Configuración de los discos RAID

La tecnología RAID (*Redundant Array of Independent Disks*) permite combinar diferentes dispositivos para que sean vistos como un solo espacio de almacenamiento para las aplicaciones. De esta manera se puede mejorar el tiempo de acceso y/o la fiabilidad de los dispositivos de almacenamiento. Las diferentes técnicas empleadas se definen con respecto al nivel de RAID usado. Los niveles más corrientes son RAID 0, 1 y 5; los que se estudiarán en el marco de la certificación.

Se puede administrar el RAID en modo hardware, con los controladores de discos especializados o en modo software, desde el sistema operativo.

Linux implementa un piloto de gestión de software del RAID, el piloto `md` (*Multiple Device piloto*), que gestiona los niveles más corrientes de RAID, 0, 1 y 5, y que debe ser estudiado en el marco de la certificación.

Otras soluciones pueden implementarse para gestionar el software RAID en Linux: RAID LVM o RAID directamente soportado por el gestor de sistema de archivos ZFS o Btrfs.

1. Los principales niveles de RAID

a. RAID 0

RAID 0 (agregación de bandas, *striping*) combina distintos discos en un solo conjunto. Los bloques de datos se reparten en bandas de tamaño idéntico que están repartidas uniformemente en los diferentes discos. Las operaciones de entrada/salida pueden ser, por lo tanto, muy rápidas, ya que los controladores de los discos las pueden efectuar de manera simultánea.

Sin embargo, la fiabilidad del conjunto es bastante baja ya que al perder un disco perderá el conjunto de los datos. No hay redundancia en los datos almacenados y la coherencia de los volúmenes lógicos se destruye en el caso de fallo en un disco.

El espacio de almacenamiento útil de un conjunto RAID 0 es igual a la capacidad útil del más pequeño de los discos multiplicado por el número de discos que lo componen, ya que no hay redundancia de datos y que las bandas de datos están repartidas de manera uniforme en los discos (cada disco tiene que tener el mismo número de bandas).

Ventajas:

- Rapidez de lectura y escritura del conjunto de los bloques.
- Uso óptimo del espacio de los discos, siempre y cuando los discos sean del mismo tamaño.

Inconvenientes:

- No hay redundancia de datos, por lo tanto, no hay tolerancia frente a fallos.
- La pérdida de un disco compromete el conjunto de los datos almacenados, la fiabilidad del conjunto es igual a la fiabilidad del menos fiable de los discos utilizados.

b. RAID 1

RAID 1 (discos en espejo, *mirroring*) combina distintos discos en un solo conjunto. Cada bloque de datos útil está escrito en cada uno de los discos. Esta redundancia asegura una fiabilidad excelente al conjunto, mayor cuanto mayor sea el número de discos. Mientras quede un disco operativo, los datos estarán intactos y mientras el controlador de ese disco funcione, esos datos seguirán estando accesibles.

Las operaciones de lectura pueden ser más rápidas, porque los controladores las pueden efectuar simultáneamente.

El espacio de almacenamiento útil del conjunto RAID 1 es igual a la capacidad útil del más pequeño de los discos.

Ventajas:

- Excelente tolerancia frente a fallos, proporcional al número de discos combinados (y al número de controladores de discos para la accesibilidad).
- Buen rendimiento en lectura.

Inconvenientes:

- El espacio de disco necesario es al menos dos veces el tamaño del espacio de disco útil.
- Puede haber un impacto en el rendimiento en escritura, incluso si en general las escrituras se hacen simultáneamente en los diferentes discos.

c. RAID 5

RAID 5 (agregación en bandas con paridad) combina al menos tres discos en un solo conjunto. Los bloques de datos están repartidos en bandas de tamaño idéntico, repartidas uniformemente en los diferentes discos, excepto en uno de ellos. Para cada conjunto de bandas, una banda de paridad estará calculada y escrita en el disco restante. La ubicación del conjunto de paridad estará repartida alternativamente en los discos.

