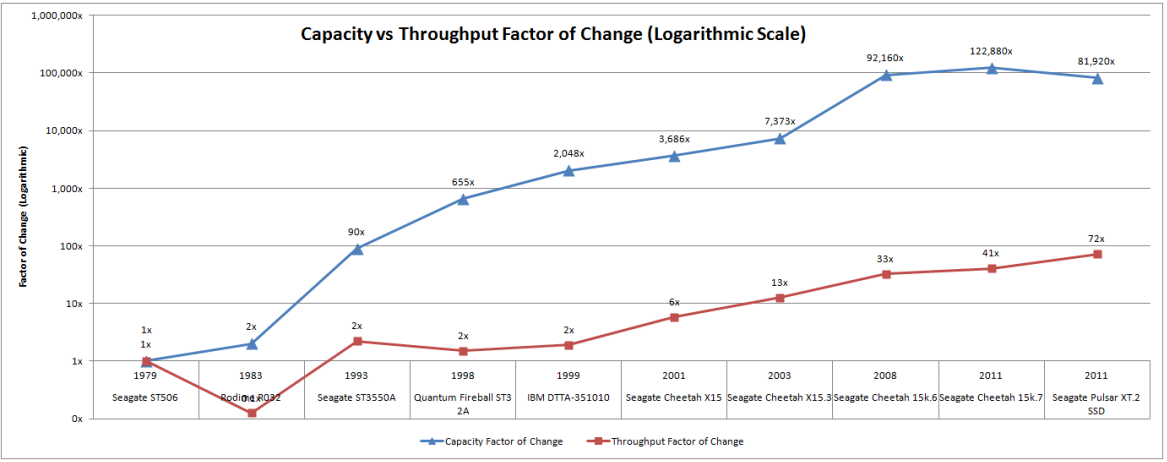
Writeability

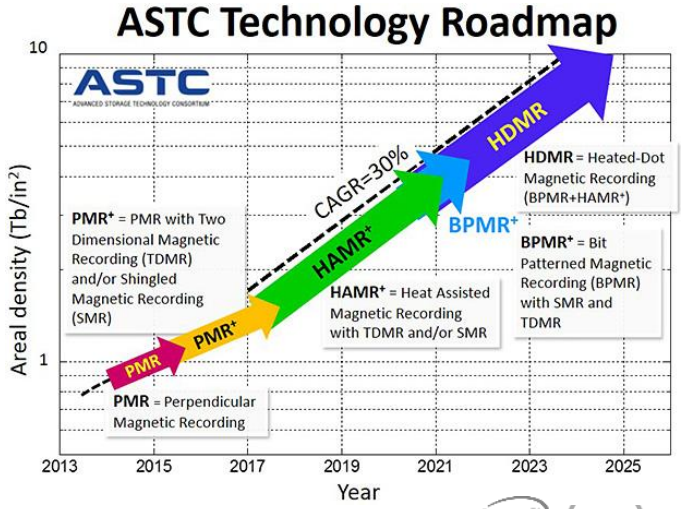
Readability

Stability

Evolution des disques mécaniques

- Evolution soutenue pour la capacité
- Evolution faible pour la performance
  - Le ratio performance/capacité est divisé par ~1000 (SSD) à 3000 (HDD)
- Dans le futur, la capacité reste le principal axe de développement.
- Les performances, au second plan, bénéficieront de technologies auxiliaires (cache SSD).



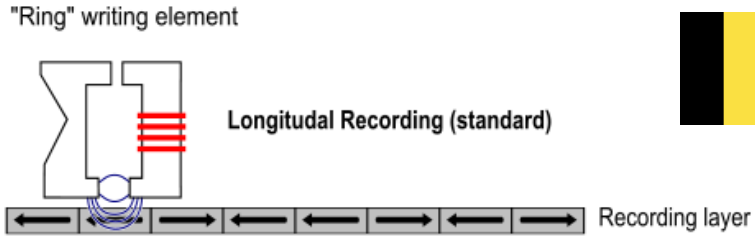


Source : ASTC

Technologies HDD courantes

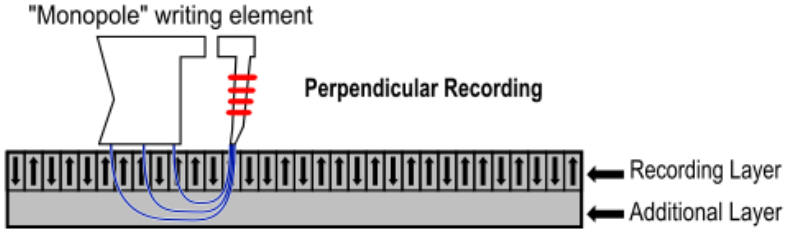
Longitudinal Recording (LMR)

