Ficheros Sistemas Operativos

Enrique Soriano

GSYC

29 de noviembre de 2018







Algunos derechos reservados. Este trabajo se entrega bajo la licencia Creative Commons Reconocimiento
NoComercial - SinObraDerivada (by-nc-nd). Para obtener la licencia completa, véase

http://creativecommons.org/licenses/by-sa/2.1/es. También puede solicitarse a Creative Commons, 559 Nathan

Abbott Way, Stanford, California 94305, USA.

Ficheros

- Tradicionalmente: datos que persisten tras la ejecución del proceso y/o del sistema (excepción: ramfs), y que son compartidos por distintos procesos y/o sistemas.
- Requisitos:
 - Gran tamaño
 - Durabilidad
 - Acceso concurrente
 - Protección
- Más general: es una interfaz. P. ej. puede representar un dispositivo.

Ficheros

- Estructura: fichero simple (Unix), Resource Forks (Mac)...
- Nombre: ¿extensión?
- ¿Tipos?
- Ya hemos usado ficheros desde nuestros programas, p. ej. printf().
- Operaciones básicas: open, read, write, close.

Discos: estructura interna

- Bloque físico

 determinado por el HW (sector).

 Tradicionalmente de 512 bytes, algunos fabricantes están pasando a 4 Kb.
- ullet Bloque lógico o determinado por el sistema de ficheros.
- Si no coinciden, empaquetamos.
- Siempre hay fragmentación interna en los bloques.
- ¿Tamaño? compromiso entre:

```
\uparrow tamaño de bloque \Rightarrow \uparrow fragmentación \uparrow tamaño de bloque \Rightarrow \uparrow tasa de transferencia
```

Discos

Acceso al disco:

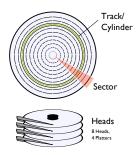
- Secuencial: se accede registro a registro (contiguos).
- Aleatorio: se accede a cualquier registro inmediatamente.

¿Y los discos duros actuales? Caches flash, discos SSD, híbridos, etc.

Discos

Discos

- Direccionamiento de bloques:
 - CHS: Cylinder, Head, Sector. Esquema viejo.
 - LBA: Linear.
- Así eran:





Discos

- Algoritmo de planificación de disco: ¿cómo organiza el SO las operaciones?
- El algoritmo depende totalmente del tipo de disco que usemos.
- ¿Cómo se relaciona el direccionamiento LBA con la geometría real del disco? ¿Se puede generalizar? ¿Hay caches internas en el disco? ¿Es un disco de verdad o es un disco virtual?
- Los clásicos:
 - FIFO: es justo, no altera el orden de las operaciones. Lento en discos mecánicos.
 - Shortest Seek First: no es justo, puede provocar hambruna, desordena operaciones.
 - Ascensor: compromiso justicia/eficiencia, no provoca hambruna, desordena operaciones.
 - ... (hay muchos otros)



Particiones

- Un disco se puede dividir en partes: particiones.
- En linux, el comando fdisk permite manipular las particiones de un disco.

Particiones PC (BIOS)

Esquema tradicional de los PC, ha sido reemplazado por UEFI:

- MBR: Primer bloque del disco (512 bytes).
 - Cargador primario: 440 bytes.
 - Tabla de particiones primarias: 4 entradas de 16 bytes.
 - Arrancable/no arrancable.
 - Dirección del primer bloque (CHS).
 - Tipo de partición (ntfs, linux, linux swap, plan9, ...).
 - Dirección del último bloque (CHS).
 - Dirección del primer bloque (LBA).
 - Número de bloques de la partición.
 - Direcciones de 32 bits: como mucho discos de 2TB.

Particiones UEFI

GUID Partition Table (GPT) :

- Usa direcciones y tamaños de 64 bits.
- Primer bloque (LBA 0): legacy MBR.
- Segundo bloque (LBA 1): GPT header, con punteros para:
 - Array de bloques con la tabla de particiones (empieza en LBA 2). El array tiene un tamaño mínimo de 16 Kb independientemente del tamaño del sector físico. Las entradas en la tabla (128 bytes) continenen: tipo de particion, LBA del primer bloque, LBA del último bloque, atributos (read-only, oculta, etc.), nombre de particion...
 - Copia de la cabecera.
 - El primer bloque usable.
 - ...
- El cargador está en una partición de tipo ESP (EFI System Partition, que es una FAT).