En caso de pérdida de una banda de datos, la banda de paridad permitirá reconstituirla, asegurando la tolerancia frente a fallos. Pero este mecanismo solamente es eficaz en el caso en que solamente haya un disco que no esté accesible. Si hubiera dos discos o más que no estuvieran accesibles, habría pérdida de datos en el conjunto de los discos. Mientras que un disco no esté operativo, no habrá tolerancia frente a fallos para las nuevas escrituras. Es por esto por lo que el conjunto de discos en RAID 5 integra generalmente un disco de emergencia (*spare disk*), que solo se utiliza para reemplazar un disco defectuoso.

Cuando el disco defectuoso haya sido reparado o reemplazado, hay que reconstruir el conjunto RAID 5 reconstituyendo los datos y las bandas de paridad para escribirlos en el disco de reemplazo.

Las operaciones de lectura pueden ser muy rápidas, ya que las efectúan diferentes controladores de discos. Las escrituras pueden ser lentas, a causa del cálculo y de la escritura de la banda de paridad.

El espacio de almacenamiento útil de un conjunto RAID 5 es igual a la capacidad del más pequeño de sus discos, multiplicada por el número de discos que lo componen, menos 1 a causa de las bandas de paridad y menos 2 si hay un disco de emergencia (*spare*).

Ventajas:

- Tolerancia frente a fallos limitada a un disco. Mientras haya un disco en fallo, ya no habrá más tolerancia frente a fallos, excepto en el caso de que exista un disco de emergencia.

Inconvenientes:

- Una parte del espacio del disco no puede ser utilizado para los datos.
- El rendimiento en escritura puede verse impactado por el cálculo de la paridad.

2. Configuración del RAID

El piloto md es un módulo del núcleo que implementa el software RAID en un conjunto de dispositivos de almacenamiento, discos duros completos y/o particiones de discos duros.

El comando `mdadm` permite configurar volúmenes RAID y administrarlos. Forma parte del paquete `mdadm`.

a. Creación de un volumen RAID

Un volumen RAID está compuesto por distintos espacios de almacenamiento que pueden ser discos duros enteros o particiones de disco duros.

La creación de un volumen RAID se hace con la opción `-C` del comando `mdadm`. Hay que especificar el nombre del nuevo volumen o su número, el nivel de RAID que se quiera implementar y la lista de los espacios de almacenamiento que se le va a destinar.

El archivo de configuración del comando, generalmente `/etc/mdadm /mdadm.conf`, es facultativo en las versiones recientes y no se crea durante la instalación del paquete.

Sintaxis

```
mdadm -C ArchivoEspecialVol -l|--level=Nivel -n|--raid-devices=NúmeroDevRaid
[ -x|--spare-devices=NúmEmergencia ] ArchivoEspecial1 ... ArchivoEspecialN
```

Principales parámetros

`-C ArchivoEspecialVol`	Archivo especial del volumen RAID creado.
`-l|--level=Nivel`	Nivel de RAID.
`-n|--raid-devices=NœmeroDevRaid`	Número de espacios de almacenamiento activos.
`-x|--spare-devices=NœmEmergencia`	Número de espacios de almacenamiento de emergencia.
`ArchivoEspecial1 ...ArchivoEspecialN`	Espacios de almacenamiento.

Descripción

La opción `-C` crea un nuevo volumen RAID. El archivo especial que se le asociará, `ArchivoEspecialVol`, se encuentra generalmente bajo la forma `/dev/mdX`, donde `X` es un número, pero esto no es obligatorio.

La opción `-n` especifica el número de espacios de almacenamiento que utilizará el volumen. Debe ser igual o superior al número de elementos de la lista `ArchivoEspecial1..., ArchivoEspecialN` menos el número de espacios de almacenamiento de emergencia, indicado con la opción `-x`.

Una vez el volumen RAID creado, este puede ser usado inmediatamente. Es visto como un dispositivo en modo bloque, podemos por lo tanto crear un sistema de archivos o hacer de

él un volumen físico LVM.