- Limites faibles
- Effet superparamagnétique
- 200-300 Gb/in^2



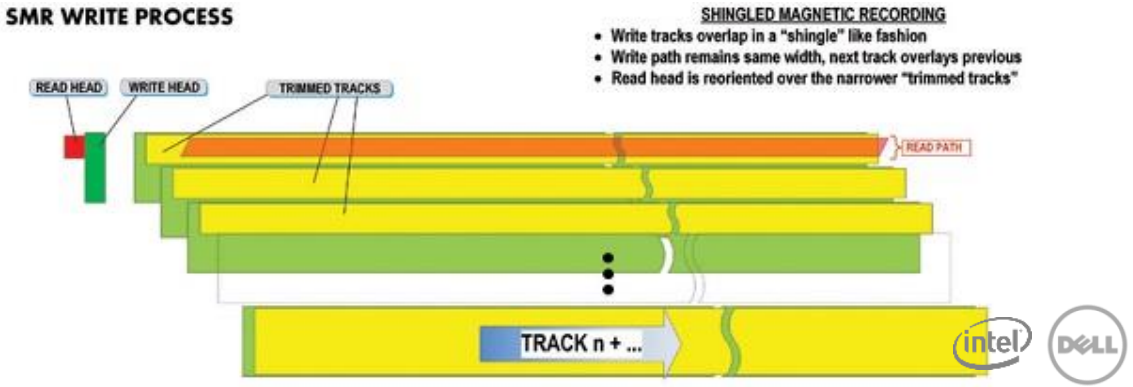
Perpendicular Recording (PMR)

- Densité plus élevée
- ~1 Tb/in^2
- Environ 1 nm/b

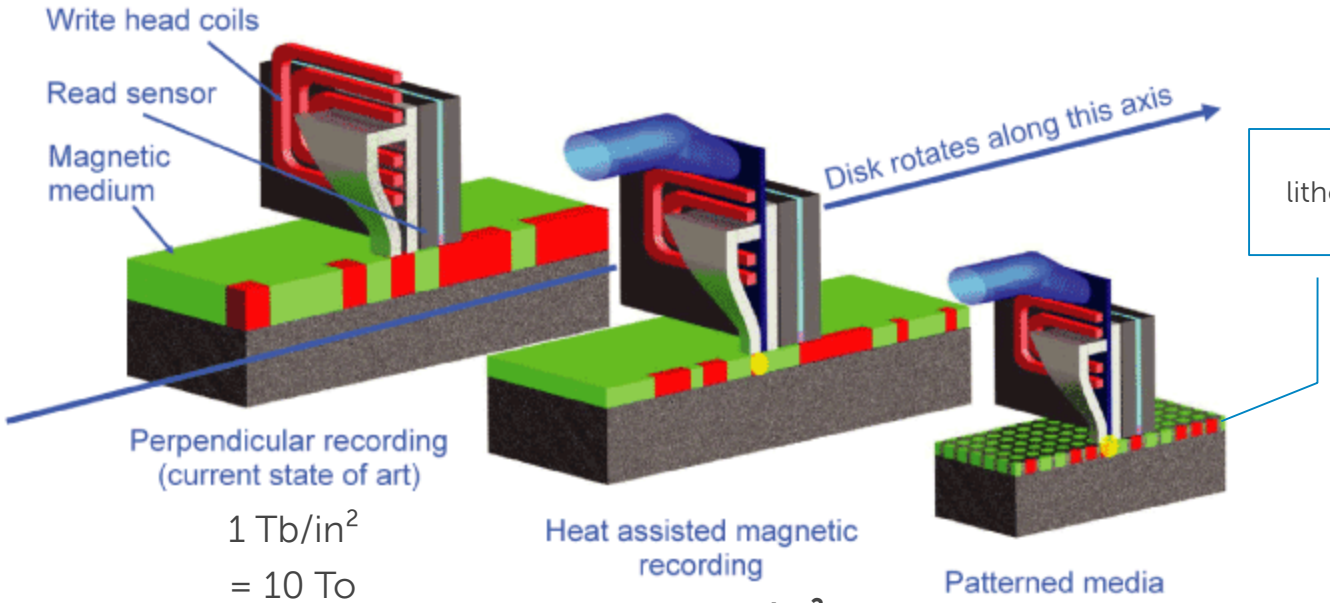


Shingled Recording (SMR)

- Densité comparable à PMR
- Ecriture et lecture de largeurs différentes
- Ré-écriture des pistes précédentes – WORM-like



