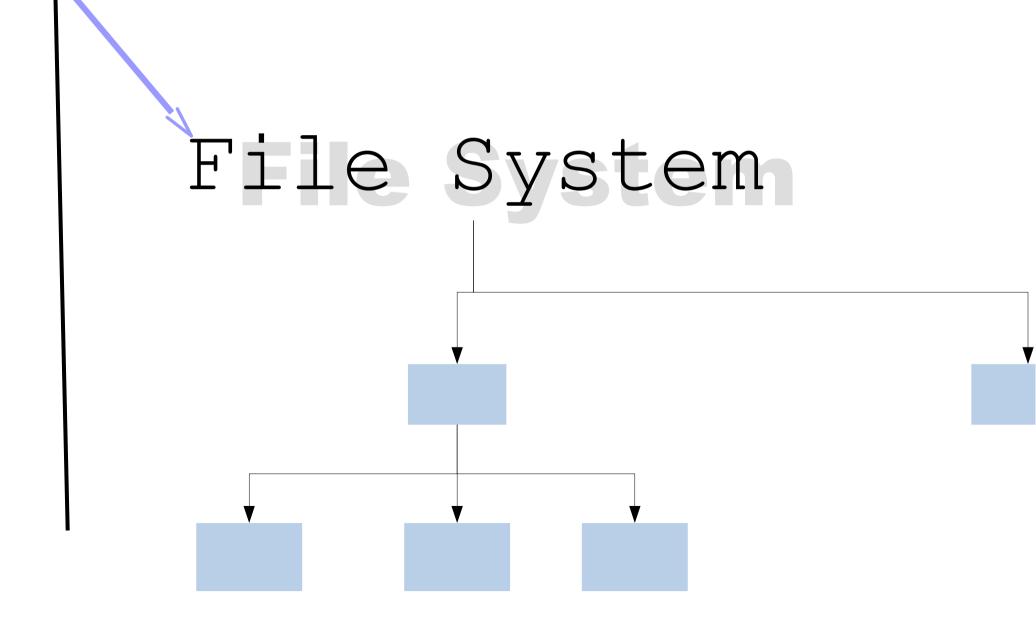
# Organización de Datos





# Indice

- Introducción de las funciones del Sistema de Archivo
- Caracteristicas del Sitema de archivo
- Tipos de Sistema de archivo
- FAT
- Unix File-Systems
  - Inodos
  - Manejo de inodos
- NTFS
  - MFT
- Abstracción del Sistema de archivo



# Introducción al Sistema de Archivo

Es el que indica como se <mark>organiza la información dentro de los medios físicos</mark>. Entre sus funciones son el guardado y recuperación de la información; Administración del espacio libre y el guardado de los atributos de seguridad del sistema operativo (En los filesystem nuevos).

El conjunto de programas del sistema operativo encargados de proveer la visión lógica de la información almacenada a usuarios y programas conforman el Sistema de Archivo.

### Estos programas se encargan de:

- Identificar y localizar archivos
- Seguridad
- Asignación de espacio
- Coordinación de transferencia (sistema multiusuario)



# Tipos de Sistema de archivo

Los sistemas de archivos pueden ser clasificados por medio, de red y de proposito especial.

- Discos de plato magnetico o Flash (acceso aleatorio): FAT, NTFS, HFS+, UFS, EXT2/3/4, XFS, BRTFS, Veritas File system, ZFS, ReiserFS
- Discos Opticos: ISO9660, UDF
- Flash file System (Especialmente embebidos): JFFS, UBIFS, LogFS, F2FS, SquashFS
- **Network File System:** NFS, AFS (Sistema de comparticion de apple), SMB (Sistema de comparticion de microsoft), FTP, Webdav, SSHFS.
- Sistemas de archivos en Cluster:
  - Shared-disk / storage area network: GFS2 (Red hat), GPFS (IBM), CSV (Microsoft), OCFS (Oracle)
  - **Distributed file systems:** GFS (Google), HDFS (Apache), Ceph (Red Hat), DFS (Microsoft), GlusterFS (Red Hat)
- Especiales:
  - procfs --> Mapeo de procesos y estructuras en linux
  - configfs y sysfs --> Muestra como archivos los parametros de consulta y configuración de un Linux
  - **devfs** --> los archivos en esta fs representa a dispositivos. Y permite utilizar las primitivas de cualquier fs para acceder al dispositivo.
  - Superpuestos --> overlayFS, UnionFS, aufs



