Particionamiento avanzado RAID por software

1. Definiciones

El **RAID** (*Redundant Array of Independent Disks*) ha sido definido por la Universidad de Berkeley en 1987 con el doble objetivo de reducir los costes y aumentar la fiabilidad del almacenamiento de datos. El objetivo es compaginar varios pequeños discos físicos independientes en una matriz (array: tabla, conjunto, fila, matriz) de discos cuya capacidad supera la del SLED (*Single Large Expensive Drive*). Hoy en día, la palabra *Independent* ha remplazado a *Inexpensive*. Una matriz aparece como una unidad lógica de almacenamiento único.

El **MTBF** (*Mean Time Between Failure* - intervalo entre fallos) del conjunto es igual al MTBF de un disco individual dividido por el número de discos en el conjunto y por lo tanto, teóricamente, una solución RAID puede no ser la adecuada en sistemas críticos. Afortunadamente, un sistema RAID es tolerante a los fallos gracias a que almacena de manera redundante su información según varios métodos:

- **RAID-0**, llamado **stripe mode**: al menos dos discos forman un único volumen. En principio, los dos discos tienen el mismo tamaño. Se fraccionará y efectuará cada operación de lectura/escritura en cada uno de los discos. Por ejemplo, se escribirán 4 KB en el disco 0, 4 KB en el disco 1, 4 KB en el disco 2, luego 4 KB en el disco 0, etc. Así, se aumentan las prestaciones, ya que se efectúan en paralelo las operaciones de lectura y escritura sobre los discos. Si N es el número de discos y P la velocidad de transferencia, la velocidad de transferencia del volumen RAID es, en principio, próxima a N*P mbps. El RAID-0 no tiene ninguna redundancia. En caso de fallo de uno de los discos, es probable que se pierda el conjunto de datos.
- **RAID-1**, llamado **mirroring**: primer modo redundante. Se puede utilizar a partir de dos discos o más con posibles discos de emergencia (*Spare Disk*). Se duplica cada información escrita en un disco sobre los demás. Si N-1 discos del RAID caen, los datos siguen intactos. Si hay un disco de emergencia, en caso de fallo, se reconstruye automáticamente y se sustituye el disco que falla. Las

prestaciones en escritura pueden ser malas: escritura en N discos al mismo tiempo, con el riesgo de saturar el controlador de disco y el bus. Los discos deben estar preferentemente en dos controladores diferentes. Las prestaciones en lectura son buenas, porque RAID emplea un algoritmo que puede leer los datos en cada disco (ya que son idénticos).

- **RAID-5**: RAID con algoritmo distribuido de paridad. Es el modo más utilizado, ya que ofrece la mejor relación entre el número de discos, el espacio disponible y la redundancia. Hacen falta al menos tres discos con posibles discos de emergencia. Hay paridad en cada uno de los discos. El tamaño final es el de N-1 discos. El RAID-5 sobrevive a un fallo de disco. En este caso, si hay un disco de emergencia, será reconstruido automáticamente. Las prestaciones en lectura son equivalentes a las del RAID-0, mientras que en escritura dependen del algoritmo empleado, así como de la memoria de la máquina.

2. Precauciones y consideraciones de uso

a. Disco de emergencia

Un disco de emergencia (*Spare Disk*) no forma parte integrante de una matriz RAID mientras no se averíe un disco. Si eso sucede, se marca el disco como defectuoso y el primer disco Spare toma el relevo. Pase lo que pase, conviene cambiar el disco averiado lo antes posible y volver a construir el RAID.

b. Disco averiado

Un disco averiado (*Faulty Disk*) es un disco que ha sido reconocido como defectuoso o como erróneo por el RAID. En este caso, RAID utiliza el primer disco SPARE para reconstruir su matriz. Los discos faulty pertenecen siempre a la matriz, pero están desactivados.

c. Boot

No se debe ubicar en una matriz RAID la partición boot (la que contiene el núcleo, la configuración del bootloader, los archivos imágenes de discos) en caso de utilizar la

versión 1 de GRUB: este cargador del sistema operativo es incapaz de montar particiones RAID. GRUB2 es capaz de arrancar de un RAID por software, empleando una configuración adaptada.