Perspectives pour les disques mécaniques



1 Tb/in²
= 10 To

Est. 50 Tb/in²
= ~50 To
~2016

Est. 20-300 Tb/in²
= ~100 To
~2020

Image : Seagate

Silicium

Mémoires Flash

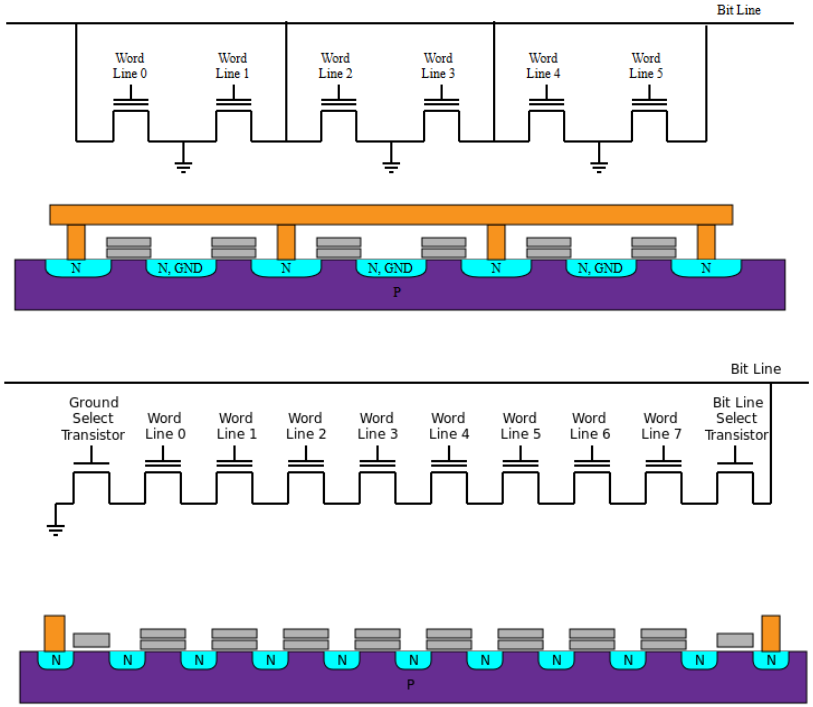
Stockage sur Silicium

Mémoires NOR : remplacement ROM

- Accès aléatoires
- Cycles lecture courts
- Systèmes embarqués/mobiles
- Taux d'erreur faibles

Mémoires NAND : remplacement HDD

- Accès par page
- Interface S/P
- Densité (x2.5)
- Taux d'erreur plus élevé :
  - Bad-block management
  - Wear-levelling
  - ECC



SLC, MLC et TLC

Densification de la capacité :

- **Single-Level Cells** = 1 bit/cellule
- **Multiple-Level Cells** = 2 bits/cellule
  - Enterprise MLC : taux d'erreurs plus faibles
  - Contrôleurs internes critiques pour la fiabilité et les performances !
- **Triple-Level Cells** = 3 bits/cellule
  - Contrôleurs toujours plus critiques !!!
- SanDisk 4x Flash = 4 bits/cellule
- Samsung : hybrides TLC + quelques cellules SLC – capacité et performances

… au détriment de la fiabilité, de la performance et de la consommation !

NVRAM

La recherche est très active autour de la mémoire non-volatile.

Mix des fonctionnalités de la DRAM et de la Flash :

- Persistance de l'information
- Rapidité d'accès (relative)
- Faible consommation électrique

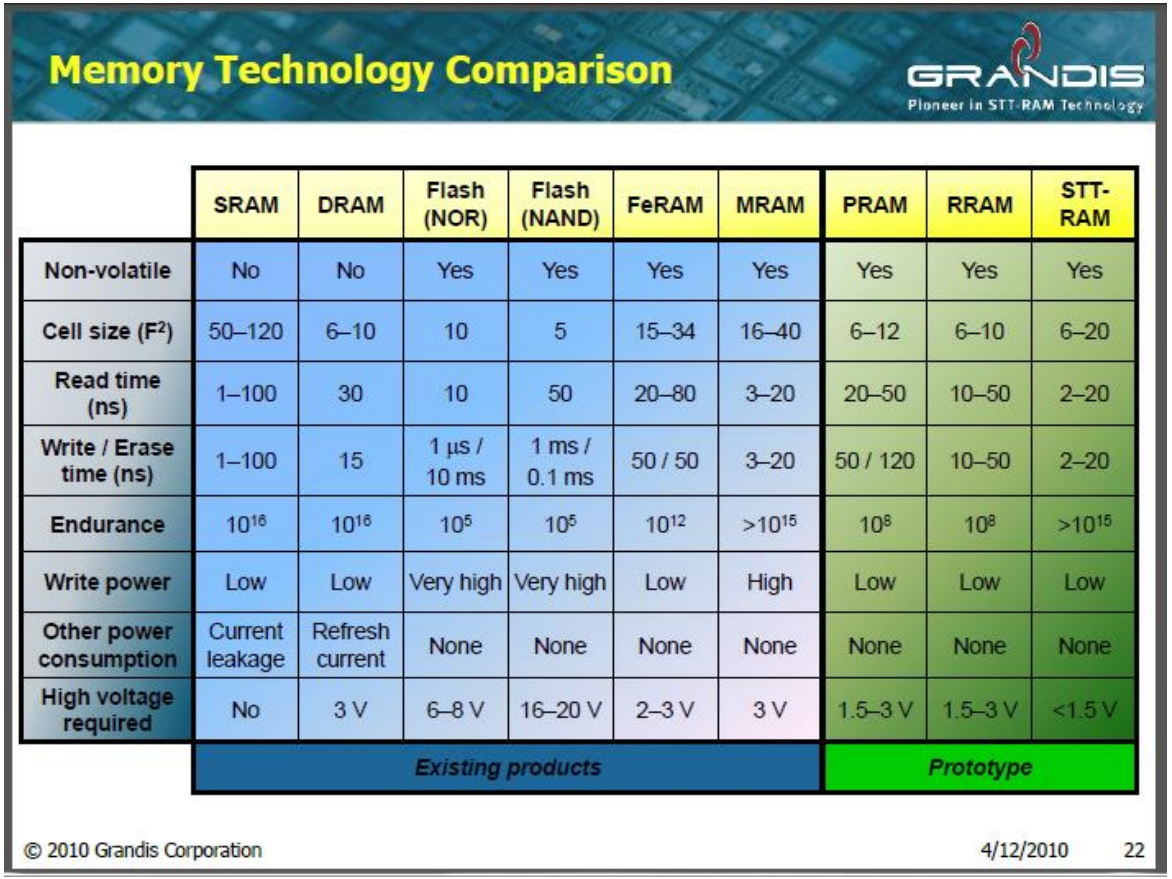
F-RAM, ReRAM, MRAM, OST-MRAM, PCM or P-RAM, STT-MRAM, CMOx, CBRAM, Racetrack memory, Memristors.