## Aspectos a evaluar en un sistema de archivos

- Manejo de espacio
  - Manejo de nombres de los objetos (directorio y archivos)
  - La estructura de directorios utilizados
  - Metadata que se le puede asociar a los objetos (directorio y archivos)
- Utilidades: Creación, Chequeos, Defrag, Tunning, Resize, Undelete, etc.
- Mecanismos de permisos y restriccion.
  - ACL
  - Capacidades
- Integridad de la información
- Limitiaciones impuestas por diseño.
- Extras
  - Encriptacion a nivel archivo o directorio
  - Compresión a nivel archivo o directorio
  - Deduplicacion



# Casos de estudio

- FAT
- •UNIX
- NTFS



## **FAT**

El sistema de archivos FAT se compone de cuatro secciones:

### • El sector de arranque.

Siempre es el primer sector de la partición (volumen) e incluye información básica, punteros a las demás secciones, y la dirección de la rutina de arranque del sistema operativo.

### · La región FAT.

Contiene dos copias de la tabla de asignación de archivos (por motivos de seguridad). Estos son mapas de la partición, indicando qué clusters están ocupados por los archivos.

### • La región del directorio raíz.

Es el índice principal de carpetas y archivos.

### La región de datos.

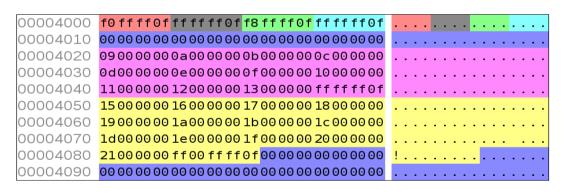
Es el lugar donde se almacena el contenido de archivos y carpetas. Por tanto, ocupa casi toda la partición. El tamaño de cualquier archivo o carpeta puede ser ampliado siempre que queden suficientes clusters libres. Cada cluster está enlazado con el siguiente mediante un puntero. Si un determinado cluster no se ocupa por completo, su espacio remanente se desperdicia.



### **FAT**

### **FAT - Estructuras**

- Boot (sector de arranque)
- Región FAT (dos copias)
- Directorio raiz (Índice principal de carpetas y directorios)
- Datos



FAT ID

Clusters vacíos

Primer cadena (un solo cluster: 0003)

Tercera cadena (16 clusters: 0014 a 0021)

Indicación de fin de cadena

El directorio raiz cabe en un sólo cluster

Segunda cadena (12 clusters: 0008 a 0013)



# FAT - Tabla de asignacion

La tabla de asignación de archivos consta de una lista de entradas. Cada entrada contiene información sobre un clúster:

- La dirección del siguiente clúster en la cadena.
- Si es pertinente, la indicación de "fin de archivo" (que es también el fin de la cadena).
- Un carácter especial para indicar que el clúster es defectuoso.
- Un carácter especial para indicar que el clúster está reservado (es decir, ocupado por un archivo).
- El número cero para indicar que el clúster está libre (puede ser usado por un archivo).
- El tamaño de estas entradas también depende de la variante FAT en uso: FAT16 usa entradas de 16 bits, FAT32 usa entradas de 32 bits, etc.



## FAT - Diretorio Raiz

Este índice es un tipo especial de archivo que almacena las subcarpetas y archivos que componen cada carpeta. Cada entrada del directorio contiene el nombre del archivo o carpeta (máximo 8 caracteres), su extensión (máximo 3 caracteres), sus atributos (archivo, carpeta, oculto, del sistema, o volumen), la fecha y hora de creación, la dirección del primer cluster donde están los datos, y por último, el tamaño que ocupa.

El directorio raíz ocupa una posición concreta en el sistema de archivos, pero los índices de otras carpetas ocupan la zona de datos como cualquier otro archivo.

Los nombres largos se almacenan ocupando varias entradas en el índice para el mismo archivo o carpeta.