Volúmenes Lógicos

 Los volúmenes lógicos (LV) pueden estar formados por distintos trozos (physical extends, PE) de distintos volúmenes físicos (PV). P. Ej. Linux LVM.

```
Typical limits for Linux LVM v1:
<- p →
                              Range of PE size (p): 8KB... 512MB
                              Range of PV size (v): 512MB... 2TB
                                 Range of PEs (N): 1... 65534
                                 Range of PVs (X): 1... 256
  Physical Volume
                                  Range of VGs(Z): 1... 99
                       Volume Group
                                               Volume Group,→
   Physical Volume.
                                    Physical Volume
        PE, PE,,
       /root
                                               /var
                          /home
  Logical Volume, Logical Volume, Logical Volume,
```

Imágenes de disco

- Un fichero que tiene dentro la estructura completa (bloque a bloque) de un disco duro, CD-ROM (ISO 9660), etc.
- Hay distintos formatos: iso, bin, dmg...
- Son útiles para máquinas virtuales, copias de seguridad, clonar sistemas, transmisión de discos por la red, etc.

Sistemas de Ficheros

- El sistema de ficheros nos ofrece la abstracción de fichero: no pensamos en los bloques, para nosotros un fichero es una secuencia de bytes.
- ¿Cómo asigna los bloques a un fichero?
- Se puede implementar de distintas formas: asignación continua, enlazada, enlazada con tabla, indexada

Asignación de espacio contigua

- Un fichero ocupa una serie de bloques contiguos en disco.
- Acceso rápido en discos de acceso secuencial.
- Rendimiento alto: un acceso para conseguir un dato cualquiera → la entrada de directorio contiene la dirección de inicio y la longitud del fichero.
- Problema: si hay asignación dinámica → fragmentación externa.
- Problema: hay que saber el tamaño máximo del fichero cuando se crea.
- Tamaños limitados por los huecos → compactación.
- ¿Se usa? Sí: CD, DVD, ...

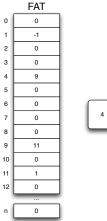


FAT

Es un caso de asignación de espacio usando una **lista enlazada** implementada con una tabla:

- Usa bloques lógicos grandes: clusters.
- La tabla FAT tiene una entrada por cluster.
- El índice de la tabla es el número de cluster.
- La entrada de directorio tiene la referencia a la entrada primer cluster.
- La entrada del cluster actual referencia la entrada del siguiente cluster.
- Mejora del acceso aleatorio respecto a una lista sin tabla: no hay que leer los clusters del fichero para conseguir la dirección del cluster N.

FAT



MYFILE.txt First: 4 clusters:

FAT

- Si queremos la tabla FAT en memoria principal \rightarrow no puede ser muy grande \rightarrow clusters grandes \rightarrow fragmentación interna.
- El acceso aleatorio es más rápido que en lista enlazada normal.
- ¿Y si se estropea la tabla FAT?

Ejemplo: FAT32

Reserved sectors	Fat Region	Data Region
------------------	---------------	-------------

- El sector 0 de la partición es el boot sector (también está duplicado en el sector 6). Contiene código del cargador secundario e información sobre el sistema de ficheros:
 - No de sectores, 4 bytes (Max. tam. de disco: 2 Tb)
 - Etiqueta del volumen.
 - Nº de copias de la tabla FAT.
 - Primer cluster del raíz.
 - ...



Ejemplo: FAT32

- Entrada de directorio (32 bytes):
 - Nombre del archivo, 8 Bytes.
 - Extensión del archivo, 3 Bytes.
 - Atributos del archivo, 1 Byte.
 - Reservado, 10 Bytes.
 - Hora de la última modificación, 2 bytes.
 - Fecha de la última modificación, 2 bytes.
 - Primer cluster del archivo, 4 bytes.
 - Tamaño del archivo, 4 bytes.
- Clusters de 32 Kb, 64 sectores.

Asignación de espacio: indexada

- Idea: se indexan los bloques de datos del fichero en un *bloque* de indirección.
- Bloque de indirección grande \rightarrow desperdicio.
- Bloque de indirección pequeño \rightarrow no soporta ficheros grandes.
- Para un acceso a datos, siempre son dos accesos a disco.

Asignación de espacio: indexada multinivel

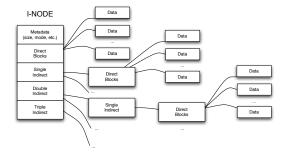
- Por niveles. P. ej. indirección doble:
 - Los bloques de índice de 1er nivel apuntan a bloques de índice de 2º nivel.
 - Los bloques de índice de 2º nivel apuntan a bloques de datos.
- Ejemplo: bloques de 4Kb con punteros de 4 bytes.
- Problema: Para los ficheros pequeños, desperdiciamos.
- Problema: Para un acceso a datos, n accesos a disco.