Mais dans les faits :

- Densité plus faible aujourd'hui
- Performances entre celles des Flash et des DRAM (fonction techno et opération)
- Sensibilité aux variations de températures et plages d'opération étroites
- Processus de fabrication de standards à complexes

Votre défi : apprécier les impacts sur le développement des infrastructures, des applications et des méthodes !

Matrice des technologies NVRAM comparées



Memory Technology Comparison

GRANDIS
Pioneer in STT-RAM Technology

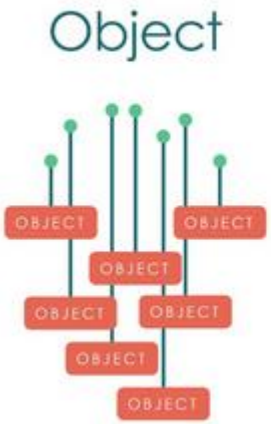
	SRAM	DRAM	Flash (NOR)	Flash (NAND)	FeRAM	MRAM	PRAM	RRAM	STT-RAM
Non-volatile	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cell size (F^2)	50–120	6–10	10	5	15–34	16–40	6–12	6–10	6–20
Read time (ns)	1–100	30	10	50	20–80	3–20	20–50	10–50	2–20
Write / Erase time (ns)	1–100	15	1 µs / 10 ms	1 ms / 0.1 ms	50 / 50	3–20	50 / 120	10–50	2–20
Endurance	10^{16}	10^{16}	10^5	10^5	10^{12}	$>10^{15}$	10^8	10^8	$>10^{15}$
Write power	Low	Low	Very high	Very high	Low	High	Low	Low	Low
Other power consumption	Current leakage	Refresh current	None	None	None	None	None	None	None
High voltage required	No	3 V	6–8 V	16–20 V	2–3 V	3 V	1.5–3 V	1.5–3 V	<1.5 V
	Existing products						*Prototype*		

4/12/2010 22

Solutions de stockage

Bloc, objet, fichier

- Bloc :
  - Stockage élémentaire (pas de mise en forme)
  - Fourniture d'un espace adressable brut
- Objet :
  - Stockage d'informations indexées
  - Métadonnées + OID
- Fichier :
  - Stockage d'information via interface (POSIX)
  - Métadonnées et hiérarchie dossiers/fichiers => système de fichiers



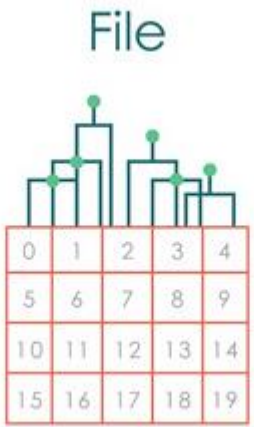
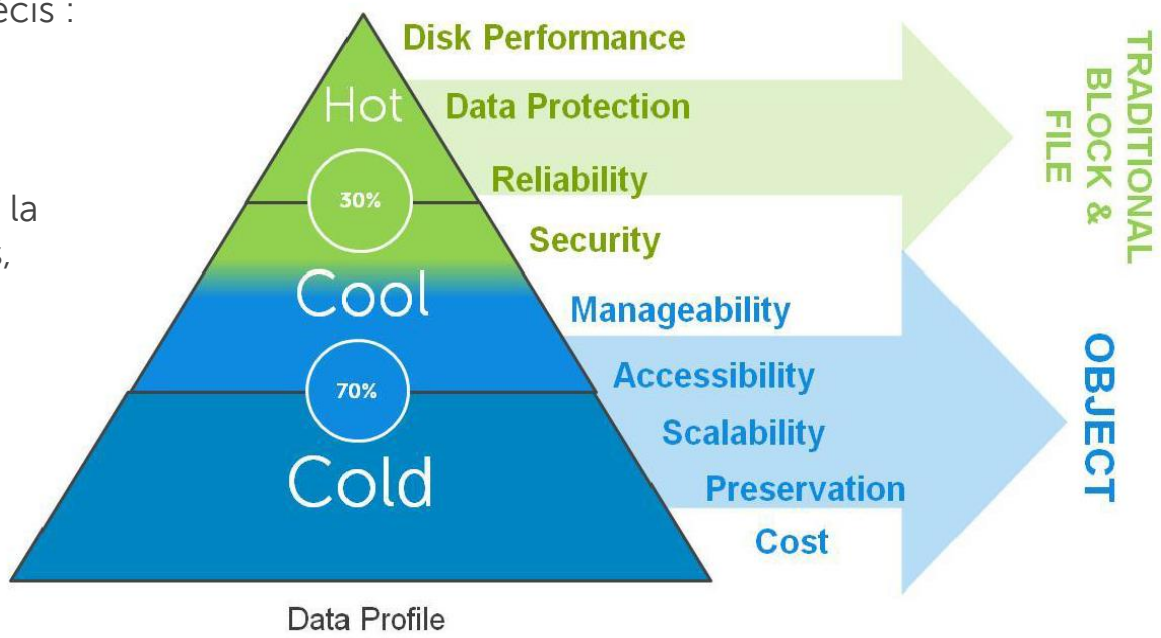