d. Swap

Se puede instalar una partición de intercambio (swap) sobre un RAID, pero no suele ser útil en los casos comunes. Linux es capaz de equilibrar el uso del swap sobre varios discos/particiones individuales. En este caso, declare n swaps en `/etc/fstab` con la misma prioridad.

```
/dev/sda2 swap swap defaults,pri=1 0 0
/dev/sdb2 swap swap defaults,pri=1 0 0
/dev/sdc2 swap swap defaults,pri=1 0 0
```

Sin embargo, en caso de necesidad de alta disponibilidad, es posible la swap en el RAID.

e. Periféricos

El sistema reconoce una matriz RAID como un periférico de tipo bloque, al igual que cualquier disco físico. Así, un RAID puede estar constituido por discos, particiones (en general, se crea una única partición en cada disco). El bus no tiene importancia alguna: puede construir una matriz RAID con discos SCSI e IDE mezclados. Asimismo, se puede construir un RAID sobre otras matrices RAID, por ejemplo RAID-0+1 (2x2 discos sobre RAID-1, formando sobre RAID-0 una nueva matriz las dos resultantes). Los periféricos RAID tienen la forma:

```
/dev/md0
/dev/md1
```

f. IDE y SATA

Si los discos IDE han sido mucho tiempo el SCSI del pobre (hardware de menor calidad, lentitud, falta de fiabilidad), hoy por hoy esto ya no es cierto. Los últimos modelos que

encontrará son totalmente idénticos a los discos SCSI, si exceptuamos el controlador. De hecho, puede montar configuraciones RAID en IDE a un coste razonable. Sin embargo, hay que recordar una regla: **un solo disco IDE por bus IDE**.

En la práctica, esto corresponde a poner un disco por bus, nada más. La razón estriba en que un bus IDE sobrevive al fallo en un disco, pero si lo que falla es el bus mismo se pierden todos los discos conectados al bus y con ellos la matriz RAID. La compra de tarjetas IDE adicionales (a bajo precio) permite compensar el problema de fiabilidad (dos discos por tarjeta).

El principio es el mismo que para los discos SATA. En teoría es más simple, ya que sólo puede conectar un disco por cable. Sin embargo, si lee el manual de su placa base o el del controlador, le recordarán que a menudo los puertos SATA van por parejas. Además, si su placa contiene varios controladores, puede ser una buena idea separar los discos: uno por controlador, o como con IDE, añadir un controlador SATA adicional por PCI-Express.

g. Hot Swap

- **IDE**: **¡Nunca desenchufe un disco IDE en caliente!** Es la mejor manera de destruir el disco, si ya no fuese el caso, y de destruir el controlador IDE (y, en ocasiones, la placa base o la IDE adicional). El disco IDE no está diseñado para ello.
- **SCSI**: los controladores SCSI no están previstos para el Hot Swap, pero en teoría deberían funcionar de todas maneras, si el disco es idéntico física y lógicamente.
- **SATA**: Se equipara el SATA con un SCSI. La especificación SATA en versión 2 soporta teóricamente el Hot Swap. Sin embargo, la mayoría de los controladores actuales o no implementan o implementan mal esta posibilidad; de ahí que haya riesgos de bug o de quemar su controlador. Compruébelo en la documentación del fabricante de su placa base (chipset). Es importante verificar si el conector de alimentación es un verdadero conector SATA completamente cableado.
- **SCA**: son discos SCSI específicos. Consulte el documento «Software RAID Howto».

