Capítulo 4
Los discos y el sistema de archivos

1. Representación de los discos

Nota previa: las unidades de medida de almacenamiento usadas en este capítulo y en todo el libro usan la representación tradicional, según la regla 1KB = 1024 bytes (2^{10}), a no ser que se indique lo contrario.

1.1 Nomenclatura

Este apartado realiza un repaso a los puntos ya vistos en el capítulo Presentación de Linux. En función del tipo de controlador e interfaz en los cuales se conectan los discos, Linux da diferentes nombres a los archivos especiales que representan discos duros.

Cada disco y cada partición está representado por un archivo especial de tipo bloque.

1.1.1 IDE

Se conserva esta sección por razones históricas, la norma SATA ha reemplazado la norma IDE sobre casi todos los ordenadores de escritorio y portátiles desde hace diez años. Los discos con controladores IDE (también llamados PATA, *Parallel Ata* o ATAPI) se llaman hdX:

- hda: IDE0, Master
- hdb: IDE0, Slave

– hdc: IDE1, Master
– hdd: IDE1, Slave
– etc.

Contrariamente a lo que se cree, no hay límite al número de controladores IDE, más allá del número de los puertos de extensión de la máquina (slots PCI). Existen muchas tarjetas adicionales y convertidores que permiten leer antiguos discos IDE. A partir de cuatro discos o lectores, los archivos se llaman hde, hdf, hdg, etc.

Linux entiende que los lectores de CD-Rom, DVD y grabadores de tipo IDE/ATAPI son discos IDE y respetan la nomenclatura citada.

Los núcleos de Linux ahora utilizan por defecto un API llamado libata para acceder al conjunto de los discos IDE, SCSI, USB, Firewire, etc. La nomenclatura sigue la de los discos SCSI, que tratamos en el punto siguiente.

1.1.2 SCSI, SATA, USB, FIREWIRE, etc.

Los discos con controladores SCSI, SCA, SAS, FiberChannel, USB, Firewire, thunderbolt (y probablemente otras interfaces exóticas) se llaman sdX. La enumeración de los discos sigue el orden de detección de las tarjetas SCSI y de los adaptadores (hosts) asociados, más la adición o supresión manual de otras interfaces de discos duros mediante hotplug o udev.

– sda: primer disco SCSI
– sdb: segundo disco SCSI
– sdc: tercer disco SCSI
– etc.

La norma SCSI marca una diferencia entre los diversos soportes. Así, los lectores de CD-Rom, DVD, HD-DVD, BlueRay y los grabadores asociados no llevan el mismo nombre. Los lectores y grabadores están en srX (sr0, sr1, etc.). También puede encontrar scd0, scd1, etc. Pero suelen ser vínculos simbólicos hacia sr0, sr1, etc.

A partir de sdz, la enumeración arranca en sdaa, sdab, etc.

El comando **lsscsi** permite enumerar los periféricos SCSI. Observe que los discos son sdX, mientras que el lector dvd es srX.

```
$ lsscsi
[4:0:0:0]    disk    ATA      ST380011A        8.01  /dev/sda
[5:0:0:0]    cd/dvd  LITE-ON  COMBO SOHC-4836V S9C1  /dev/sr0
[31:0:0:0]   disk    USB2.0   Mobile Disk      1.00  /dev/sdb
```

1.2 Casos especiales

1.2.1 Controladores específicos

Algunos controladores no siguen esta nomenclatura. Por ejemplo, es el caso de algunos controladores RAID físicos. Hay que verlo caso por caso. Un controlador Smart Array en un servidor HP, que utilice el controlador cciss, coloca sus archivos de periféricos en /dev/cciss con los nombres cXdYpZ, donde X es el slot, Y el disco y Z la partición... Los nuevos controladores usan el piloto hpsa, explotan la capa SCSI del núcleo y usan el nombrado estándar de los periféricos.

1.2.2 Virtualización

La representación de discos de sistemas invitados (*guests*) virtualizados depende del tipo de controlador simulado. La mayoría son de tipo IDE o SCSI, y en ambos casos muy a menudo con libata son vistos como SCSI. Sin embargo, algunos sistemas, como por ejemplo KVM o XEN (así como los utilizan los entornos cloud, como AWS) que ofrece paravirtualización, disponen de un controlador específico que presenta los discos con el nombre vdX (virtual disk x o xvdX):

– vda: primer disco virtualizado, o vxda.

