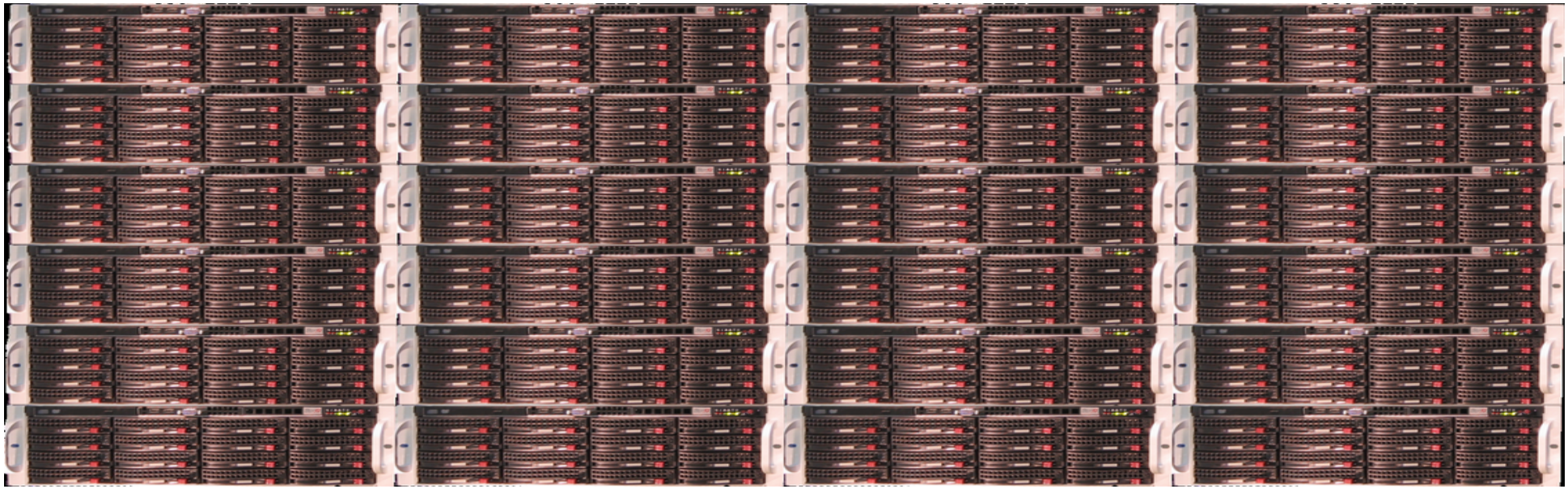
intellique
Open Source Storage Solutions

Le RAID



Le RAID?

Notre sujet n’est pas l’unité d’élite de la police...

intellique
Open Source Storage Solutions

Le RAID?

Notre sujet n’est pas non plus ...

intellique
Open Source Storage Solutions

Le RAID?

- signifie "*Redundant Array of Inexpensive Drives*".
- inventé au début des années 80
- moins coûteux
- plus performant

intellique
Open Source Storage Solutions

Les niveaux de RAID

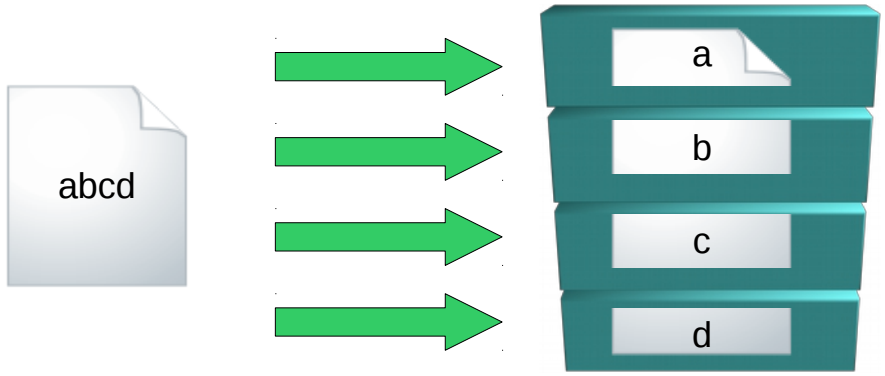
- RAID 0 (*striping)*
- RAID 1 (miroir)
- RAID 10
- RAID 4
- RAID 5
- RAID 6
- RAID Z et autres....