3. RAID con mdadm

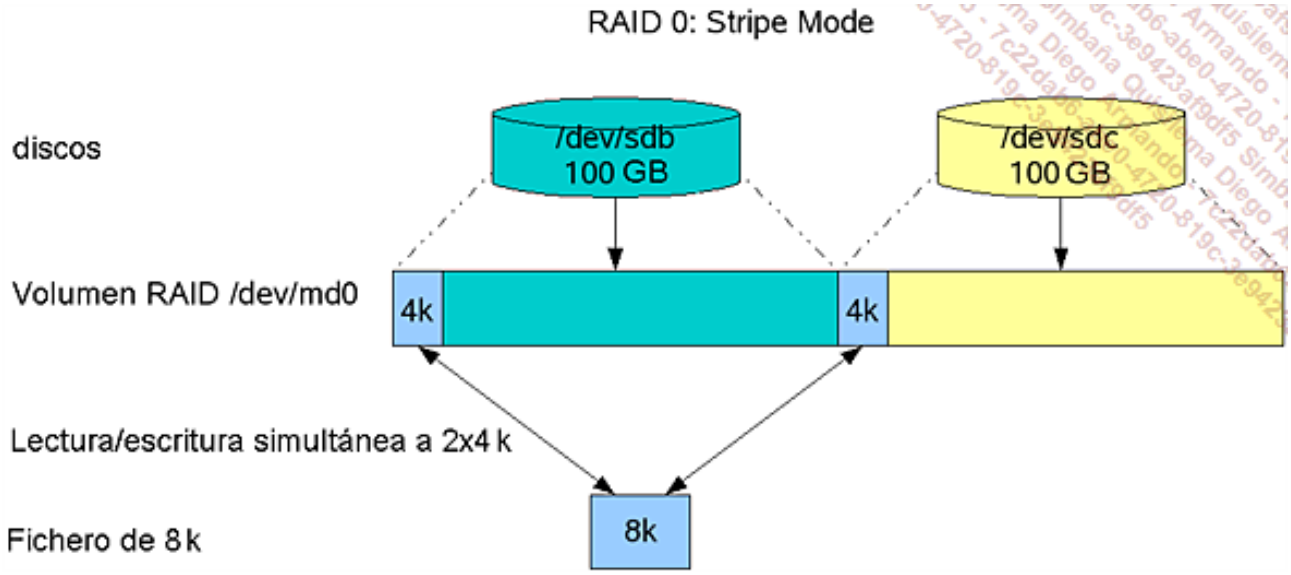
a. Preparación

La herramienta **mdadm** permite efectuar todas las operaciones. Su archivo de configuración es `/etc/mdadm.conf` o `/etc/mdadm/mdadm.conf` .

Con el fin de crear matrices RAID, es necesario que las particiones que servirán para crear la matriz sean de tipo **0xFD** (Linux RAID autodetect) para un particionado de tipo MBR, o **0xFD00** para un particionado de tipo GPT. Las particiones deben estar físicamente en discos diferentes, pero para hacer pruebas el soporte RAID autoriza particiones en un mismo disco. En ese caso, deberá tener en cuenta que las particiones dispongan del mismo tamaño.

b. Creación

RAID-0



Fundamentos de RAID 0: Striping

O sea, dos particiones `/dev/sdb1` y `/dev/sdc1` . Va a crear una partición RAID-0, ensamblaje de estas dos particiones.

```
# mdadm --create /dev/md0 --level=raid0 --raid-devices=2 /dev/sdb1 /dev/sdc1
```

`--create`	Crea un RAID.
`/dev/md0`	Nombre del archivo de tipo bloque que representa la matriz RAID.
`--level`	Tipo de RAID que se va a crear: `0`, `raid0` y `stripe` para RAID0.

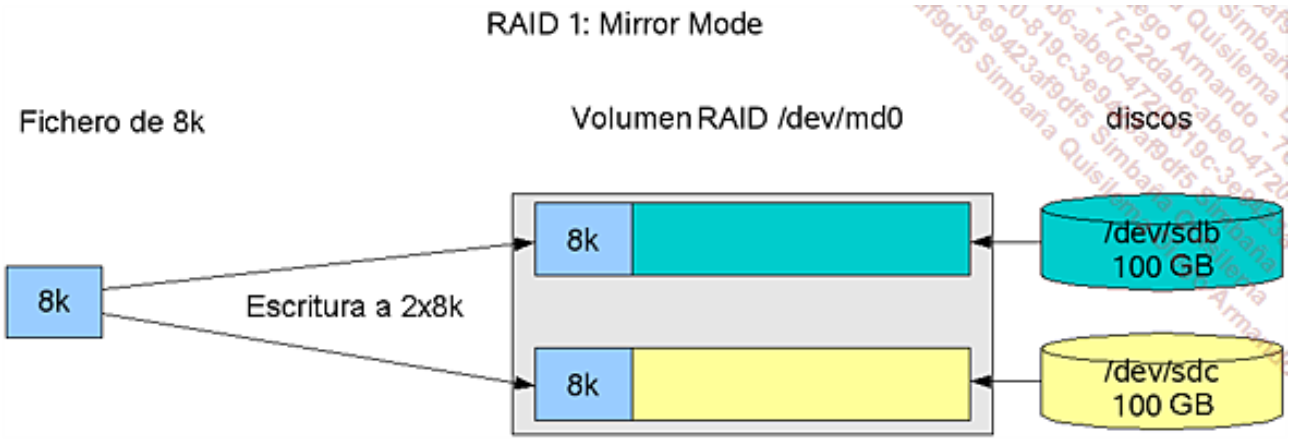
`linear` no es RAID0 (llenado poco a poco).