– vdb: segundo disco virtualizado, o vxbd.

– etc.

1.2.3 SAN, iSCSI, multipathing

Los discos conectados a través de una SAN (*Storage Area Network*, generalmente con fibra óptica) o mediante iSCSI se ven como discos SCSI y conservan esta nomenclatura. Sin embargo, los sistemas de gestión de rutas múltiples (*multipathing*) se ubican por debajo, proporcionando otros nombres. Powerpath llamará a los discos emcpowerx (emcpowera, emcpowerb, etc.) mientras que el sistema por defecto de Linux llamado multipath los llamará mpathx (mpath0, mpath1, etc.) o de cualquier otro modo elegido por el administrador.

2. Operaciones de bajo nivel

2.1 Información

El comando **hdparm** permite efectuar un gran número de operaciones directamente en los discos duros gestionados por la librería libata, o sea todos los discos SATA, ATA (IDE) y SAS. El comando **sdparm** puede hacer más o menos lo mismo para los discos SCSI. Observe que, a pesar de que los nombres de periféricos de la libata sean idénticos a los del SCSI, es más que probable que muchas opciones de configuración de hdparm no funcionen en discos SCSI. Lo mismo vale para sdparm con los discos SATA o IDE. Los ejemplos que damos a continuación se basan en hdparm.

Para obtener información completa relativa a un disco, utilice los parámetros `-i` o `-I`. El primero recupera la información, desde el núcleo, que se obtiene en el momento del arranque. El segundo interroga directamente al disco. Es preferible `-I` porque da una información muy detallada.

```
# hdparm -I /dev/sda

/dev/sda:

ATA device, with non-removable media
    Model Number:       VBOX HARDDISK
    Serial Number:      VB91a2e953-933cdc65
    Firmware Revision:  1.0
Standards:
    Used: ATA/ATAPI-6 published, ANSI INCITS 361-2002
    Supported: 6 5 4
Configuration:
    Logical         max     current
    cylinders     16383     16383
    heads         16    16
    sectors/track     63    63
    --
    CHS current addressable sectors:   16514064
    LBA    user addressable sectors:   63152320
    LBA48  user addressable sectors:   63152320
    Logical/Physical Sector size:           512 bytes
    device size with M = 1024*1024:       30836 MBytes
    device size with M = 1000*1000:       32333 MBytes (32 GB)
    cache/buffer size  = 256 KBytes (type=DualPortCache)
Capabilities:
    LBA, IORDY(cannot be disabled)
    Queue depth: 32
    Standby timer values: spec'd by Vendor, no device specific minimum
    R/W multiple sector transfer: Max = 128        Current = 128
    DMA: mdma0 mdma1 mdma2 udma0 udma1 udma2 udma3 udma4 udma5 *udma6
```

```
          Cycle time: min=120ns recommended=120ns
     PIO: pio0 pio1 pio2 pio3 pio4
          Cycle time: no flow control=120ns  IORDY flow control=120ns
Commands/features:
     Enabled    Supported:
        *    Power Management feature set
        *    Write cache
        *    Look-ahead
        *    48-bit Address feature set
        *    Mandatory FLUSH_CACHE
        *    FLUSH_CACHE_EXT
        *    Gen2 signaling speed (3.0Gb/s)
        *    Native Command Queueing (NCQ)
Checksum: correct
```

2.2 Modificación de los valores

Se puede modificar varios parámetros de los discos. Sin embargo, ¡cuidado! Algunas opciones de **hdparm** pueden resultar peligrosas tanto para los datos contenidos en el disco como para el propio disco. La mayoría de los parámetros son de lectura y escritura. Si no se especifica ningún valor, **hdparm** muestra el estado del disco (o del bus) para este comando. A continuación le presentamos algunos ejemplos de opciones interesantes.