Le RAID 0

- Pas vraiment RAID : pas de redondance

Les plus:
• amélioration de la capacité
• Amélioration des performances

Les moins:
• Sécurité dégradée!



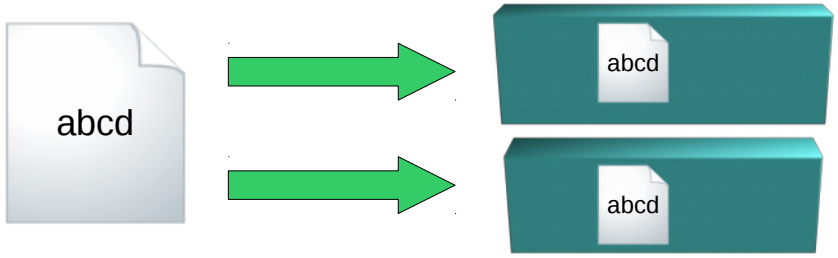
intellique
Open Source Storage Solutions

Le RAID 1

- Miroir, ou copie complète

Les plus:
• amélioration de la sécurité

Les moins:
• Au détriment de la capacité



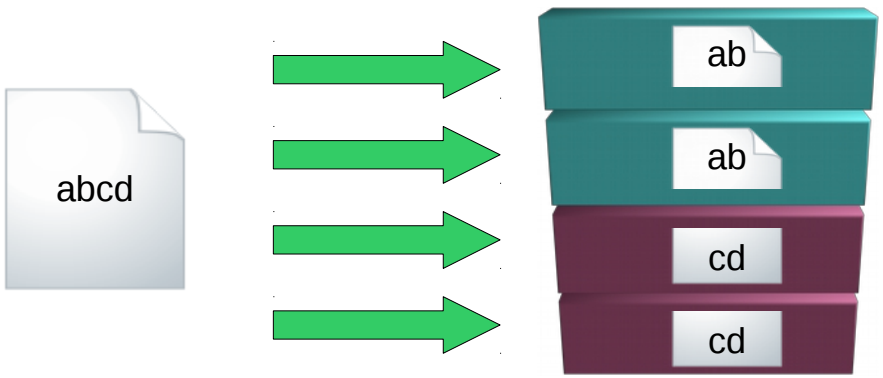
intellique
Open Source Storage Solutions

Le RAID 10

- aussi RAID 0+1 ou RAID 1+0

Les plus:
• performant et capacitif comme RAID-0
• aussi sûr que RAID-1

Les moins:
• aussi coûteux que RAID-1 aussi...



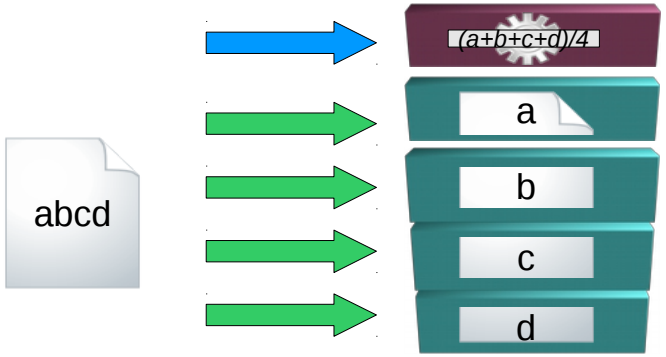
intellique
Open Source Storage Solutions

Le RAID 4

- Un disque de parité pour *N* disques
- Souvent improprement appelé RAID 3 dans le passé (*Ciprico*, *Maxstar*...)
- Souvent non assumé (*DatadirectNetworks*...)