`--raid-devices`	Número de particiones utilizadas para crear la matriz.
`/dev/sdb1, /dev/sdc1`	Particiones que forman la matriz, según el número indicado en `--raid-devices` .

Sólo queda instalar el sistema de archivos en el disco RAID:

```
# mkfs -t ext4 /dev/md0
```

RAID-1



El principio es el mismo. Esta vez, tendrá que añadir una partición de emergencia **`/dev/sdd1`** .

```
# mdadm --create /dev/md0 --level=raid1 --raid-devices=2 /dev/sdb1
/dev/sdc1 --spare-devices=1 /dev/sdd1
```

`--level 1`, `mirror` o `raid1` son valores aceptados para un RAID-1.

`--spare-devices` número de discos de emergencia para utilizar.

`/dev/sdd1` partición o particiones de emergencia, según el número indicado en `-spare-devices` .

Luego:

```
# mkfs -t ext4 /dev/md0
```

RAID-0+1

Hacen falta al menos cuatro particiones. Debe crear dos matrices RAID-1, que va a agrupar en una sola matriz RAID-0.

```
# mdadm --create /dev/md0 --level=raid1 --raid-devices=2 /dev/sdb1
/dev/sdc1
```

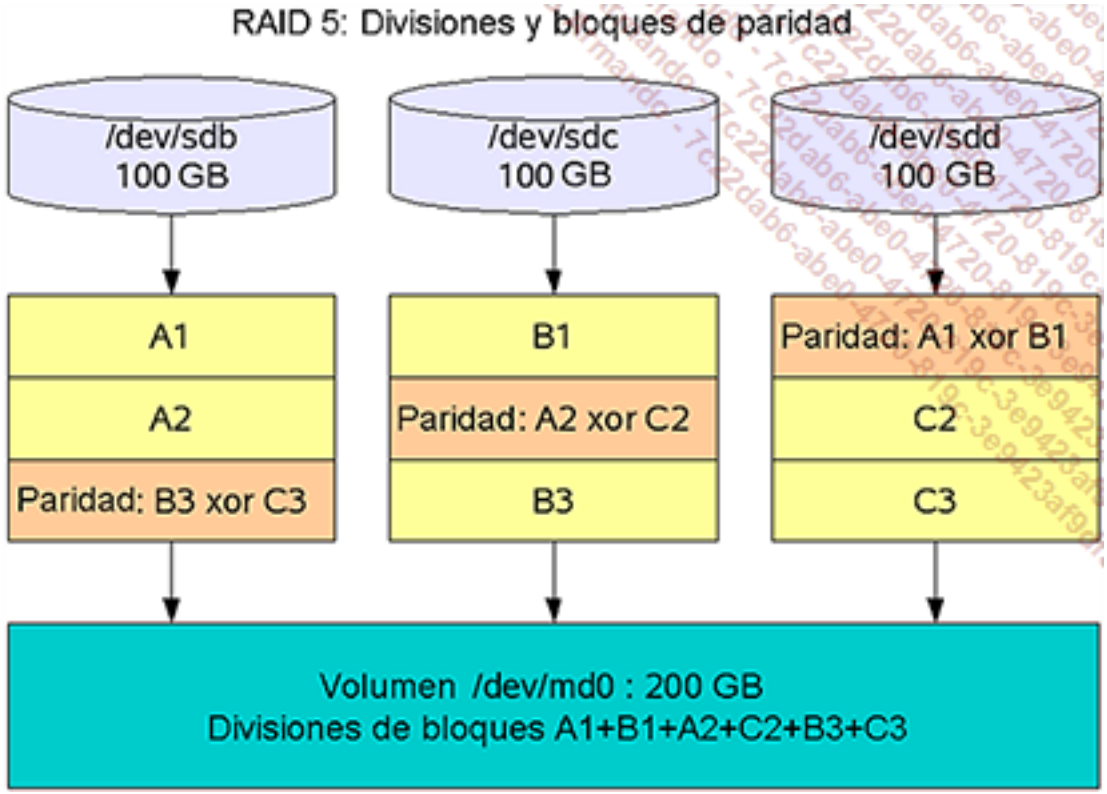
```
# mdadm --create /dev/md1 --level=raid1 --raid-devices=2 /dev/sdd1
/dev/sde1
```

```
# mdadm --create /dev/md2 --level=raid0 --raid-devices=2 /dev/md0
/dev/md1
```