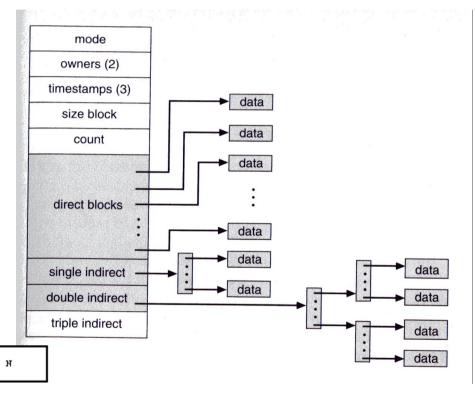
Los conceptos centrales son:

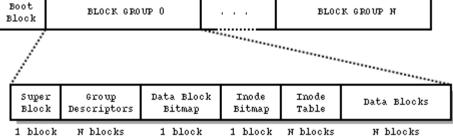
- 1.- superbloque,
- 2.- nodo-i,
- 3.- bloque de datos,
- 4.- bloque de directorio, y
- 5.- bloque de indirección.



## Ext2/3/4

- Superbloque,
- Tabla del grupo
- Tabla bitmap de datos
- Tabla bitmap de inodes
- Nodo-i,
- Bloque de datos,





## **Caracteristicas Ext4**

- Permite archivos de 1(EiB) y hasta 16(TiB)
- Extends
- Compatibilidad con Ext3/2
- Pre alocamiento de espacio
- Alocamiento demorado
- Sin limitaciones de directorios
- Mejoramiento en el timestamp
- Encriptación transparente
- Journal Checksum



El superbloque tiene información del sistema de archivos en conjunto, como su tamaño (la información precisa aquí depende del sistema de archivos).

Un nodo-i tiene toda la información de un archivo, salvo su nombre. El nombre se almacena en el directorio, junto con el número de nodo-i. Una entrada de directorio consiste en un nombre de archivo y el número de nodo-i que representa al archivo.

El nodo-i contiene los números de varios bloques de datos, que se utilizan para almacenar los datos en el archivo. Sólo hay espacio para unos pocos números de bloques de datos en el nodo-i; en cualquier caso, si se necesitan más, más espacio para punteros a los bloques de datos son colocados de forma dinámica.

Estos bloques colocados dinámicamente son bloques indirectos; el nombre indica que para encontrar el bloque de datos, primero hay que encontrar su número en un bloque indirecto.



- **Superblock**: Contiene toda la información de la configuración del sistema de archivo. Contiene información el total de inodos y bloques del sistemas de archivo y cuantos estan libres, cuando fue montado, etc.
- Inode: Cada archivo esta representado por una estructura llamado un inode.
   Cada inodo contiene la descripción del fichero, tipo de archivo, los derechos de acceso, marcas de tiempo, el tamaño y los punteros a los bloques de datos. El bloque de indirección es un inodo que se usa para apuntar a otros inodos cuando la cantidad de inodos se acaban.
- Directorios: Los directorios se estructuran en un arbol jerárquico. Cada directorio puede contener archivos o directorios. Los directorios son un tipo especial de archivos.
- **Enlaces:** Sistemas de archivos Unix implementar el concepto de enlace. Varios nombres se pueden asociar con un inodo. El inode contiene un campo que contiene el número asociado con el archivo.
- Archivos especiales de dispositivo: En los sistemas operativos tipo Unix, los dispositivos pueden acceder a través de ficheros especiales. Un archivo especial de dispositivo no utiliza ningún espacio en el sistema de ficheros. Es sólo un punto de acceso al controlador de dispositivo. Existen dos tipos de archivos especiales: carácter y bloquear archivos especiales. La primera permite operaciones I / O en modo de caracteres mientras que el segundo requiere datos a escribir en modo bloque a través de las funciones de caché del búfer. Cuando una solicitud de E / S se realiza en un archivo especial, se envía a un (pseudo) controlador de dispositivo. Un archivo especial hace referencia a un número importante, que identifica el tipo de dispositivo, y un número menor de edad, que identifica la unidad.