Les plus:
- performance des N disques en lecture
- performance des N disques en écriture séquentielle

Les moins:
- performance d'un seul disque en écriture aléatoire



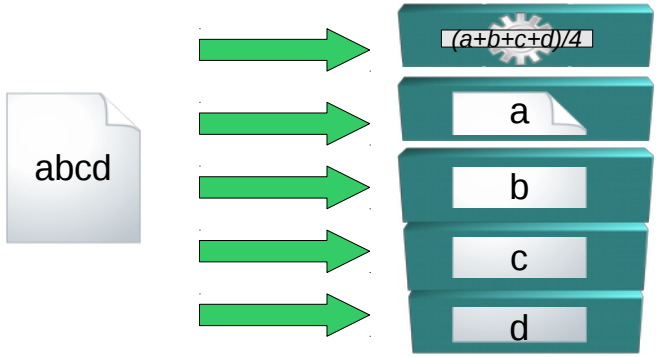
intellique
Open Source Storage Solutions

Le RAID 5

- un bloc de parité pour N blocs
- parité répartie sur tous les N+1 disques

Les plus:
• séparation possible des E/S sur disques en lecture

Les moins:
• Performances en écriture très basse si blocs trop grands (*lecture/modification/écriture*)



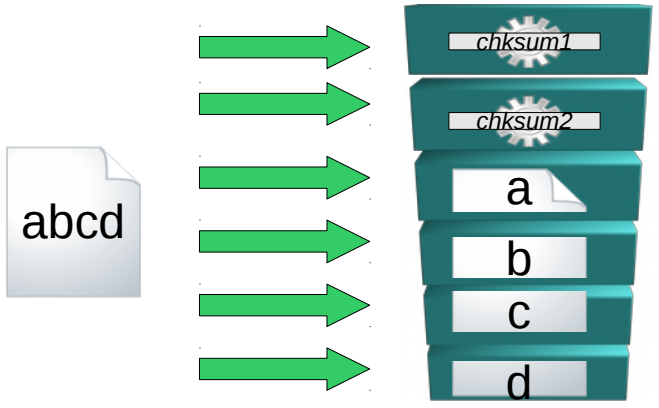
intellique
Open Source Storage Solutions

Le RAID 6

- deux blocs de parité pour N blocs
- parité répartie sur tous les N+2 disques

Les plus:
• sécurité encore meilleure statistiquement que RAID-10
• surcoût raisonnable

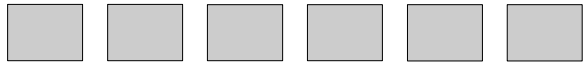
Les moins:
• performances limitées en écriture
(*lecture/modification/écriture* fréquent)



Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.



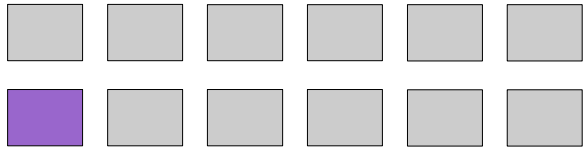
intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...



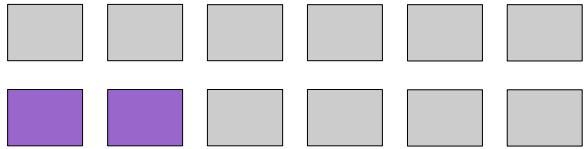
intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...

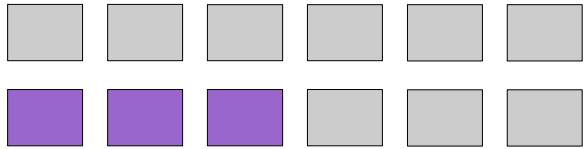


Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...



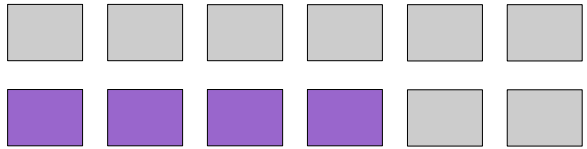
intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...