- **-c**: anchura del bus de transferencia EIDE en 16 o 32 bits. 0=16, 1=32, 3=32 compatible.
- **-d**: utilización del DMA. 0=no DMA, 1=DMA activado.
- **-X**: modifica el modo DMA (mdma0 mdma1 mdma2 udma0 udma1 udma2 udma3 udma4 udma5). Puede utilizar cualquiera de los modos anteriores o valores numéricos: 32+n para los modos mdma (n varía de 0 a 2) y 64+n para los modos udma.
- **-C**: modo de ahorro de energía en el disco (unknown, active/idle, standby, sleeping). Se puede modificar el estado con -S, -y, -Y y -Z.
- **-g**: muestra la geometría del disco.
- **-M**: indica o modifica el estado del Automatic Acoustic Management (AAM). 0=off, 128=quiet y 254=fast. No todos los discos lo soportan.
- **-r**: pasa el disco en sólo lectura.
- **-T**: bench de lectura de la caché del disco, ideal para probar la eficacia de transferencia entre Linux y la caché del disco. Hay que volver a ejecutar el comando dos o tres veces.
- **-t**: bench de lectura del disco, fuera de la caché. Mismas observaciones que para la opción anterior.

Así, el comando siguiente pasa el bus de transferencia a 32 bits, activa el modo DMA en modo Ultra DMA 5 para el disco sda:

```
# hdparm -c1 -d3 -X udma5 /dev/sda
```

Le mostramos a continuación otros ejemplos:

```
# hdparm -c /dev/sda

/dev/sda:
 IO_support    =  0 (default 16-bit)

# hdparm -C /dev/sda

/dev/sda:
 drive state is:  active/idle

# hdparm -g /dev/sda

/dev/sda:
 geometry      = 3931/255/63, sectors = 63152320, start = 0

# hdparm -T /dev/sda

/dev/sda:
 Timing cached reads:   23868 MB in  2.00 seconds = 11950.45 MB/sec
# hdparm -t /dev/sda

/dev/sda:
 Timing buffered disk reads: 308 MB in  3.02 seconds = 101.87 MB/sec
```

3. Elegir un sistema de archivos

3.1 Fundamentos

3.1.1 Definición de sistema de archivos

La acción de "formatear" un disco, un pendrive o cualquier soporte de datos consiste únicamente en crear en un soporte de memoria secundaria (volumen de almacenamiento) la organización lógica que permite colocar datos en él. La palabra "formateo" en Linux se utiliza para describir la creación de un sistema de archivos. Hablamos de un sistema de archivos que representa a la vez la organización lógica de los soportes tanto a un nivel inferior como a un nivel de usuario.

No se escribe la información en los discos de cualquier manera. Se requiere una mínima organización para colocar en ellos tanto la información relativa a los archivos como los datos almacenados. El sistema de archivos (y los controladores asociados) es el que define esta organización. Si bien los fundamentos organizativos suelen ser los mismos en los diferentes sistemas de archivos presentes soportados por Linux, las implementaciones y organizaciones lógicas de los datos en el disco varían bastante de uno a otro. De esta manera, no hay un único tipo de sistema de archivos, sino varios, puestos a disposición del usuario, el administrador o el ingeniero.

Todos los sistemas de archivos de Linux deben respetar las normas POSIX. Como POSIX define un conjunto de reglas básico, un sistema de archivos puede ir más lejos de esta norma ofreciendo extensiones. La mayoría de estas conciernen a elementos de seguridad, como las ACL o selinux.

El principio básico es asociar un nombre de archivo con su contenido y autorizar su acceso: creación, modificación, supresión, desplazamiento, apertura, lectura, escritura, cierre. Conforme a este principio, el sistema de archivos debe gestionar lo que deriva de ello: mecanismos de protección de los accesos (permisos, propietarios), accesos concurrentes, etc.

3.1.2 Representación

Además de la organización y el almacenamiento de la información y datos en los archivos, el sistema de archivos debe facilitar al usuario una visión estructurada de sus datos, que permite distinguirlos, encontrarlos, tratarlos y trabajar con ellos en forma de archivos dentro de una estructura de directorios con los comandos asociados. Asimismo, cada sistema de archivos debe proporcionar lo necesario para que los programas puedan acceder a él.

Un sistema de archivos Unix se organiza en forma de un árbol de directorios y subdirectorios desde una raíz común. Es una estructura en árbol. Gran parte de estas se presentaron en el capítulo anterior. Cada directorio forma parte de una organización y propone, a su vez, una organización: el sistema de archivos dispone de una jerarquía ordenada. Se puede repartir la propia estructura en árbol entre varios soportes y sistemas de archivos.