Engineer

RAID

☰ plusieurs disques réunis, souvent pour tolérer une panne

Ce que c’est

RAID fait fonctionner plusieurs disques comme un seul périphérique bloc, afin que le groupe puisse être plus rapide, de plus grande capacité ou plus résistant qu’un seul de ses membres. Le nom vient d’un rapport technique de 1987 de l’université de Californie à Berkeley, rédigé par David Patterson, Garth Gibson et Randy Katz : *Plaidoyer pour des ensembles redondants de disques peu coûteux (RAID)*. Linux écrit les niveaux en minuscules et sans espace, raid1 ou raid5, comme les affiche son pilote md.

Le rapport part d’un décalage. Les processeurs et la mémoire devenaient plus rapides, et le résumé prévient que ce gain est perdu si les entrées-sorties ne suivent pas. Il oppose le grand disque unique coûteux, Single Large Expensive Disk ou SLED, dont la capacité avait rapidement augmenté mais les performances seulement modestement, aux ensembles de disques bon marché destinés aux ordinateurs personnels, dont il attend, face au grand disque, des gains d’un facteur d’environ dix en matière de vitesse, de fiabilité, de consommation et de possibilités d’extension. La comparaison est concrète : un disque de grand système IBM 3380, un Fujitsu Super Eagle pour mini-ordinateurs et le Conner CP 3100 de 100 Mo pour ordinateurs personnels.

Vient ensuite le revers, dans une section que les auteurs présentent comme une mauvaise nouvelle. Si les disques tombent en panne indépendamment et à taux constant, hypothèse que le rapport attribue aux fabricants lorsqu’ils annoncent un temps moyen avant panne (MTTF), le MTTF d’un ensemble est celui d’un disque divisé par le nombre de disques. Cent CP 3100, chacun donné pour 30 000 heures, donnent 300 heures, moins de deux semaines ; mille donnent 30 heures. La conclusion est nette : « Sans tolérance aux pannes, les grands ensembles de disques peu coûteux sont trop peu fiables pour être utiles. »

La réponse est la redondance. Les disques sont répartis en groupes de fiabilité, chacun doté de disques de contrôle supplémentaires contenant de l'information redondante. Lorsqu'un disque tombe en panne, il est remplacé et son contenu est reconstruit sur le nouveau à partir de cette information ; la durée nécessaire est le temps moyen de réparation, que réduisent des disques de secours à chaud activés électroniquement lorsqu'un disque meurt. Le rapport présente ensuite cinq modes d'organisation et appelle chacun d'eux un niveau. Le niveau 1 met chaque disque en miroir. Le niveau 2 corrige les erreurs avec un code de Hamming, comme les systèmes de mémoire. Le niveau 3 ne nécessite qu'un disque de contrôle par groupe. Le niveau 4 permet de petites lectures et écritures simultanées sur des disques distincts. Le niveau 5 répartit données et informations de contrôle sur tous les disques pour qu'aucun disque de contrôle unique ne ralentisse les écritures.

La pratique a enrichi la liste. mdadm, l'outil de gestion du RAID logiciel de Linux, accepte notamment linear, raid0 (ou stripe), raid1, raid4, raid5, raid6 et raid10. La répartition par bandes sans redondance, aujourd'hui appelée niveau 0, ne figure pas dans la classification de 1987, qui commence par les disques en miroir. Le rapport laissait ouverte la réalisation matérielle ou logicielle d'un ensemble, en la présentant comme une simple question de coûts et d'avantages. Linux md est logiciel ; une carte contrôleur qui présente au système d'exploitation un seul grand disque est la solution matérielle.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce que RAID ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Des sauvegardes testées plutôt que supposées fiables, et un plan écrit pour le matin où un serveur ne redémarre pas.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. 100 sauvegardes critiques automatisées : Barman — 100 sauvegardes critiques automatisées dans les bases