Ejemplos

Se utilizan dos particiones de discos duros, `/dev/sda4` y `/dev/sdd1`, para crear un volumen `RAID` de nivel `1` (espejo). Como las particiones contienen sistemas de archivos y son de tamaños diferentes, el comando muestra una advertencia y solicita una confirmación:

```
mdadm -C /dev/md0 -l 1 -n 2 /dev/sda4 /dev/sdd1
mdadm: /dev/sda4 appears to contain an ext2fs file system
    size=5237760K  mtime=Thu Jan  1 01:00:00 1970
mdadm: Note: this array has metadata at the start and
  may not be suitable as a boot device.  If you plan to
  store '/boot' on this device please ensure that
  your boot-loader understands md/v1.x metadata, or use
  --metadata=0.90
mdadm: largest drive (/dev/sdd1) exceeds size (5232640K) by more than 1%
Continue creating array? y
mdadm: Defaulting to version 1.2 metadata
mdadm: array /dev/md0 started.
```

Se ha creado el volumen RAID:

```
ls -l /dev/md0
brw-rw---- 1 root disk 9, 0 ago 19 17:49 /dev/md0
El archivo /dev/md0 es un archivo especial de bloques.
```

El comando `blkid` muestra información de los dos espacios de almacenamiento que componen el volumen RAID:

```
blkid /dev/sda4 /dev/sdd1
/dev/sda4: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="679cb515-
2f5c-6078-6829-4b9f0f928907" LABEL="beta:0" TYPE="linux_raid_member"
PARTUUID="12deb3a0-04"
/dev/sdd1: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="e77f5f03-
448a-1b48-feb6-9bedc62e40b5" LABEL="beta:0" TYPE="linux_raid_member"
```

```
PARTUUID="c3072e18-01"
```

Las dos particiones son de tipo RAID Linux y han recibido una etiqueta `beta:0`, `beta` *es el nombre de la máquina.*

b. Tipo de partición RAID

Cuando se crea una partición en un disco duro, se le puede atribuir un tipo bajo la forma de un valor numérico predefinido. Para las particiones que usan el modo de particionamiento tradicional en el mundo del PC, llamado MBR (*Master Boot Record*), los tipos principales son:

`0x82`	Linux swap
`0x83`	Linux
`0xFD`	Linux RAID auto
`0xE8`	LUKS (cifrado)
`0x07`	NTFS
`0x0C`	FAT32

Para las particiones que usan el modo de particionamiento más reciente y general, GPT (*GUID Partition Table*, con *GUID= Globally Unique IDentifier*), el tipo `0xFD00` está reservado para las particiones RAID Linux.

En las antiguas versiones de RAID Linux, el tipo de partición `0xFD` (Linux RAID auto) servía en algunos casos (en particular para la carga inicial del núcleo) para detectar las particiones que componen los conjuntos RAID. Hoy día, este tipo de partición ya no es

necesaria para que se activen los volúmenes RAID.

Se puede, sin embargo, fijar con este tipo las particiones usadas en los volúmenes RAID, para reconocerlas cuando se usen herramientas de gestión de particiones u algunas otras herramientas.

Para cambiar el tipo de una partición existente, se pueden usar distintos comandos.

Ejemplo

Para cambiar el tipo de la partición MBR `/dev/sdc1` a `RAID`:

```
sfdisk --id /dev/sdc1 fd
```

o con el comando interactivo `fdisk`:

```
fdisk /dev/sdc
t
Partition number (1,2, default 2): 1
Partition type (type L to list all types): fd
```

c. Estado de un volumen RAID

La opción `-D` del comando `mdadm` muestra las características y el estado de un volumen RAID.

Sintaxis

```
mdadm -D|--detail ArchivoEspecialVol
```

Descripción

El comando muestra todas las características del volumen RAID indicado: su nivel de RAID, su estado así como el de sus componentes y el estado de estos.