Luego:

```
# mkfs -t ext4 /dev/md2
```

RAID5



El RAID va a emplear tres discos de datos /dev/sdb1, /dev/sdc1, /dev/sdd1 y un disco de emergencia /dev/sde1.

```
# mdadm --create /dev/md0 --level=raid5 --raid-devices=3 /dev/sdb1
/dev/sdc1 /dev/sdd1



--spare-devices=1 /dev/sde1
```

Luego se instala el sistema de archivos:

```
# mkfs -t ext4 /dev/md2
```

c. Guardar la configuración

Para facilitar la tarea de la herramienta **mdadm**, puede crear (no es obligatorio) el archivo de configuración `/etc/mdadm.conf` . Se puede crear este archivo manualmente, pero la herramienta **mdadm** sabe generarlo. Es preferible hacerlo DESPUÉS de la creación de las matrices RAID.

```
# echo "DEVICE partitions" > /etc/mdadm.conf



# mdadm --detail --scan>> /etc/mdadm.conf
```

4. Estado del RAID

El archivo virtual `/proc/mdstat` contiene información sobre el RAID. En este archivo puede ver el detalle de un RAID, en particular si alguno de los volúmenes de la matriz está averiado (Faulty).

```
Personalities : [linear] [multipath] [raid0] [raid1] [raid6] [raid5] [raid4] [raid10]
md0 : active raid1 sdd1[2](S) sdc1[1] sdb1[0]
```

```
      203136 blocks super 1.2 [2/2] [UU]

unused devices: <none>
```

El comando **watch** permite comprobar un estado en tiempo real:

```
# watch cat /proc/mdstat
```

También puede utilizar **mdadm** con el parámetro `--detail`:

```
/dev/md0:
        Version :1.2
  Creation Time : Sun Feb23 16:56:49 2020
     Raid Level : raid1
     Array Size : 203136 (198.38 MiB 208.01 MB)
  Used Dev Size : 203136 (198.38 MiB 208.01 MB)
   Raid Devices : 2
  Total Devices : 3
    Persistence : Superblock is persistent

    Update Time : Sun Feb 23 16:56:52 2020
          State : clean
 Active Devices : 2
Working Devices : 3
 Failed Devices : 0
  Spare Devices : 1

Consistency Policy : resync

           Name : ubuntu:0 (local to host ubuntu)
           UUID : 7b2cbaee:eccefa2a:c66b1fc7:7ec7432a
         Events : 17

  Number   Major   Minor   RaidDevice State
     0       8       17        0      active sync   /dev/sdb1
     1       8       33        1      active sync   /dev/sdc1
```

```
       2     8     49        -  spare  /dev/sdd1
```

Cabe señalar que, con este último comando, puede obtener muchos más detalles, en particular cuáles son los discos «spare» y «faulty».