100 sauvegardes critiques automatisées : Barman Contactez-nous →

Pourquoi c'est important

Un disque s'use, et sur un serveur à disque unique sa panne arrête la machine et laisse les données dépendre de la sauvegarde existante, s'il y en a une. Le rapport ouvre sa section sur la fiabilité par ce constat : les disques sont assez peu fiables pour que les administrateurs sauvegardent tout chaque nuit. Avec la redondance, une panne de disque devient une réparation plutôt qu'une

interruption. L'ensemble continue avec ses membres restants, dans un état que Linux appelle degraded, puis retrouve sa pleine capacité de fonctionnement lorsqu'un disque de remplacement a été ajouté et reconstruit.

Cette protection ne dure que tant que les disques restants tiennent. Le calcul du rapport considère les données perdues lorsqu'un deuxième disque du même groupe tombe en panne avant la réparation du premier : la durée de la fenêtre de réparation compte donc autant que le nombre de disques de contrôle. Avec des disques de secours à chaud et des réparations courtes, les chiffres deviennent impressionnants : le rapport calcule qu'un ensemble de niveau 5 de 1 000 disques, par groupes de dix et avec 5 pour cent de disques de secours, aurait encore un MTTF de 90 ans. Ailleurs, il avertit qu'un séisme ou une surtension peut faire tomber ensemble les disques d'un même ensemble, alors que ses formules supposent le contraire.

Choisir un niveau est un compromis entre capacité, vitesse et coût. Le rapport conclut que les disques en miroir conviennent au traitement transactionnel qui n'utilise pas plus de la moitié du stockage, et que le niveau 5 semble préférable lorsqu'une plus grande capacité est nécessaire, pour le calcul sur superordinateur ou pour un mélange de ce calcul et de traitement transactionnel.

Lorsque la redondance ne suffit plus, il reste la récupération. La réalisation consacrée à la récupération de données sur cartes SD, disques durs, SSD et RAID traite un ensemble RAID défaillant comme un problème distinct d'un SSD hors service ou d'un disque à plateaux en panne, l'approche appropriée dépendant du support.

À quoi cela ressemble en pratique

Prenons un serveur avec une partition sur chacun de deux disques, `/dev/sda1` et `/dev/sdb1`, à mettre en miroir pour un volume de données. Avec Linux md, trois commandes construisent l'ensemble, y créent un système de fichiers et l'enregistrent dans la configuration :

```
mdadm --create /dev/md0 --level=1 --raid-devices=2 /dev/sda1 /dev/sdb1
mke2fs -t ext4 -L data /dev/md0
mdadm --detail --scan >> /etc/mdadm/mdadm.conf
```

La première crée `/dev/md0`, un ensemble RAID1 composé des deux partitions, et le noyau lance une resynchronisation, qui recalcule la redondance comme après toute création. La deuxième formate l'ensemble comme un disque unique : le système de fichiers ne voit qu'un périphérique. La troisième ajoute une ligne décrivant l'ensemble au fichier de configuration de mdadm, situé à `/etc/mdadm.conf` ou `/etc/mdadm/mdadm.conf`. Le manuel demande de vérifier cette sortie avant utilisation, car elle peut contenir des détails indésirables.

Trois fichiers indiquent l'état de l'ensemble :

```
cat /proc/mdstat
cat /sys/block/md0/md/degraded
cat /sys/block/md0/md/sync_action
```

`/proc/mdstat` répertorie tous les périphériques md actifs et leur état ; mdadm le lit pour trouver les ensembles et suivre leur reconstruction. Chaque ensemble possède aussi un répertoire sous `/sys/block/`, où `degraded` contient le nombre de membres manquants : 0 pour un ensemble complet, 1 lorsqu'un disque manque. `sync_action` vaut `idle` lorsqu'il ne se passe rien, `resync` après une création ou un arrêt incorrect, `recover` lorsqu'un disque de secours à chaud récupère le contenu d'un membre perdu, et `check` ou `repair` pendant une vérification complète. Y écrire `check` lance cette vérification, qui lit tous les blocs de l'ensemble :

```
echo check > /sys/block/md0/md/sync_action
cat /sys/block/md0/md/mismatch_cnt
```

Ensuite, `mismatch_cnt` indique le nombre de secteurs réécrits, ou qui l'auraient été. La documentation du noyau précise que la plupart des niveaux travaillent par pages plutôt que par secteurs : le compteur peut donc dépasser le nombre d'erreurs réelles.