• Estructura de Ext4

https://ext4.wiki.kernel.org/index.php/Ext4\_Disk\_Layout



### • Enlaces:

- http://web.mit.edu/tytso/www/linux/ext2intro.html
- http://www.nongnu.org/ext2-doc/ext2.html#DEFINITIONS
- http://upload.wikimedia.org/wikipedia/commons/a/a2/Ext2-inode.gif
- http://cs.smith.edu/~nhowe/262/oldlabs/img/ext2.png
- https://www.cs.cmu.edu/~mihaib/fs/fs.html

# Sistema de Archivo

Del ejercicio 5 represéntese mediante un esquema esa misma situación del disco pero suponiendo que en vez de una FAT se dispone de un sistema de asignación tipo UNIX. Supóngase una estructura de inodo con 3punteros directos, 1 índice simple indirecto, 1 índice doble indirecto y 1 índice triple indirecto.

Dir Simplificada	FAT		Asignado a
A.txt → 6	0	X	
B.txt → 8	1	EOF	C.txt
C.txt → 12	2	BAD	
	3	EOF	
	4	15	A.txt
	5	FREE	
	6	4	A.txt
	7	13	B.txt
	8	7	B.txt
	9	EOF	A.txt
	10	14	C.txt
	11	FREE	
	12	16	C.txt
	13	EOF	B.txt
	14	1	C.txt
	15	9	A.txt
	16	10	C.txt

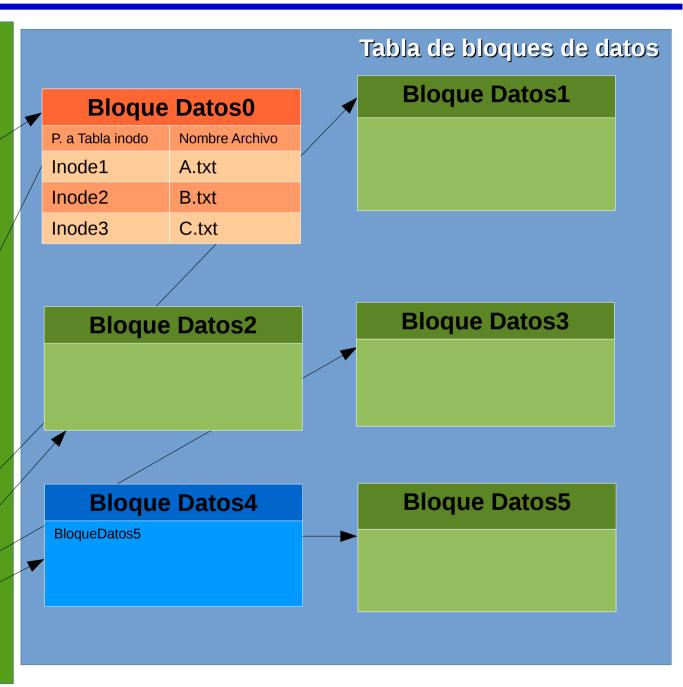


# Sistema de Archivo

### Tabla de inodos

Inode0 (root directory)				
Tipo	Dir			
P. Dir 1	Bdatos 0			
P. Dir 2	Null			
P. Dir 3	Null			
P. Ind. Simple	Null			
P. Ind. Doble	Null			
P. Ind. Triple	Null			

Inode1 (root directory)			
Tipo	Dir		
P. Dir 1	Bdatos 1		
P. Dir 2	Bdatos 2		
P. Dir 3	Bdatos 3		
P. Ind. Simple	Bdatos 4		
P. Ind. Doble	Null		
P. Ind. Triple	Null		





#### Caracteristicas NTFS

- Cuotas
- Maximo Tamaño: 256TiB con cluster 64KiB. Tamaño maximo del archivo: 16EiB a 1KiB o 16TiB a 64KiB.
- Date resolution 100ns.
- Jornaling
- Hard Links
- Alternative Data Stream: Permite mas de un archivo con el mismo nombre. Formato filename:streamname. Comandos del PowerShell Add-Content, Clear-Content, Get-Content, Get-Item, Out-String, Remove-Item, Set-Content.
- Compresión
- Sparse File. Archivos con espacios libres.
- Volumen Shadow Copy (o snapshot). Realiza una copia de todo archivo o directorio modificado.
- Encryption
- Puntos de Montado
- Deduplicación.