Asignación de espacio: indexada con esquema combinado

- Algunos bloques directos de datos.
- Algunos bloques de indirección simple.
- Algunos bloques de indirección doble.
- ...
- Nos quedamos con lo mejor de cada modelo anterior.

Ficheros en Unix: i-nodos

Es un caso de asignación indexada con esquema combinado:



Unix: Partición



- Bloque de arranque (boot block): código de arranque del sistema.
- Superbloque: tamaño del volumen, número de bloques libres, lista de bloques libres, tamaño del vector de i-nodos, siguiente i-nodo libre, cierres...
- Vector de i-nodos: representación de los ficheros.
- Bloques de datos.

- Cada fichero/directorio tiene un i-nodo asociado, definido por un número de i-nodo. El directorio raíz siempre tiene el número de i-nodo 2.
- La estructura se localiza en el vector indexando por el número de i-nodo.
- El OS mantiene una cache de i-nodos en memoria.

- La estructura contiene:
 - permisos
 - tiempos
 - acceso a datos (atime)
 - modificación de los datos (mtime)
 - modificación del i-nodo (ctime)
 - tamaño
 - dueño
 - tipo
 - número de bloques
 - contador de referencias (links)
- No contiene el nombre de fichero.



- Un directorio relaciona un nombre con un i-nodo: entrada de directorio.
- Un directorio tiene:
 - Su propio i-nodo
 - Bloques de datos con la lista de entradas de directorio.
- Entre las entradas, tiene la entrada de . (su i-nodo) y ... (i-nodo del padre).
- El kernel mantiene una cache de entradas de directorio en memoria.

/home/joe/test.txt



2	
2	
3	usr
5	home

DATA BLOCK 321

5	
2	
7	joe
9	jane

DATA BLOCK 239



I-NODE 5

METADATA direct-block 321 I-NODE 7

METADATA direct-block 239

- Enlaces duros: es otro nombre para el fichero, la entrada de directorio apunta al mismo i-nodo que el antiguo nombre. En general, no se permite crear enlaces duros para directorios: rompen la jerarquía, crean bucles y crea ambigüedad en dot-dot.
- Enlaces simbólicos: es un fichero cuyos datos contienen la ruta al fichero enlazado → pueden romperse.

Unix: Permisos

Se establecen esos permisos para:

- dueño.
- grupo.
- resto de usuarios.

rwx rwx rwx

Tipo de acceso:

- r: permiso de lectura. En directorio: se pueden leer las entradas de directorio.
- w: permiso de escritura. En directorio: se pueden escribir las entradas del directorio (borrar, renombrar, añadir ficheros).
- x: permiso de ejecución. En directorios: se puede entrar o atravesar el directorio cuando se evalúa una ruta. Es necesario para acceder a un fichero del directorio (datos y metadatos).

Unix: Permisos

- Los permisos se representan normalmente en octal:
 P. ej: 0664 es 110 para el dueño, 110 para el grupo, 100 para el resto.
- Hay otros permisos
 - sticky bit (+t): es para directorios: no puedes borrar una entrada si no eres el dueño del directorio, del fichero/directorio que representa la entra, o root.
 - **setuid/setgid bit** (+s): el proceso que ejecute el fichero adoptará el UID/GID del dueño/grupo del fichero.
- Podemos ver los permisos con el comando 1s -1. Si los queremos ver del directorio y no de las entradas: -d.

Unix: Permisos

- El comando chmod cambia los permisos de un fichero. Solo lo puede hacer el dueño del fichero y root.
- Inicialmente, el creador de un fichero es su dueño y grupo.
- El comando chown cambia el dueño de un fichero. Hay que tener privilegios especiales para hacer esto.
- El comando chgrp cambia el grupo de un fichero. El dueño puede cambiarlo a un grupo al que él pertenezca.