Il faut aussi que quelqu'un soit informé d'une panne. `mdadm --monitor` interroge les ensembles et transforme ses observations en événements, qu'il peut envoyer à une adresse électronique ou transmettre à un programme. Sur un système géré par systemd, le manuel désigne `mdmonitor.service` comme voie principale et `MAILADDR` dans `mdadm.conf` comme paramètre à renseigner pour les

courriels. Seuls cinq événements envoient un courriel : `Fail` , `FailSpare` , `DegradedArray` , `SparesMissing` et `TestMessage` . Le dernier sert à prouver que la chaîne fonctionne :

```
MAILADDR ops@example.com
```

```
mdadm --monitor --scan --oneshot --test
mdadm --detail --test /dev/md0
```

La première commande envoie un message de test pour chaque ensemble trouvé puis se termine. La deuxième convient à un script : avec `--test` , son code de sortie vaut 0 pour un ensemble sain, 1 si au moins un périphérique est en panne, 2 si les défaillances sont assez nombreuses pour rendre l’ensemble inutilisable et 4 si mdadm rencontre une erreur en recueillant des informations sur l’ensemble.

L’ensemble couvre un événement, la panne d’un disque, et rien d’autre. La réalisation sur l’automatisation des sauvegardes et des tests de restauration décrit une conception de récupération distincte : des instantanés restic chiffrés et dédupliqués vers le stockage d’un autre fournisseur par SFTP, des minuteurs systemd et un processus mensuel qui restaure un fichier connu dans un emplacement temporaire et compare ses octets. Il s’agit d’un test de restauration prévu dans la conception, pas d’une preuve d’exécution. Le relevé d’implémentation daté du 2026-10-27 indique que les minuteurs étaient désactivés faute de dépôt configuré et ne consigne aucune restauration réelle achevée.

Où les problèmes surviennent

Un ensemble est conçu pour un événement, la panne d’un disque membre, et la plupart de ses causes de perte de données se situent en dehors de ce cas.

L’ensemble est pris pour une sauvegarde. Un miroir écrit simultanément chaque modification sur les deux disques, y compris un fichier supprimé par erreur, une table supprimée par une mauvaise migration ou un volume chiffré par un rançongiciel. Tous les membres portent alors les mêmes dégâts. La

correction est une sauvegarde distincte conservée hors de l'hôte, avec des restaurations testées régulièrement : l'ensemble protège contre les pannes matérielles et la sauvegarde contre tout le reste.

Un ensemble dégradé passe inaperçu. Un ensemble dégradé continue de servir les données, sans différence apparente jusqu'à la panne du disque suivant. mdadm n'envoie aucun courriel sans adresse ou programme : avec `--scan` et aucun des deux configuré, le manuel précise qu'il ne surveille rien. `DegradedArray` ne couvre en outre que les ensembles déjà dégradés lorsque mdadm les découvre ; un disque qui tombe en panne plus tard produit un événement `Fail` . La correction consiste à définir `MAILADDR` , vérifier que `mdmonitor.service` fonctionne et envoyer un message `--test` pour prouver que le courriel arrive.