Image : Caringo

Positionnement des modes de stockage

Chaque mode répond à un besoin précis :

- Modes fichiers et blocs pour les **performances**
- Mode objets pour la **scalabilité** de la capacité et du nombre d'éléments, **préservation** et **évolutivité**



Stockage mode bloc

- L'interface d'accès est la plus élémentaire :
  - Les blocs sont indexés de manière similaire aux adresses mémoire
  - L'application ou l'OS en font usage
  - Sans système de fichiers ou autre couche d'abstraction (virtualisation), ce stockage a peu de sens
- En interne, une solution de stockage en mode bloc :
  - Implique potentiellement une sécurisation des informations : RAID, réplication …
  - Une gestion hiérarchique des niveaux de stockage : HSM, DLCM …
  - Un protocole d'accès variable : DAS, SAN …
  - Une gestion de la résilience : multipathing en particulier

Stockage en mode objet

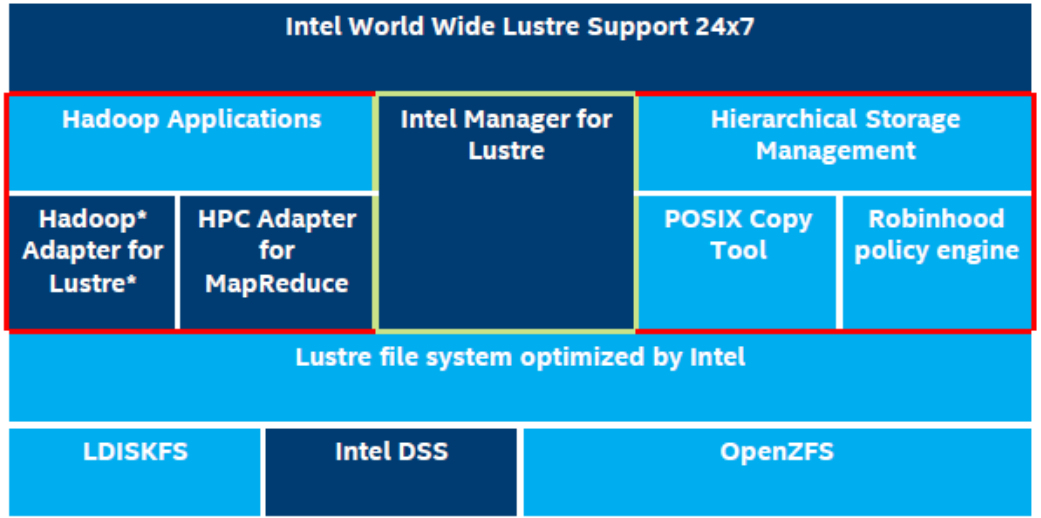
- L'interface d'accès permet d'organiser les données :
  - Typiquement : PUT(), GET() et DELETE()
  - Parfois (souvent) des interfaces plus traditionnelles : NFS, CIFS, FTP …
  - Un identifiant unique par objet (OID) est remis à l'insertion
  - L'application ou l'infrastructure globale doivent préserver l'identifiant (OID)
- Le stockage en mode objet est une infrastructure distribuée :
  - Les objets sont hébergés sur les moyens de stockage rendus **banalisés**
  - La **préservation** est la responsabilité de l'infrastructure : réplication, codage d'erreur, …
  - Les mécanismes du stockage en mode objet autorisent une **évolutivité forte** en capacité et en nombre d'objets
- L'organisation des données est « centralisée » :
  - Généralement on construit une interface et une base de données « métier » permettant de rechercher les OID en fonction des métadonnées => effort d'intégration
  - Spécialisation => confort d'utilisation pour les utilisateurs et expertise administrateurs

