Departament d'Arquitectura de Computadors

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Conceptes Avançats de Sistemes Operatius

Facultat d'Informàtica de Barcelona
Dept. d'Arquitectura de Computadors

Curs 2019/20 Q2

Sistemes de fitxers

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Índex

- Gestió de quotes
- Journaling
- Sistemes de fitxers en xarxa
- Mecanisme de *boot*

Índex

- Gestió de quotes
- Journaling
- Sistemes de fitxers en xarxa
- Mecanisme de *boot*

Gestió de quotes

- Quota, què és?
  - Habilitat de limitar la quantitat de dades que un usuari (o grup d'usuaris) té en un sistema de fitxers (partició)
- Mecanisme
  - Independent del sistema de fitxers
- Requereix
  - Que el sistema de fitxers les suporti
  - Que el kernel les suporti

Gestió de quotes

- Preparació de la partició
  - Ha de ser muntada amb les opcions 'usrquota' i/o 'grpquota'
    - Es pot usar /etc/fstab
      - /dev/sda9 /home ext4 defaults,usrquota,grpquota 1 1
  - Comanda quotacheck per crear els fitxers de quota
    - quotacheck -vagum
      - verbose, all, group, user, no-remount
    - Crea
      - /aquota.user
      - /aquota.group

Gestió de quotes

- Activació i aturada de les quotes
  - /sbin/quotaon -avug (all, verbose, user, group)
    - Activa el mecanisme de quotes
  - /sbin/quotaoff per desactivar-lo
- Edició de quotes
  - edquota, obre un editor, estil crontab

```
Disk quotas for user xavim (uid 500):
  Filesystem                blocks       soft       hard     inodes     soft     hard
  /dev/loop0                    32         16         32          2        0        0
```

- Examinar les quotes: quota -v

```
Disk quotas for user xavim (uid 500):
     Filesystem  blocks   quota   limit   grace   files   quota   limit
grace
     /dev/loop0      32*       16      32     6days       2       0       0
```

Gestió de quotes

- "Grace period"
  - Temps durant el qual l'usuari pot arribar al límit "hard", només amb warnings per part del sistema
  - Si expira el "grace period", llavors el sistema de quotes ja no deixa passar del "soft" límit

Índex

- Gestió de quotes
- Journaling
- Sistemes de fitxers en xarxa
- Mecanisme de *boot*

Journaling

- Habitualment, les operacions sobre fitxers inclouen diverses operacions al disc
  - Exemple: esborrar un fitxer
    - Esborrar l'entrada del directori
    - Marcar l'i-node com a lliure a la taula d'i-nodes
    - Marcar els blocs de dades com a lliures a la taula de blocs
  - Si el sistema s'apaga en mig d'aquests passos...
    - Pot quedar un conjunt de blocs de dades ocupats i sense nom
    - L'entrada del directori pot quedar apuntant a un conjunt de blocs alliberats → poca seguretat!