Ejemplo

Características y estado del volumen RAID nivel `1 /dev/md0` :

```
mdadm -D /dev/md0
/dev/md0:
        Version: 1.2
  Creation Time: Tue Mar 10 18:08:05 2020
     Raid Level: raid1
     Array Size: 5232640 (4.99 GiB 5.36 GB)
  Used Dev Size: 5232640 (4.99 GiB 5.36 GB)
   Raid Devices: 2
  Total Devices: 2
    Persistence: Superblock is persistent

    Update Time: Tue Mar 10 18:15:51 2020
          State: clean, resyncing
 Active Devices: 2
Working Devices: 2
 Failed Devices: 0
  Spare Devices: 0

Consistency Policy: resync

   Resync Status: 50% complete

           Name: beta:0  (local to host beta)
           UUID: 760b8ab1:90415cd1:7e1e1308:fa7e75fd
         Events: 8

    Number   Major   Minor   RaidDevice State
       0       8        4        0      active sync   /dev/sda4
       1       8       49        1      active sync   /dev/sdd1
```

Se puede ver que el volumen RAID está compuesto por dos particiones, `/dev/sda4` y `/dev/sdd1` . El tamaño útil del volumen es de 5,36 GB. Su estado es `resyncing` (sincronizando), porque el volumen acaba de crearse y el piloto `md` copia una partición encima de la otra para establecer el espejo.

Un poco más tarde, el mismo comando da este resultado:

mdadm -D /dev/md0

```
/dev/md0:
        Version: 1.2
  Creation Time: Tue Mar 10 18:08:05 2020
     Raid Level: raid1
     Array Size: 5232640 (4.99 GiB 5.36 GB)
  Used Dev Size: 5232640 (4.99 GiB 5.36 GB)
   Raid Devices: 2
  Total Devices: 2
    Persistence: Superblock is persistent

    Update Time: Tue Mar 10 18:23:32 2020
          State: clean
 Active Devices: 2
Working Devices: 2
 Failed Devices: 0
  Spare Devices: 0

Consistency Policy: resync

           Name: beta:0  (local to host beta)
           UUID: 760b8ab1:90415cd1:7e1e1308:fa7e75fd
         Events: 17

    Number   Major   Minor   RaidDevice State
       0       8        4        0      active sync   /dev/sda4
       1       8       49        1      active sync   /dev/sdd1
```

Podemos observar ve que el volumen RAID está en estado `clean`, la sincronización de las dos particiones ha terminado.

El archivo `mdstat` del sistema de archivos virtual `proc` provee dinámicamente la información de los volúmenes RAID y su sestado:

Ejemplo

cat /proc/mdstat

```
Personalities: [raid1]
md0: active raid1 sdd1[1] sda4[0]
    5232640 blocks super 1.2 [2/2] [UU]
    [===>.................]  resync = 15.7% (824064/5232640)
finish=12.9min speed=5672K/sec
unused devices: <none>
```

Se puede ver que el volumen `md0` está sincronizándose. Está compuesto por dos particiones: `sdd1` y `sda4`.

d. Paro y supresión de un volumen RAID

El paro de un volumen RAID se hace con la opción `-S` (*stop*) del comando `mdadm`.

Sintaxis

```
mdadm -S ArchivoEspecialVol
```

Descripción

El volumen RAID `ArchivoEspecialVol` no tiene que estar siendo utilizado (sistema de archivos montado).

Una vez que se haya desactivado el volumen RAID, los espacios de almacenamiento que lo constituían pueden volver a ser usados, pero primero hay que suprimir de su superbloque la información relativa al RAID, usando el comando siguiente:

```
mdadm --zero-superblock ArchivoEspecial
```

Ejemplo

Se desactiva un volumen RAID de nivel `1` y se "limpian" los espacios de almacenamiento que lo componen:

```
mdadm -S /dev/md0
mdadm: stopped /dev/md0
ls -l /dev/md0
ls: no se puede acceder a '/dev/md0': No existe el fichero o el directorio
```

El archivo especial del volumen RAID ya no existe, pero sus dos componentes todavía se consideran como formando parte de un volumen RAID:

```
blkid /dev/sda4 /dev/sdd1
/dev/sda4: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="679cb515-
2f5c-6078-6829-4b9f0f928907" LABEL="beta:0" TYPE="linux_raid_member"
PARTUUID="12deb3a0-04"
/dev/sdd1: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="e77f5f03-
448a-1b48-feb6-9bedc62e40b5" LABEL="beta:0" TYPE="linux_raid_member"
PARTUUID="c3072e18-01"
```

Se vacían los superbloques de la información RAID:

```
mdadm --zero-superblock /dev/sda4 /dev/sdd1
blkid /dev/sda4 /dev/sdd1
/dev/sda4: UUID="493a1a1c-0163-4193-a55e-2488445ec96a" TYPE="ext4"
PARTUUID="12deb3a0-04"
/dev/sdd1: LABEL="SWAP-emerg" UUID="30d42e87-3f2a-4cfe-906b-4ebf5a8138f2"
TYPE="swap" PARTUUID="c3072e18-01"
```

Las dos particiones han vuelto a encontrar sus características iniciales.