Stockage en mode fichiers

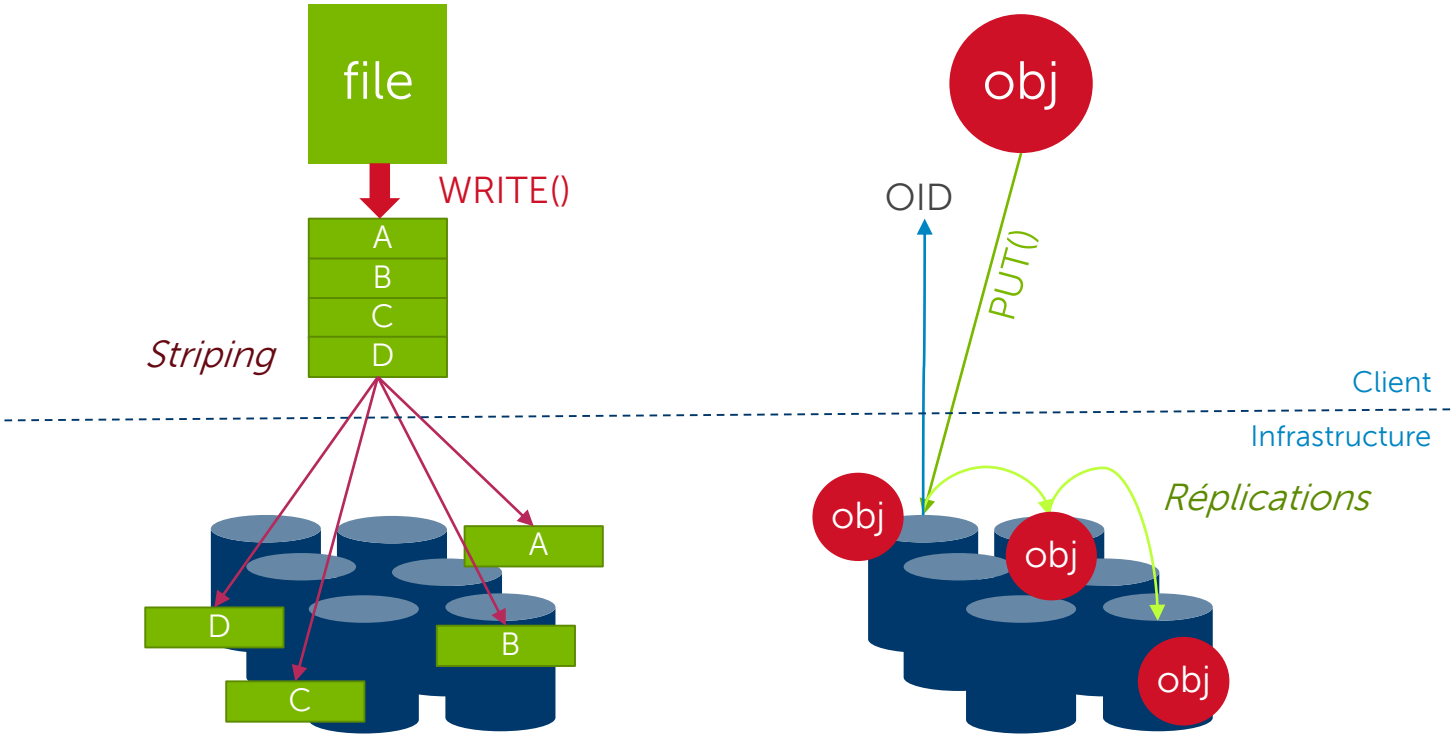
- Pour l'interface d'accès :
  - API standard d'un système de fichiers virtuel, puis système de fichiers physique
  - Grande liberté d'organisation pour l'utilisateur
  - Faible contrôle pour les administrateurs; limité à la supervision et quota
- Le stockage en mode fichiers est centralisé par nature :
  - L'ensemble des utilisateurs partagent de multiples ressources
  - Les contentions d'accès sont gérés par des verrous
  - Certains cas permettent la distribution et la **parallélisassions** (☺) du FS
- L'activité et l'organisation des données appliquée par les utilisateurs impactent fortement les performances du systèmes :
  - Nombres d'entrées (inodes) limitées et/ou fixes
  - Goulots d'étranglements multiples : stockage, serveurs, réseaux …

Intel® Enterprise Edition for Lustre*

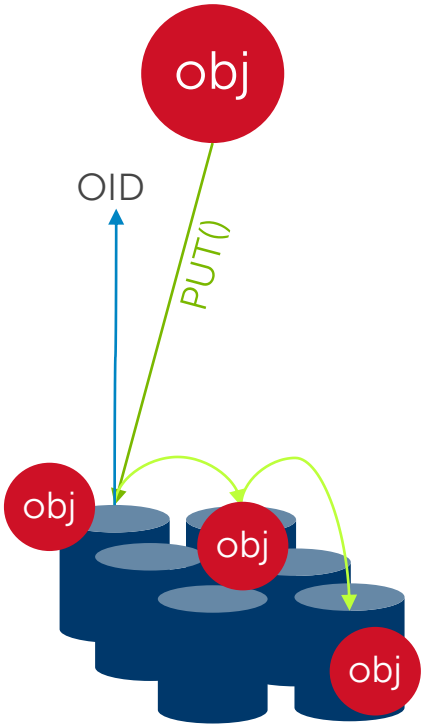
- 24x7 Enterprise support
- Intel Lustre adapters for Big Data
- Intel Manager for Lustre
- HSM support, including the Robinhood policy engine
- Production quality Lustre file system enhanced by Intel
- Maximum data protection using ZFS
- Improving small files performance using Differentiated Storage Services
- Professional services and training