### Organización de Volumen NTFS

- Se tiene cuatro estructuras basicas:
  - NTFS Boot Sector --> almacena información sobre el diseño del volumen y las estructuras del sistema de archivos, el codigo de arranque de algunos windows servers.
  - Master File Table (MTF) --> Contiene la información necesaria para recuperar los archivos de la partición NTFS, como los atributos de un archivo.
  - File System Data --> Almacena los datos que no se encuentran dentro de la tabla maestra de archivos.
  - Master File Table Copy --> Copia del MFT.

#### Estructura NTFS

https://technet.microsoft.com/en-us/library/cc781134(WS.10).aspx



### MFT (Metadata File Table)

- Habrá una fila o record (entrada en la tabla) por cada archivo que exista en el sistema de archivos.
- Ahora bien, el concepto de archivo es bastante amplio: toda entidad almacenada individualmente es un archivo. La propia MFT es considerada un archivo y tiene su propia entrada en la MFT.
- Las primeras 16 entradas de la MFT están reservadas para almacenar archivos del sistema, es decir, información sobre el propio sistema de archivos. El resto de entradas se corresponden con los archivos y carpetas de datos.



System File	File Name	MFT Record	Purpose of the File
Master file table	\$Mft	0	Contiene en el registro de archivo todos los archivos y directorios en el volumen NTFS. Si supera el registro se utiliza otro.
MTF Mirro	\$MtfMirr	1	Duplica el registro MFT para mejor los problemas de falla.
Log file	\$LogFile	2	Contiene información para recuperacion. Mantiene consistencia de los metadatos despues de una falla.
Volume	\$Volume	3	Contine información sobre el volumen, como el nombre o la versión
Attribute definitions	\$AttrDef	4	Lista de nombres, numeros y descriptores.
Root file name index		5	La carpeta raiz
Cluster bitmap	\$Bitmap	6	Representa en bits los cluster usados y no usuados.
Boot sector	\$Boot	7	Incluye el BPB y el codigo de carga adicional si el volumen es booteable.
Bad cluster file	\$BadClus	8	Contine la lista de clusters dañados.
Security file	\$Secure	9	Contiene los descriptores unicos de seguridad para todos los archivos del volumen.
Upcase table	\$Upcase	10	Convierte caracteres para machear los caracteres UNICODE upcase.
NTFS extension file	\$Extend	11	Información adicional, como cuotas.
		12-15	Reservado para futuros usos.



### **Enlaces Sugeridos:**

- <a href="http://ftp.kolibrios.org/users/Asper/docs/NTFS/ntfsdoc.html">http://ftp.kolibrios.org/users/Asper/docs/NTFS/ntfsdoc.html</a>
- https://technet.microsoft.com/en-us/library/cc781134%28v=ws.10%29.aspx
- <a href="http://www.tuxera.com/community/ntfs-3g-manual/">http://www.tuxera.com/community/ntfs-3g-manual/</a>



Es una capa virtual que sirve como capa de abstracción del sistema de archivo real. Este capa es la que utiliza el usuario para acceder al sistema de archivo real. Esta capa permite controlar y/o manejar de la misma manera un sistema de archivo local como un sistema de archivo de red.

El primer mecanismo de sistema de archivo virtual en un sistema Unix fue instroducido en el SunOS de 1985. Esto permitia acceder a su sistema de archivo local llamado UFS como acceder a sistemas remotos NFS en forma transparente.

Podría decirse que el servicio más importante que la capa VFS ofrece una caché de datos I/O uniforme. Linux mantiene cuatro cachés de datos de E/S: page cache , i-node cache , buffer cache y directory cache. El cache de páginas (page cache) combina datos de la memoria y archivos virtuales. El buffer cache es una interfaz con el dispositivo de bloques y cachea los meta-datos de discos de bloques. El cache de i-nodo mantiene el acceso reciente a los i-nodos de archivo. El cache de directorio (d-cache) mantiene en la memoria de un árbol que representa una parte de la estructura de directorios del sistema de archivos.

La estructura de VFS de linux se parece a la estructura de ext2.