3. Explotación de un volumen RAID

Una vez que el volumen RAID haya sido creado, se utilizará como un espacio de almacenamiento clásico, a través de su archivo especial. En su interior se puede crear un sistema de archivos o integrarlo en un grupo de volúmenes LVM y después ponerlo a disposición de las aplicaciones de manera transparente.

a. Reemplazo de un espacio de disco

Si un disco miembro de un volumen RAID ya no está operativo y el volumen RAID soporta la tolerancia frente a fallos (RAID 1 o 5), el volumen sigue estando operativo y utilizable para las aplicaciones.

Una vez que el disco haya sido reemplazado, hay que reactivarlo dentro del volumen RAID, usando la opción `--add` del comando `mdadm`:

```
mdadm --add VolRAID ArchivoEspecialMiembro
```

b. Ejemplo de uso de un volumen RAID 1

Se crea un volumen RAID de nivel `1`, `/dev/md0`, combinando dos particiones de 8 GB aproximadamente en dos memorias USB distintas, `/dev/sdc1` y `/dev/sdd1`. Las dos particiones han sido creadas de tipo RAID Linux:

```
fdisk -l /dev/sdc /dev/sdd
Disco /dev/sdc: 29,9 GiB, 32078036992 bytes, 62652416 sectores
Unidades: sectores de 1 * 512 = 512 bytes
Tamaño de sector (lógico/físico): 512 bytes / 512 bytes
Tamaño de E/S (mínimo/óptimo): 512 bytes / 512 bytes
Tipo de etiqueta de disco: dos
Identificador del disco: 0xbeac1b4e

Disposit.  Inicio Comienzo    Final Sectores Tamaño Id Tipo
/dev/sdc1          2048 16779263 16777216     8G fd RAID Linux autodetectado

Disco /dev/sdd: 7,6 GiB, 8100249600 bytes, 15820800 sectores
Unidades: sectores de 1 * 512 = 512 bytes
Tamaño de sector (lógico/físico): 512 bytes / 512 bytes
Tamaño de E/S (mínimo/óptimo): 512 bytes / 512 bytes
Tipo de etiqueta de disco: dos
Identificador del disco: 0xc3072e18

Disposit.  Inicio Comienzo    Final Sectores Tamaño Id Tipo
/dev/sdd1  *       12144 15820799 15808656   7,6G fd RAID Linux autodetectado
```

Creación del volumen RAID `/dev/md0`:

```
mdadm -C /dev/md0 -l 1 -n 2 /dev/sdc1 /dev/sdd1
mdadm: Note: this array has metadata at the start and
  may not be suitable as a boot device.  If you plan to
  store '/boot' on this device please ensure that
  your boot-loader understands md/v1.x metadata, or use
  --metadata=0.90
mdadm: largest drive (/dev/sdc1) exceeds size (7895104K) by more than 1%
Continue creating array? y
mdadm: Defaulting to version 1.2 metadata
mdadm: array /dev/md0 started.
```

Se crea un sistema de archivos ext4 en el volumen RAID `/dev/md0`:

```
mkfs -t ext4 /dev/md0
mke2fs 1.44.6 (5-Mar-2019)
Creating filesystem with 1973776 4k blocks y 493856inodes
Filesystem UUID: 908429ce-eecf-4350-8b25-ab43b61ff67d
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632

Allocating group tables: done
Writing inode tables: done
Creating journal (16384 blocks): done
Writing superblocks and filesystem accounting information: done
```