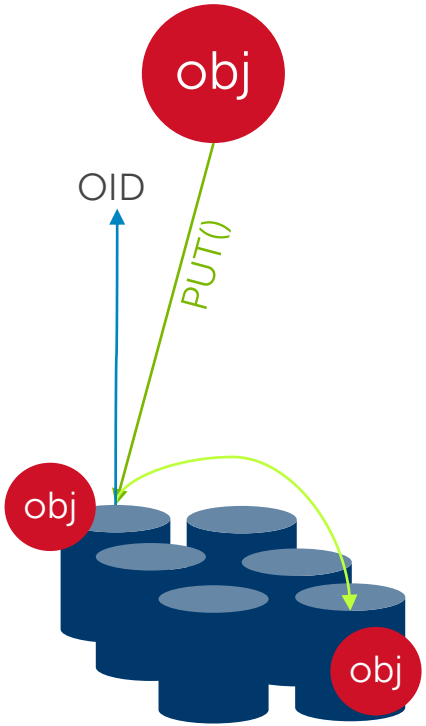
Distribution en mode fichiers et en mode objets



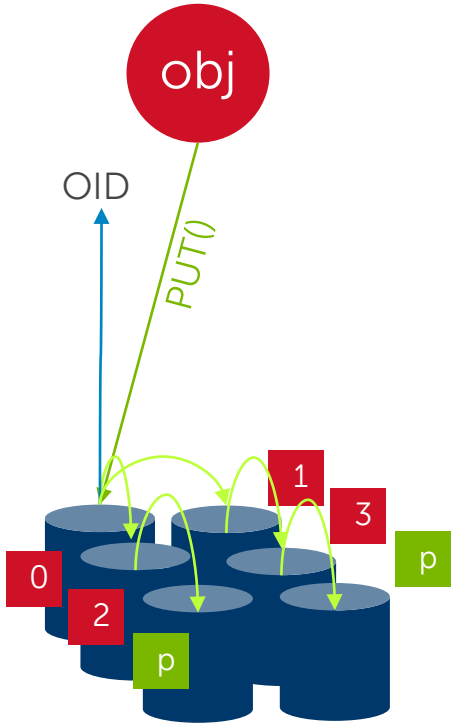
Stockage distribué en mode objet



Réplication 3x



Réplication 2x



Erasure Coding

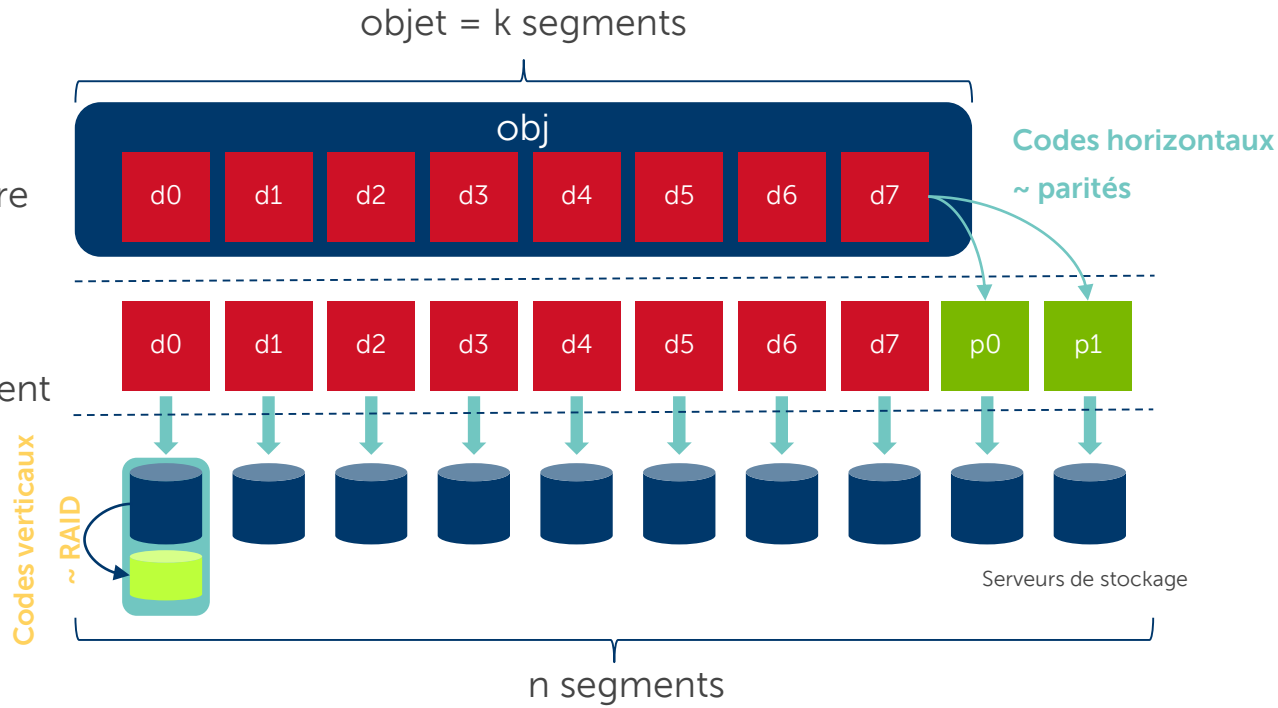
Erasure Coding

Système d'équations linéaires assurant l… sécurisation du stockage de l'informati…

Cette technique autorise une meilleure utilisation de l'espace brut disponible sur l'infrastructure distribuée tout en renforçant sa résilience.

Certains arrangements dans les algorithmes permettent également d'améliorer les temps de reconstruction et de soulager le réseau avec des transferts de données limités.

Mots clés : systèmatiques, *maximum distance separable (MDS)*, horizontal/horizontal-vertical …



Tolérance aux pannes : (n – k)

Vitesse de reconstruction : fonction des algorithmes appliqués

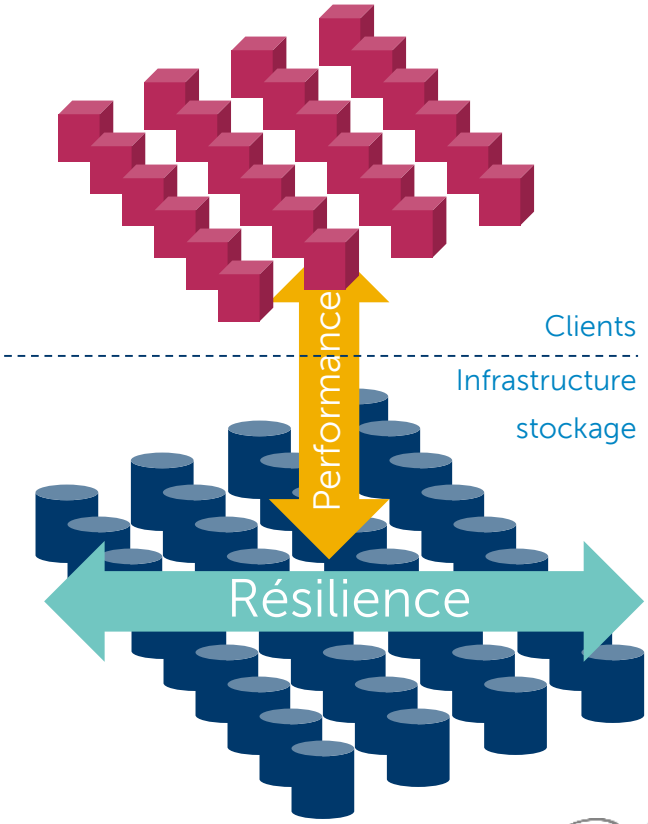
Impacts sur l'infrastructure de stockage distribuée

Les solutions de stockage distribué répondent aux besoins croissants :

- Capacité
- Nombre d'éléments
- Performances
- Résilience
- Pérennisation de l'infrastructure

Cependant elles nécessitent la prise en compte de contraintes fortes pour leur évolutivité :

- Performances par élément de stockage (serveurs)
- **Capacité des réseaux d'interconnexion**
- Supervision de l'infrastructure composée de nombreux éléments augmentant le taux de panne



Conclusions