### VFS Superblock

Todo sistema de archivo montado esta representado por el VFS superblock. El VFS contiene:

- Device/Dispositivo: Este es el identificador de dispositivo. por ejemplo /dev/sda1
- Puntero a inodos/Inode pointers: Apunta al primer inodo del sistema de archivo.
   El puntero de inodo cubierto representa el inodo del directorio donde esta montado dentro del sistema. Si es el root file system no tiene puntero cubierto.covered pointer,
- Tamaño de Bloque/Blocksize: es el tamaño de bloque en el sistema de archivo. ej 4096 bytes.
- Operaciones de superbloque/Superblock operations: Apunta a grupo de rutinas de superbloque. Entre otras cosas estas rutinas son las que permite la lectura y escritura de inodos y superbloques.
- Tipo de sistema de archivo/File System type: Es el puntero a los datos de estrutura del tipo de sistema de archivo.
- Especifico del sistema de archvio/File System specific: Es el puntero a la información necesario para este sistema de archivo.



#### El VFS inodo

Como el sistema de archivo Ext2, todo archivo, directorio y otros es represnetado por algún VFS inodo. FS inode.

La información de cada inodo VFS es construida desde la información extraida del sistema de archivo mediante rutinas especificas. Cada VFS inodo existe solo en memoria del sistema y es mantenida dentro del cache de inodo. VFS inodo contiene los siguientes datos:

- Dispositivo/device: Es el identificador de dispositivo donde se encuentra el inodo real.ode represents,
- Numero de Inodo/inode number: Es un numero unico de inodo dentro del sistema de archivo. Es una convinación entre el dispositivo y el el numero de inodo.
- Modo/mode: Como el EXT2 este campo describe los permisos al inodo.
- ID de usuario/user ids: Describe los ids de usuario
- Tiempos/times: De creación, modificación, y de escritura.
- Tamaño de Bloque/block size: Tamaño de bloques.
- Operaciones de inodo/inode operations: Es un puntero a la direcciones de los procedimientos de inodo. Por ejemplo la opeacion de truncar el archivo que representa el inodo.
- Cuenta/count: Es el numero de componetes de sistema que estan usuando este inodo. Cuando la suma llega a cero puede ser descartado o reusado.
- lock: Este campo es usado cuando el inodo es lockeado cuando es usado por el sistema.
- Sucio/dirty: Indica si el inodo VFS ha sido escrito y el sistema de archivo por debajo necesita ser modificado.



Registro de sistema de archivos

Todo en linux todo nuevo sistema de archivo debe ser registrado. Los sistemas de archivos pueden ser creados como modulos y cargados bajo demanda. Cuando son cargados el kernel registra el sistema de archivo.

Uno puede ver si el archivo esta registrado haciendo un cat al archivo /proc/filesystems.



### Bibliografía

- http://www.ibm.com/developerworks/library/l-fuse/
- http://www.tldp.org/LDP/tlk/fs/filesystem.html
- http://www.tldp.org/LDP/khg/HyperNews/get/fs/vfstour.html
- https://kaiwantech.files.wordpress.com/2009/08/vfs\_msdos\_31.png
- http://www.inf.fu-berlin.de/lehre/SS01/OS/Lectures/Lecture16.pdf
- The Linux VFS, Chapter 4 of Linux File Systems by Moshe Bar (McGraw-Hill, 2001). ISBN 0-07-212955-7
- Chapter 12 of Understanding the Linux Kernel by Daniel P. Bovet, Marco Cesati (O'Reilly Media, 2005). ISBN 0-596-00565-2



FUSE permite implementar un sistema de archivo funcional en el espacio de usuario de programas. Las caracteristicas que incluyen son:

- Una libreria simple.
- El modulo esta incorpado al kernel.
- Correr el sistema de archivo en "userspace", esto impliaca que no necesita ser usuario privilegiado.
- Corre en linux 2.4.X en adelante.
- Ha probado ser estable.