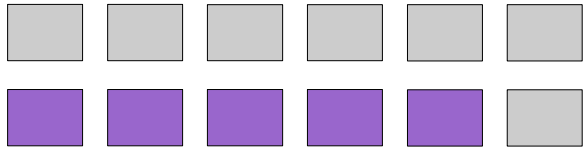
intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...



intellique
Open Source Storage Solutions

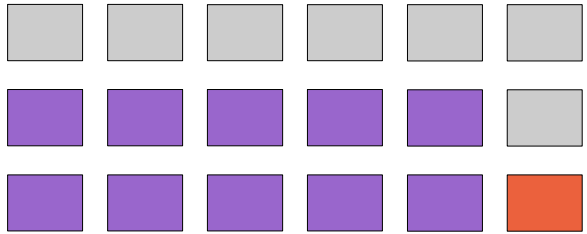
Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

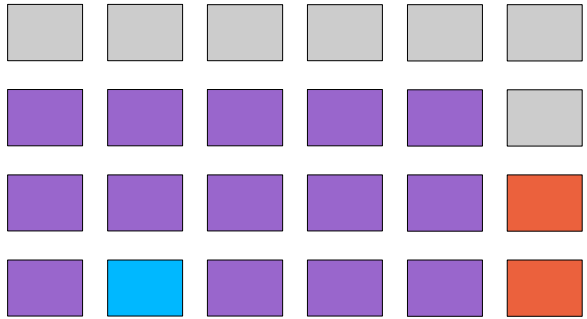
On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité

Temps 4 : on modifie un des blocs (mais pas tous)



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

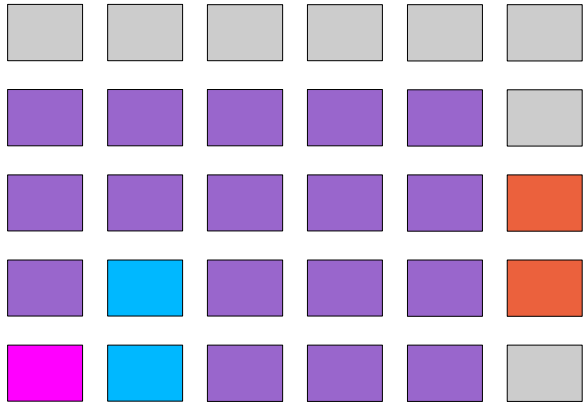
Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

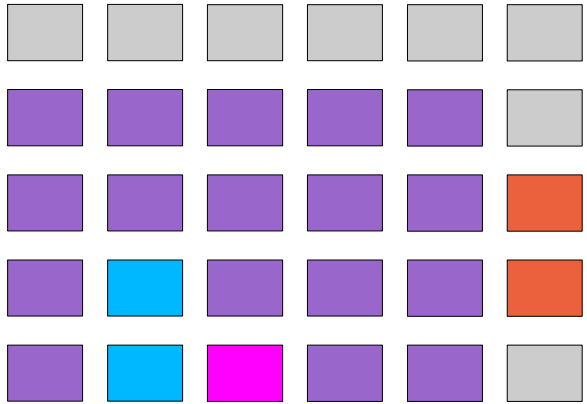
Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...



Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

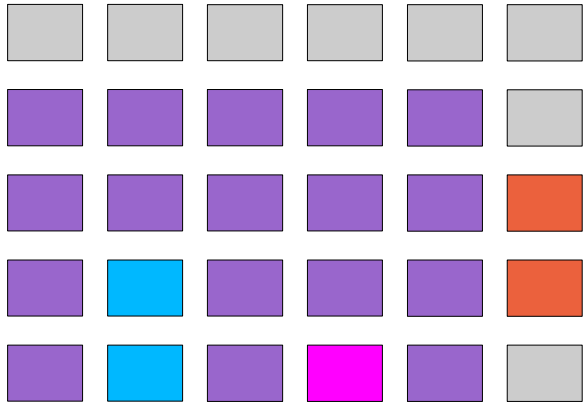
Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

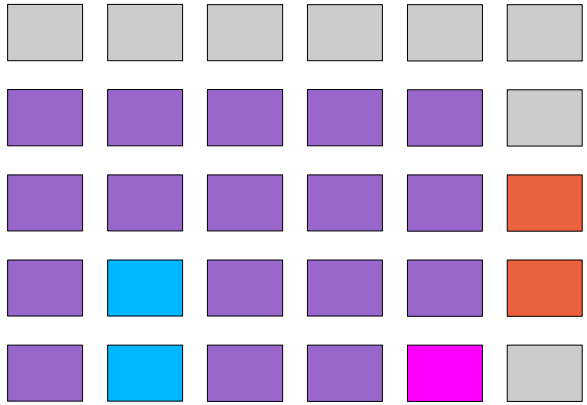
Temps 1 : pas de données.

Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

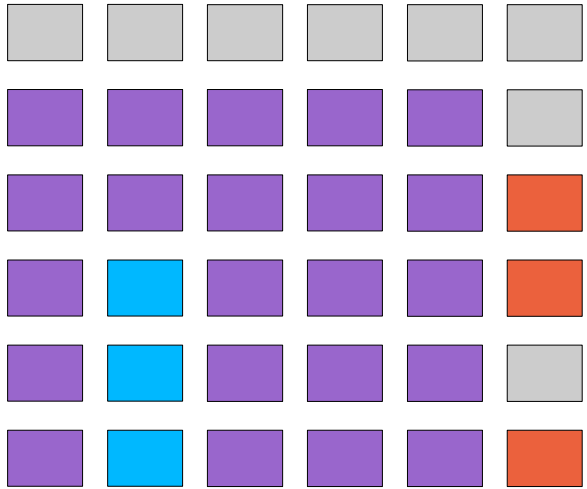
Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...

Temps 6 : pour pouvoir récrire la nouvelle parité !



intellique
Open Source Storage Solutions

Digression : Le cycle lecture/modification/écriture

On répartit les données sur en blocs de taille égale (par ex. 64 Ko), et on ajoute un bloc (ou deux) de parité de même taille, puis on stocke un bloc par disque. Exemple avec 6 disques et une parité (RAID-5) :

Temps 1 : pas de données.

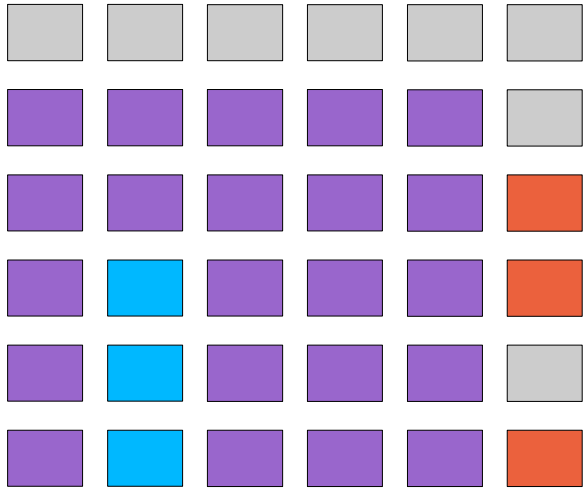
Temps 2 : on écrit les données...

Temps 3 : puis le bloc de parité.

Temps 4 : on modifie un des blocs (mais pas tous)...

Temps 5 : on doit relire les autres blocs de la bande...

Temps 6 : pour pouvoir récrire la nouvelle parité !



Résultat : pour modifier quelques Ko, on peut devoir relire plusieurs Mo et récrire plusieurs dizaines de Ko (jusqu'à trois blocs).

intellique
Open Source Storage Solutions

Le RAID Z et autres

- peut cumuler le fonctionnement des différents RAID

Les plus:
• parité calculée uniquement sur l'espace utilisé
• pas d'effet *lecture/modification/écriture*
• reconstruction plus rapide
• bénéfices additionnels selon les solutions : clichés, mise en grappe, etc.

Les moins:
• presque entièrement logiciel (contention de ressources)
• enveloppe de performances souvent restreinte

intellique
Open Source Storage Solutions

Le risque de panne

- soit x la probabilité qu'un disque tombe en panne sur la période considérée;
- soit $P(raid)$ la probabilité de panne de la grappe;
- soit N le nombre de disques :

$$P0 = Nx$$

$$P1 = x^2$$

$$P10 = x^2 . N/2$$

$$P4 = x^2 . (N - 1)$$

$$P5 = x^2 . (N - 1)$$

$$P6 = x^3 . (N - 2)$$

Quel mode utiliser?

Soit N le nombre de disques, t le débit et Ta le temps d'accès d'un disque :

- RAID-0 : performance séquentielle pure
  - débit : $N . t$
  - latence : Ta / N

- RAID-10 : performance en écriture garantie
  - débit en écriture : $N/2 . t$
  - débit en lecture : $N . t$
  - Latence : $Ta / (N/2)$

intellique
Open Source Storage Solutions

Quel mode utiliser?