5. Simular una avería

Ahora va a simular une avería en **`/dev/sdb1`** :

```
# mdadm /dev/md0 -f /dev/sdb1
```

```
mdadm: set /dev/sdb1 faulty in /dev/md0
```

Mire el estado del RAID en **`/proc/mdstat`** durante la ejecución:

```
md0 : active raid1 sdd1[2] sdc1[1] sdb1[0](F)
      203136 blocks super 1.2 [2/2] [U_]
      [=>.....................] recovery = 8.8% (9216/104320) finish=0.1min
speed=9216K/sec
```

Ha aparecido una «(F)» cerca de `sdb1`, lo que indica un disco Faulty. Observamos también que, de los dos discos, uno tiene una avería y que el RAID reconstruye su matriz con el spare disk. Después de la ejecución, obtenemos:

```
md0 : active raid1 sdd1[2] sdc1[1] sdb1[0](F)
      203136 blocks super 1.2 [2/2] [UU]
```

El RAID está reconstruido y funciona de maravilla.

```
# mdadm --detail /dev/md0
...
        State : clean
 Active Devices : 2
Working Devices : 2
 Failed Devices : 1
  Spare Devices : 0

    Number   Major   Minor   RaidDevice   State
       2       8      49          0      active sync  /dev/sdd1
       1       8      33          1      active sync  /dev/sdc1
       0       8      17          -        faulty  /dev/sdb1
...
```

El disco Faulty es realmente **`/dev/sdb1`** ; **`/dev/sdc1`** lo sustituyó como disco de emergencia. Así, el disco de emergencia se convierte en un disco RAID de la matriz.

6. Sustituir un disco

Ya que **`/dev/sdb1`** tiene un problema, lo va a sustituir. Sáquelo con `-r (o --remove)` :

```
# mdadm /dev/md0 -r /dev/sdb1
mdadm: hot removed /dev/sdb1 from /dev/md0


# cat /proc/mdstat
md0 : active raid1 sdd1[2] sdc1[1]
      203136 blocks super 1.2 [2/2] [UU]
```

Constate que `sdb1` ha desaparecido. Puede apagar la máquina y luego sustituir el disco defectuoso. Encienda de nuevo la máquina y vuelva a particionar el disco correctamente. Sólo falta añadir de nuevo el disco arreglado en la matriz RAID con `-a (--add)` :

```
# mdadm /dev/md0 -a /dev/sdb1
mdadm: added /dev/sdb1

# cat /proc/mdstat
Personalities : [raid1]
md0 : active raid1 sdb1[3](S) sdd1[2] sdc1[1]
      203136 blocks super 1.2 [2/2] [UU]
...
```

El disco `sdb1` aparece de nuevo. Vea el detalle:

```
# mdadm --detail /dev/md0
...
        State : clean
 Active Devices : 2
Working Devices : 3
 Failed Devices : 0
  Spare Devices : 1


    Number   Major   Minor   RaidDevice  State
       2       8      49          0      active sync  /dev/sdd1
       1       8      33          1      active sync  /dev/sdc1
       3       8      17          -      spare        /dev/sdb1
...
```

Se ha añadido de nuevo el disco **`/dev/sdb1`** ¡y el sistema lo ha convertido en el nuevo disco de emergencia!

7. Apagado y puesta en marcha manual

Puede apagar de manera puntual una matriz RAID con `-S (--stop)` DESPUÉS de haber desmontado el periférico:

```
# mdadm --stop /dev/md0
```

La matriz RAID se pone en marcha de nuevo con `-As (--assemble -scan)`. Eso implica que el archivo **`/etc/mdadm.conf`** está correctamente configurado (`--scan` es una opción que busca la información en ese archivo).

```
# mdadm --assemble --scan /dev/md0
```

Si el RAID no arranca de nuevo, puede intentar con `-R (--run)`: es probable que falte un disco o que no se haya acabado aún una reconstrucción en curso:

```
# mdadm --run /dev/md0
```

8. Destrucción del RAID

Después de haber parado la matriz RAID como se vio anteriormente, debe limpiar los metadatos de los discos que la componen de la manera siguiente:

```
# mdadm --zero-superblock /dev/sdb1
# mdadm --zero-superblock /dev/sdc1
```