Le deuxième disque tombe en panne pendant la reconstruction. Une reconstruction reconstitue le disque manquant à partir de tous les autres, et le modèle du rapport perd les données précisément lorsqu'un autre disque du groupe tombe en panne pendant ce temps de réparation. Laisser un disque défaillant en place pendant une semaine étend cette fenêtre à une semaine. La correction consiste à la réduire : conservez un disque de secours à chaud dans l'ensemble, que mdadm peut configurer à la création avec `--spare-devices` , remplacez rapidement les disques défaillants et exécutez un `check` planifié pour repérer un disque défectueux tant que l'ensemble peut encore corriger les erreurs.

Les disques ne sont pas indépendants. Le calcul du MTTF suppose des pannes indépendantes, et le rapport cite lui-même un séisme ou une surtension comme contre-exemples. Deux disques dans un boîtier partagent une alimentation, un contrôleur, une pièce et un bâtiment. La correction consiste à considérer le chiffre calculé comme une borne supérieure et à couvrir les causes communes par quelque chose d'extérieur au boîtier, par exemple une copie sur un autre site.

Un ensemble RAID 5 ou 6 est à la fois non propre (« dirty ») et dégradé. Si la machine s'arrête incorrectement alors qu'un membre manque, la parité n'est plus fiable et certains blocs ne peuvent pas être reconstruits : la documentation du noyau avertit d'une corruption indétectable, et md refuse de démarrer l'ensemble. Le forcer avec `mdadm --assemble --force` , ou avec le paramètre de module `start_dirty_degraded` pour un système de fichiers racine, revient à accepter ce risque. La correction consiste à choisir à l'avance une politique de cohérence : md peut journaliser les écritures sur un périphérique dédié pour les niveaux 4, 5 et 6, et propose pour le niveau 5 un journal partiel de parité qui,

selon la documentation, ferme la fenêtre d'incohérence des écritures. Lorsque le démarrage forcé est la seule issue, considérez les données comme suspectes et comparez-les à une sauvegarde.

RAID 0 est compté comme une protection. md indique que les ensembles raid0 et linear ne possèdent aucune information de redondance, et le moniteur de mdadm les signale par `Wrong-Level`, car ils ne prennent en charge aucune des opérations de panne, de secours et de resynchronisation qu'il surveille. Rien ne permet de reconstruire un membre manquant. La correction consiste à réserver le niveau 0 aux données jetables ou recréables et à ne jamais le compter comme du RAID au sens du rapport de 1987.

Un contrôleur matériel masque les disques. Derrière une carte contrôleur, le système d'exploitation voit un seul périphérique logique, et interroger ce périphérique sur l'état des disques atteint le contrôleur, pas les disques. smartctl, par exemple, nécessite `-d megaraid,N` pour s'adresser au disque N derrière un contrôleur MegaRAID, ce qui, selon son manuel, couvre aussi les cartes Dell PERC. La correction consiste à surveiller chaque disque physique par l'interface du contrôleur lui-même, afin de remarquer un membre défaillant tant que l'ensemble est encore complet.

Pour aller plus loin

- Plaidoyer pour des ensembles redondants de disques peu coûteux (RAID) — la page du rapport technique de Berkeley, avec son résumé et le PDF.
- UCB/CSD-87-391, rapport complet (PDF) — le calcul du MTTF, les groupes de fiabilité et les cinq niveaux.
- Ensembles RAID, documentation du noyau Linux — les fichiers de md sous `/sys`, les ensembles non propres et dégradés et les politiques de cohérence.
- mdadm(8) — création des ensembles, événements de surveillance, courriels et codes de sortie de `--detail --test`.
- smartctl(8), smartmontools — l'accès aux disques derrière les contrôleurs RAID avec l'option `-d`.

Glossaire

Choisissez selon vos contraintes réelles

Présentez le parcours de déploiement, les risques opérationnels et les personnes d'astreinte. Nous pouvons définir un changement avec des vérifications, une procédure de reprise et une responsabilité claire.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce que RAID ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

welcome@engineer.company +45 50 10 61 84 Copenhague, Danemark

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par **D.**

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/raid/