FUSE fue desarrollado para soportar el sistema de archivo AVFS, posteriomente se convirtio en un proyecto aparte.



```
Struct fuse operations {
   int (*getattr) (const char *, struct stat *);
   int (*readlink) (const char *, char *, size_t);
   int (*getdir) (const char *, fuse dirh t, fuse dirfil t);
   int (*mknod) (const char *, mode t, dev t);
   int (*mkdir) (const char *, mode t);
   int (*unlink) (const char *);
   int (*rmdir) (const char *);
   int (*symlink) (const char *, const char *);
   int (*rename) (const char *, const char *);
   int (*link) (const char *, const char *);
   int (*chmod) (const char *, mode t);
   int (*chown) (const char *, uid t, gid t);
   int (*truncate) (const char *, off t);
   int (*utime) (const char *, struct utimbuf *);
   int (*open) (const char *, struct fuse file info *);
   int (*read) (const char *, char *, size t, off t, struct fuse file info *);
   int (*write) (const char *, const char *, size t, off t, struct fuse file info *);
   int (*statfs) (const char *, struct statfs *);
   int (*flush) (const char *, struct fuse file info *);
   int (*release) (const char *, struct fuse file info *);
   int (*fsync) (const char *, int, struct fuse file info *);
   int (*setxattr) (const char *, const char *, const char *, size t, int);
   int (*getxattr) (const char *, const char *, char *, size t);
   int (*listxattr) (const char *, char *, size t);
   int (*removexattr) (const char *, const char *);
};
```



```
Ejemplo: Hello Word en FUSE
                        ----- hello.c -----
/*
  FUSE: Filesystem in Userspace
  Copyright (C) 2001-2007 Miklos Szeredi <miklos@szeredi.hu>
  This program can be distributed under the terms of the GNU GPL.
  See the file COPYING.
  gcc -Wall hello.c `pkg-config fuse --cflags --libs` -o hello
*/
#define FUSE USE VERSION 26
#include <fuse.h>
#include <stdio.h>
#include <string.h>
#include <errno.h>
#include <fcntl.h>
static const char *hello str = "Hello World!\n";
static const char *hello path = "/hello";
```



```
// Operacion de lectura de atributos.
static int hello getattr(const char *path, struct stat *stbuf)
     int res = 0;
    memset(stbuf, 0, sizeof(struct stat));
     if (strcmp(path, "/") == 0) {
         stbuf->st mode = S IFDIR | 0755;
         stbuf->st nlink = 2;
     } else if (strcmp(path, hello path) == 0) {
         stbuf->st mode = S IFREG | 0444;
         stbuf->st nlink = 1;
         stbuf->st size = strlen(hello str);
     } else
         res = -ENOENT;
     return res;
}
// Operacion de lectura de directorio
static int hello readdir(const char *path, void *buf, fuse fill dir t filler,
                off t offset, struct fuse file info *fi)
{
     (void) offset;
     (void) fi;
     if (strcmp(path, "/") != 0)
         return -ENOENT;
     filler(buf, ".", NULL, 0);
     filler(buf, "..", NULL, 0);
     filler(buf, hello path + 1, NULL, 0);
    return 0;
}
```



```
// Operacion de apertura.
static int hello open(const char *path, struct fuse file info *fi)
     if (strcmp(path, hello path) != 0)
          return -ENOENT;
     if ((fi->flags & 3) != O RDONLY)
          return -EACCES;
     return 0;
}
// Operacion de lectura
static int hello read(const char *path, char *buf, size t size, off t offset,
                struct fuse file info *fi)
{
     size t len;
     (void) fi;
     if(strcmp(path, hello path) != 0)
          return -ENOENT;
     len = strlen(hello str);
     if (offset < len) {</pre>
          if (offset + size > len)
               size = len - offset;
          memcpy(buf, hello str + offset, size);
     } else
          size = 0;
     return size;
```

Organización de Datos

```
// Definicion de operaciones...
static struct fuse_operations hello_oper = {
          .getattr = hello_getattr,
          .readdir = hello_readdir,
          .open = hello_open,
          .read = hello_read,
};
int main(int argc, char *argv[])
{
    return fuse_main(argc, argv, &hello_oper, NULL);
}
```





#### Bibliografía

http://fuse.sourceforge.net/

http://fuse.sourceforge.net/doxygen/null 8c.html

https://code.google.com/p/pngdrive/

http://www.buanzo.com.ar/lin/FUSE.html