Fallos

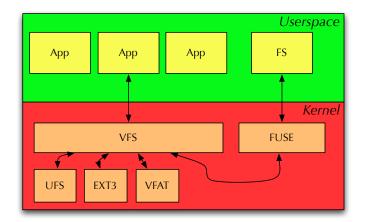
- Una pérdida de coherencia en los metadatos (estructuras del sistema de ficheros) es peor que perder los datos.
- El comando fsck sirve para reparar el sistema de ficheros después de un fallo.
- Hay que recorrer los bloques de los ficheros para detectar errores, p. ej.:
 - Un bloque está asignado a un fichero y en la lista de bloques libres a la vez.
 - Un bloque en dos ficheros a la vez.
- Esto es caro.
- Los sistemas de ficheros modernos usan técnicas para no acabar con metadatos incoherentes: journaling.



Sistemas de Ficheros: implementación

- VFS (Virtual File System): interfaz común con el kernel para todos los sistemas de ficheros.
- Lo normal es que el sistema de ficheros esté implementado como un módulo del kernel.
- Pero también puede estar implementado como un programa de espacio de usuario.
- FUSE: módulo del kernel para implementar sistemas de ficheros en espacio de usuario que se puedan integrar con VFS.

Sistemas de Ficheros: implementación



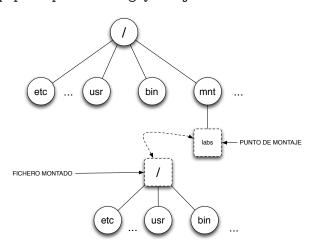
Espacio de nombres

- El árbol de ficheros del sistema.
- Se "camina" (walk) por la ruta, resolviendo cada parte.
- Se pueden *montar* nuevos árboles al espacio de nombres.
- El comando mount nos permite ver y modificar el espacio de nombres.
- Unix: un espacio de nombres común para todos los procesos del el sistema.
- Linux y otros sistemas modernos dejan tener diferentes espacios de nombres a grupos de procesos.

Espacio de nombres

Ejemplo:

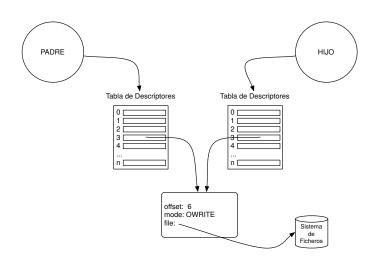
sshfs pepe@alpha.aulas.gsyc.urjc.es://mnt/labs



Procesos y ficheros

- Por cada proceso, el kernel mantiene una tabla con los ficheros que tiene abiertos: tabla de descriptores de fichero.
- Un fichero abierto tiene un descriptor de fichero (file descriptor) que lo identifica.
- El número de descriptor de fichero es el índice en la tabla.
- Un fichero abierto tiene offset, modo de apertura, ...
- Los hijos heredan una copia de la tabla del padre.
- Hay tres posiciones especiales en la tabla: (0) entrada estándar, (1) salida estándar, y (2) salida de errores.

Descriptores de fichero



Terminal

- Los procesos que creamos en el shell normalmente tienen como entrada, salida y salida de errores al **terminal**.
- No siempre es así, y se puede cambiar.
- El terminal es un fichero como cualquier otro.
- Cuando se lee, se lee del teclado.
- Cuando se escribe, se escribe en la pantalla.
- /dev/tty* son los terminales, /dev/pts/* son pseudotermianles (terminales emulados).
- Para un proceso, su terminal siempre se llama:

/dev/tty



Read

 read: lee como mucho el número de bytes indicado del fichero abierto en ese descriptor. Puede leer menos bytes sin ser un error: lectura corta. Si devuelve 0 bytes, se ha llegado al final del fichero.

```
ssize_t read(int fd, void *buf, size_t count);
```

Write

 write: escribe el número de bytes del buffer indicado en el fichero abierto en ese descriptor. Si devuelve un número distinto al número solicitado, se debe considerar un error.

```
ssize_t write(int fd, const void *buf, size_t count);
```

Pread/Pwrite

• Hay versiones de read y write a las que se le da el offset:

Lseek

- 1seek: cambia el offset del descriptor de fichero. Retorna el offset resultante, -1 en error. Tiene tres modos de operación (tercer parámetro):
 - SEEK_SET: el offset es absoluto.
 - SEEK_CUR: se suma al offset actual.
 - SEEK_END: se suma al offset del final del fichero.

off_t lseek(int fd, off_t offset, int whence);