Conclusion : 1er tour

- Technologies :
  - Le marché des équipements grand-public gouverne !
  - A court terme (2020), les capacités seront gagnantes et l'impact des nouvelles technologies disque pourra être sévère pour les performances (SMR)
  - A plus long terme (2020+), les mémoires Flash (SSD) pourraient prendre le dessus au bénéfice du rapport performance/capacité
  - Toujours à long terme, les mémoires non-volatiles vont constituer une rupture technologique franche
- Solutions :
  - Les modes distribué ou parallèle s'imposent de fait pour un nombre croissant d'applications
  - La banalisation des systèmes de stockage est une nécessité pour garantir la pérennité des installations
  - Le stockage d'objets conduirait au remplacement des silos de stockage par des silos métier

Conclusion de la conclusion

Placer son niveau d'exigences clairement pour chaque poste :

- Activité planifiée (*guess work* ≠ audit)
- Intégration à l'existant
- Sécurisation
- Performances
- Capacité
- Evolutivité
- Facilité de mise en œuvre et de supervision

Le domaine du stockage est traditionnellement conservateur :

- Cas d'études
- Maquettes
- Démonstrations
- Attention aux benchmarks : ils ne démontrent souvent qu'un cas particulier de votre usage !

Les nouvelles technologies promettent de belles choses mais cachent de sérieux compromis et la nécessité de rapprocher utilisateurs, administrateurs et fournisseurs !

Merci !

The power to do more