Se monta el sistema de archivos:

```
mkdir safe
mount /dev/md0 /root/safe
```

Se crean algunos archivos y directorios:

```
mkdir /root/safe/etc /root/safe/datos /root/safe/bin
cp /etc/hosts /etc/passwd /etc/fstab /root/safe/etc
```

```
cp /usr/bin/date /root/safe/bin/prog
```

Comprobamos el estado del volumen RAID, unos instantes más tarde:

```
mdadm -D /dev/md0
/dev/md0:
        Version: 1.2
  Creation Time: Thu Mar 12 10:38:18 2020
     Raid Level: raid1
     Array Size: 7895104 (7.53 GiB 8.08 GB)
  Used Dev Size: 7895104 (7.53 GiB 8.08 GB)
   Raid Devices: 2
  Total Devices: 2
    Persistence: Superblock is persistent

    Update Time: Thu Mar 12 11:03:12 2020
          State: clean
 Active Devices: 2
Working Devices: 2
 Failed Devices: 0
  Spare Devices: 0

Consistency Policy: resync

           Name: beta:0  (local to host beta)
           UUID: 7961964d:b7fc5583:1b84ac89:b28fd813
         Events: 155

    Number   Major   Minor   RaidDevice State
       0       8       33        0      active sync   /dev/sdc1
       1       8       49        1      active sync   /dev/sdd1
```

El volumen RAID es correcto.

Desconectamos una de las dos memorias USB y se comprueba el estado del volumen RAID:

```
mdadm -D /dev/md0
/dev/md0:
        Version: 1.2
  Creation Time: Thu Mar 12 10:38:18 2020
     Raid Level: raid1
     Array Size: 7895104 (7.53 GiB 8.08 GB)
  Used Dev Size: 7895104 (7.53 GiB 8.08 GB)
   Raid Devices: 2
  Total Devices: 1
    Persistence: Superblock is persistent

    Update Time: Thu Mar 12 11:11:19 2020
          State: clean, degraded
 Active Devices: 1
Working Devices: 1
 Failed Devices: 0
  Spare Devices: 0

Consistency Policy: resync

           Name: beta:0  (local to host beta)
           UUID: 7961964d:b7fc5583:1b84ac89:b28fd813
         Events: 158

  Number   Major   Minor   RaidDevice State
     0       8       33        0      active sync   /dev/sdc1
     -       0        0        1      removed
```

El volumen RAID está en modo degradado, pero utilizable.

Se comprueba que los datos siguen siendo correctos y accesibles:

```
ls -l /root/safe
total 28
drwxr-xr-x. 2 root root  4096 12 marzo 10:44 bin
drwxr-xr-x. 2 root root  4096 12 marzo 10:44 datos
drwxr-xr-x. 2 root root  4096 12 marzo 10:44 etc
drwx------. 2 root root 16384 12 marzo 10:40 lost+found
echo hola > /root/safe/datos/archivo.txt
```

```
cat /root/safe/datos/archivo.txt
hola
```

Se vuelve a montar la memoria USB y se reactiva su partición en el volumen RAID:

```
mdadm --add /dev/md0 /dev/sdd1
```

Comprobamos el estado del volumen RAID:

```
mdadm -D /dev/md0
/dev/md0:
        Version: 1.2
  Creation Time: Thu Mar 12 10:38:18 2020
     Raid Level: raid1
     Array Size: 7895104 (7.53 GiB 8.08 GB)
  Used Dev Size: 7895104 (7.53 GiB 8.08 GB)
   Raid Devices: 2
  Total Devices: 2
    Persistence: Superblock is persistent

    Update Time: Thu Mar 12 11:27:22 2020
          State: clean, degraded, recovering
 Active Devices: 1
Working Devices: 2
 Failed Devices: 0
  Spare Devices: 1

Consistency Policy: resync

 Rebuild Status: 0% complete

           Name: beta:0  (local to host beta)
           UUID: 7961964d:b7fc5583:1b84ac89:b28fd813
         Events: 168

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

```
       2     8     49      1      spare rebuilding   /dev/sdd1
```

La memoria USB se ha integrado en el volumen RAID y está sincronizándose.