Open

- open: abre un fichero en el modo indicado por el segundo parámetro. Se usará el primer descriptor de fichero libre.
 Retorna el número del descriptor, -1 en error.
- El segundo parámetro es una combinación de flags:
 - O_CREAT: el fichero se quiere crear si no existe. En este caso, hay que proporcionar el tercer parámetro.
 - O_RDONLY: se va a leer.
 - O_WRONLY: se va a escribir.
 - O_RDWR: se va a leer y escribir.
 - O_TRUNC: si el fichero existe y se abre para escribir (O_RDWR, O_WRONLY), entonces se trunca a longitud 0.
 - O_APPEND: se va a escribir siempre al final del fichero (lseek+write).
 - O_CLOEXEC: se debe cerrar automáticamente si se llama a exec.
 - ...

```
int open(const char *pathname, int flags);
int open(const char *pathname, int flags, mode_t mode);
```



Open

- Los permisos se comprueban en la apertura.
- Cuando se abre un fichero, el sistema mantiene el offset, que se actualiza cada vez que se lee o se escribe en él (empieza a 0).
- Si se crea el fichero, el dueño del fichero será el UID del proceso. Dependiendo del sistema, el grupo puede ser el GID del proceso o el grupo del directorio en el que se crea.
- En creación, se abrirá el fichero en el modo indicado, independientemente de los permisos que se le asignen.
- Los permisos finales del fichero están sujetos a la máscara umask.

Dup

- dup: duplica un descriptor de fichero en la primera posición disponible en la tabla. Retorna el número del descriptor en el que se ha duplicado, -1 en error.
- dup2: duplica el descriptor en la posición indicada en el segundo parámetro, cerrándolo antes si estaba abierto.

```
int dup(int fildes);
int dup2(int fildes, int fildes2);
```

Close

close: cierra un fichero abierto y libera los recursos. Retorna

 1 si hay un error. Hay que cerrar los ficheros que ya no
 vayamos a usar. Las tres primeras posiciones en la tabla de
 descriptores no se deben quedar cerradas: el proceso siempre
 debe tener entrada, salida y salida de errores.

int close(int fd);

Umask

- umask: cambia la máscara de creación y devuelve la antigua.
 Nunca falla.
- La máscara umask es una propiedad del proceso, que hereda del padre.
- Siendo perm los permisos que se desean establecer en la creación, los permisos finales del fichero serán:

```
perm & (~ umask)
```

```
mode_t umask(mode_t mask);
```

Access

- access: devuelve cero si el fichero puede accederse en el modo indicado, -1 en otro caso.
- Bits para mode:
 - F_OK: si existe.
 - R_OK: si se puede leer.
 - W_OK: si se puede escribir.
 - \bullet X_OK: si se puede ejecutar.

int access(const char *pathname, int mode);

Unlink

 unlink: elimina un nombre (enlace duro) de un fichero. Si es el último que le queda a un fichero, y ningún proceso lo tiene abierto, el fichero se elimina y se liberan sus recursos.

```
int unlink(const char *pathname);
```

Stat

- stat: lee los metadatos de un fichero, los deja en la estructura que se pasar por referencia. Si falla retorna -1.
- Cuidado: atraviesa los enlaces simbólicos. Si queremos los metadatos del enlace y no del fichero apuntado, hay que usar 1stat.

int stat(const char *pathname, struct stat *statbuf);

Stat

Campos de struct stat:

- st_ino: i-nodo del fichero.
- st_mode: entero con el modo, que son permisos, etc.
- st_nlinks: número de nombres (enlaces duros).
- st_uid: dueño.
- st_gid: grupo.
- st_size: tamaño del fichero.
- st_atime: hora del último acceso al fichero.
- st_mtime: hora de la última modificacón de sus datos.
- st_ctime: hora del último cambio en sus datos o metadatos (excepto en el tiempo de acceso).
- ...



Stat

Aplicando la máscara S_IFMT al campo st_mode podemos saber el tipo de fichero, comparando con estas constantes:

- S_IFDIR: es un directorio.
- S_IFREG: es un fichero normal.
- S_IFLNK: es un enlace simbólico.
- ...

Hay otras máscaras para acceder a los permisos, etc. man 7 inode

Ejemplo:

```
if(st.st_mode & S_IFMT == S_IFREG)
    printf("it's a regular file!\n");
```



Mkdir

 mkdir: crea un directorio. Igual que en el caso anterior, los permisos están sujetos al valor de la máscara umask.

```
int mkdir(const char *path, mode_t mode);
```

Opendir

 opendir: abre un directorio para lectura. El directorio abierto se representa con el tipo DIR. En caso de error devuelve NULL.

```
DIR *opendir(const char *dirname);
```