Recuperació costosa

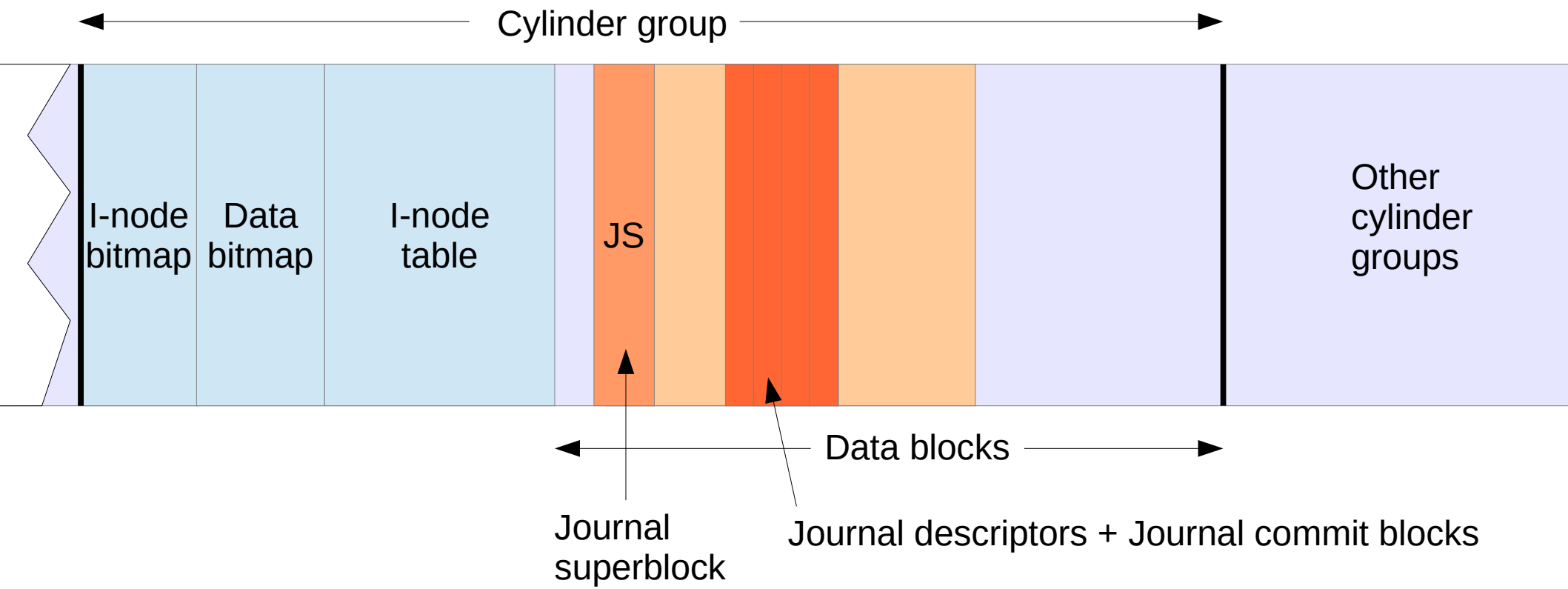
- Per arreglar aquests errors, cal recórrer completament
  - l'arbre de directoris i fitxers
  - les estructures d'i-nodes i mapes de blocs de dades
- Temps de test i recuperació
- Solució: journal

Journaling

- Garanteix consistència del sistema de fitxers
  - Fallades de corrent elèctric
  - Fallades del sistema
- Pot guardar-se en un disc diferent
  - Per minimitzar contenció al disc
    - Lectures i escriptures de dades i journal al mateix temps
- Les escriptures al journal es fan **asíncrones** per reduir l'impacte en el rendiment

Journaling

- Inclou una estructura de dades de suport a la recuperació del sistema de fitxers



Analysis and Evolution of Journaling File Systems
Vijayan Prabhakaran, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau
http://www.cs.wisc.edu/wind/Publications/sba-usenix05.pdf

Gestió del journal

- Al journal cal escriure per avançat, respecte a la resta del disc
  - Introduir dependències entre operacions
- Els canvis escrits en el journal són atòmics:
  - En recuperació, mai es repetirà una seqüència d'operacions que no estigui sencera en el journal
  - Cada seqüència inclou una suma de comprovació
    - Si la suma és incorrecte, no es reproduirà durant la recuperació

Tipus de journals

- Físic
  - Grava una còpia de cada bloc!
- Lògic
  - Grava només els canvis a les metadades del sistema de fitxers
  - Poden tenir corrupció de dades
    - Però no de l'estructura del sistema de fitxers

Journaling

- Ext4
  - Basat en ext3, sense afegir incompatibilitats
    - Limit ext3: 8TB punters a blocs de 32 bits
  - Suport per sistemes de fitxers grans
    - 48 bits d'adreça a bloc → 1 EB (2^{60} bytes)
    - Extents: milloren el tractament de blocs contigus en el fitxer i en disc
  - Tamany de block gran: 4KB – 1MB
    - **En realització, impacta a les utilitats, i al sistema**
    - **Fragmentació interna: també pels directoris…**
  - Temps en alta resolució: nanosegons
  - Incorpora suport per quotes en el propi sistema