Soit N le nombre de disques, t le débit et Ta le temps d'accès d'un disque :

- RAID-4 : performance d'écriture séquentielle et sécurité
  - débit : $(N - 1) . t$
  - Latence : $(N . Ta) / (N - 1)$

- RAID-5 : performance en E/S et sécurité
  - débit en écriture : $N/2 . t$
  - débit en lecture : $(N - 1) . t$
  - Latence : $Ta / (N - 1)$

- RAID-6 : performance en lecture et forte sécurité
  - débit en écriture : $N/3 . t$
  - débit en lecture : $(N - 2) . t$
  - Latence : $Ta / (N - 2)$

intellique
Open Source Storage Solutions

RAID et taille des disques

- plus les disques sont gros,

  plus la durée de reconstruction est longue
- en 1988, un disque de 40 Mo se remplissait en 2 à 3 minutes
- en 1998, un disque de 9 Go se remplissait en 15 à 20 minutes
- en 2008, un disque de 1 To se remplissait en 3 à 4 heures
- en 2010, un disque de 2 To se remplissait en 5 à 6 heures
- En 2013, un disque de 4 To se remplit en 12 à 14 heures

intellique
Open Source Storage Solutions

RAID et taille des disques

- plus les disques sont gros,

  plus la durée de reconstruction est longue

- Pour éviter de rester longtemps en mode dégradé, on utilise les modes à double parité (RAID 6).

intellique
Open Source Storage Solutions

RAID et taille des disques

- Le taux d'erreur non recouvrable n'évolue plus : plus les disques sont gros plus le risque est grand
- Le RAID 5 est **OBSOLÈTE.**
- N'utilisez **JAMAIS** le RAID 5 avec des disques plus gros que 400 Go.

intellique
Open Source Storage Solutions

Taille de « bande »

- une taille de bande optimale par rapport à l'application permet de maximiser la performance
- pour de petites entrées/sorties, on utilisera une bande large pour accéder aux disques successivement.
- pour de grosses entrées/sorties, on utilisera une bande étroite pour accéder aux disques simultanément.

intellique
Open Source Storage Solutions

Calcul de la parité

- Le calcul de la parité est intensif en terme de ressources.
- le processeur en charge de ce calcul peut être dédié
  - c'est le **RAID matériel**
- ou ce peut être le processeur central
  - c'est le **RAID logiciel**

intellique
Open Source Storage Solutions

RAID matériel et RAID matériel

- tous les RAID matériels ne sont pas égaux
- certains sont très peu performants car sous-dimensionnés (processeur, mémoire)
- certains utilisent une ressource interne (mémoire) pouvant entraîner une contention

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

- le RAID logiciel est généralement plus performant lors des tests de vitesse disque

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

MAIS

RAID matériel et RAID logiciel

- il entre en concurrence pour les ressources systèmes avec les autres tâches de façon difficile à prévoir
  - sur un système déjà chargé, la panne d'un disque en RAID logiciel peut entraîner une chute catastrophique des performances
  - la performance du RAID matériel est stable quelle que soit l'activité du système par ailleurs

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

DONC

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

- pour des stations de travail ou des systèmes n'exécutant qu'**une seule tâche** principale, ou sous une charge modérée, le **RAID logiciel est très satisfaisant.**
- pour des systèmes serveurs avec de **nombreux utilisateurs** et travaillant parfois sous une **charge très élevée**, le RAID logiciel est **inadapté**.

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

- Devinette

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

- Devinette
- dans quel cas peut-on utiliser du RAID logiciel sur un système de haute performances ?

intellique
Open Source Storage Solutions

RAID matériel et RAID logiciel

- Devinette
- dans quel cas peut-on utiliser du RAID logiciel sur un système de haute performances ?
- lorsque c'est un système dédié dont la seule tâche est de traiter les entrées/sorties (*Isilon*, *NetApp*, etc).

intellique
Open Source Storage Solutions

Le RAID

- C’est fini!
- Des questions ?

info@intellique.com

- Tél: 01 78 94 84 00