Readdir

- readdir: retorna un puntero a la siguiente entrada de directorio, representada por la estructura dirent. Retorna NULL en caso de que no haya más o en error.
- En Linux, la estructura dirent tiene estos campos de interés:
 - d_ino: i-nodo.
 - d_name: string con el nombre del fichero.
 - d_type: tipo de entrada de directorio.

```
struct dirent *readdir(DIR *dirp);
```

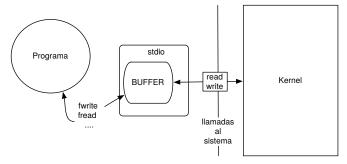
Closedir

• closedir: cierra el directorio. Si no se cierra, tenemos un *leak* de memoria. En caso de error retorna -1.

```
int closedir(DIR *dirp);
```

E/S con buffering

- A veces no es fácil leer de un fichero: p. ej. leer líneas de un fichero.
- Realizar una llamada al sistema sale caro: p. ej. leer carácter a carácter.
- Las operaciones de stdio ofrecen buffering.



E/S con buffering

- Stdio usa streams, representados por la estructura FILE, para representar un fichero abierto.
- Nos ofrece streams para la entrada estándar (stdin), salida (stdout) y salida de errores (stderr).

Fopen

- fopen: abre un fichero para hacer E/S con buffering. El modo de apertura es una string, por ejemplo "r" para lecura, "r+" para lectura-escritura, etc.
- fdopen: configura un stream a partir de un fichero abierto (de su fd).

Fclose

• fclose: vacía el stream, lo cierra y cierra el descriptor de fichero subyacente.

```
int fclose(FILE *stream);
```

Fread

• fread: lee del fichero un número de elementos (nitems) de tamaño size, los guarda a partir de la dirección ptr. Retorna el número de elementos leídos. Retorna menos elementos leídos (o cero) que los que se querían leer cuando llega a fin de fichero, pero también cunado tiene un error. Para diferenciar entre esos dos casos, hay que usar las funciones feof(3) y/o ferror(3). Tanto esas funciones como fread no actualizan la variable errno.

Fwrite

 fwrite: escribe en el fichero un número de elementos (nitems) de tamaño size, de la dirección ptr. Retorna el número de elementos escritos.

Fflush

• fflush: fuerza el vaciado del buffer y su escritura en el fichero.

```
int fflush(FILE *stream);
```

Fgets

 fgets: lee una línea de hasta size bytes. y la guarda en la string str. Siempre deja una string. Si no hay más líneas o hay error, retorna NULL.

Fprintf

• fprintf: igual que printf, pero escribe en el stream indicado (printf siempre lo hace en stdout).

```
int fprintf(FILE *stream, const char *format, ...);
```

Redirecciones en el Shell

- > y < son caracteres especiales para el shell.
- > para redirigir la salida estándar a otro fichero.
- < para redirigir la entrada estándar a otro fichero.
- P. ej.
 - \$> ps > procesos.txt
 - \$> wc -l procesos.txt
 - \$> echo hola que tal > hola.txt
 - \$> cat < /NOTICE</pre>

Redirecciones en Shell

- La mayoría de los comandos que aceptan ficheros como argumentos también leen de su entrada estándar cuando se invocan sin argumentos:
 - \$> cat < procesos.txt</pre>
 - \$> cat procesos.txt # no es lo mismo!!

/dev/null

- /dev/null es "ninguna parte".
 \$> echo hola > /dev/null
- Los comandos que ejecutan en background (&) tiene su entrada estándar redirigida a /dev/null. P. ej:
 \$> cat &

Salida de errores

- Existe para que no se mezclen las salidas. P. ej.
 - \$> ls /kokoko > /tmp/afile
 - ls: /kokoko does not exist
 - \$> cat /tmp/afile # no hay nada!

Más redirecciones

- >> Redirige la salida sin truncar, añade al final del fichero.
- número >
 Redirige la posición especificada en la tabla de descriptores. P. ej.:
 - $\$ ls * 2> /tmp/errores
- número1>&número2
 Duplica el descriptor número2 en el número1. Atención: se aplican de izquierda a derecha. P. ej.
 \$> ping 172.26.0.1 > /dev/null 2>&1

Redirecciones en el Shell

- Se pueden conjugar:
 - \$> cat < /tmp/afile > /tmp/copia-afile
 - \$> wc -l /tmp/afile > /tmp/cuentas 2> /dev/null
- \$> echo 'error!!!' 2>&1
- \$> cat < afile > afile

Error: afile se trunca.

Solución: Usa un fichero temporal.