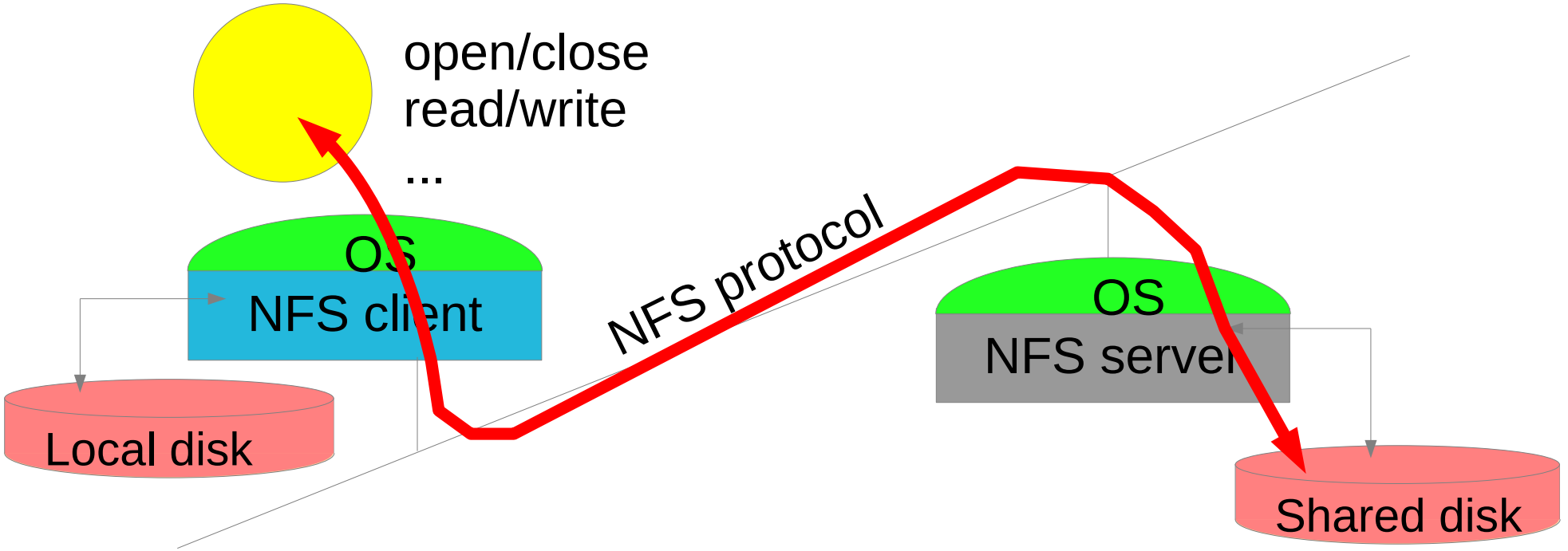
https://ext4.wiki.kernel.org/index.php/Ext4_Design

Índex

- Gestió de quotes
- Journaling
- Sistemes de fitxers en xarxa
- Mecanisme de *boot*

Sistemes de fitxers en xarxa

- Network File System (NFS)
  - Tranparent als usuaris
  - Implementat sobre Remote Proc. Calls (RPCs)
  - Centralitzat en un servidor



NAS vs SAN

- Network-attached storage (NAS)
  - Servidor de dades a nivell de fitxer
  - Unitats d'emmagatzematge sovint en RAID
  - Connectat als seus clients en xarxa via NFS, SMB o AFP
- Storage area network (SAN)
  - Xarxa dedicada que serveix dades a nivell de block
  - Arrays de discos en RAID
  - Connectats als seus clients via un switch Fiber Channel, firmware i drivers.

Sistemes de fitxers en xarxa

- AFS, Andrew File System
  - Distribuït en diferents servidors
    - Presenta una visió homogènia dels fitxers independent de la localització de l'usuari
  - OpenAFS – Linux, MAC, Windows

Object Exchange (OBEX)

- Protocol d'intercanvi de dades amb dispositius mòbils
  - fitxers de tot tipus
  - Entrades de calendari
  - Tarjetes de visita

http://openobex.sourceforge.net/about.html

Object Exchange (OBEX)

- Xarxes
  - USB
  - Infrared (IrDA, IrLAN)
  - Bluetooth
  - Serial ports / ttys
- Sistemes de fitxers
  - FUSE – Filesystem in User Space

http://fuse.sourceforge.net/

FUSE

- Configuració i mòdul en el kernel
  - CONFIG_FUSE_FS=m fuse.ko
- Llibreries
  - /lib64/libfuse.so
- Comanda
  - /bin/fusermount – setuid a root
- Interfície
  - fuse_main(argc, argv, &operations, NULL);

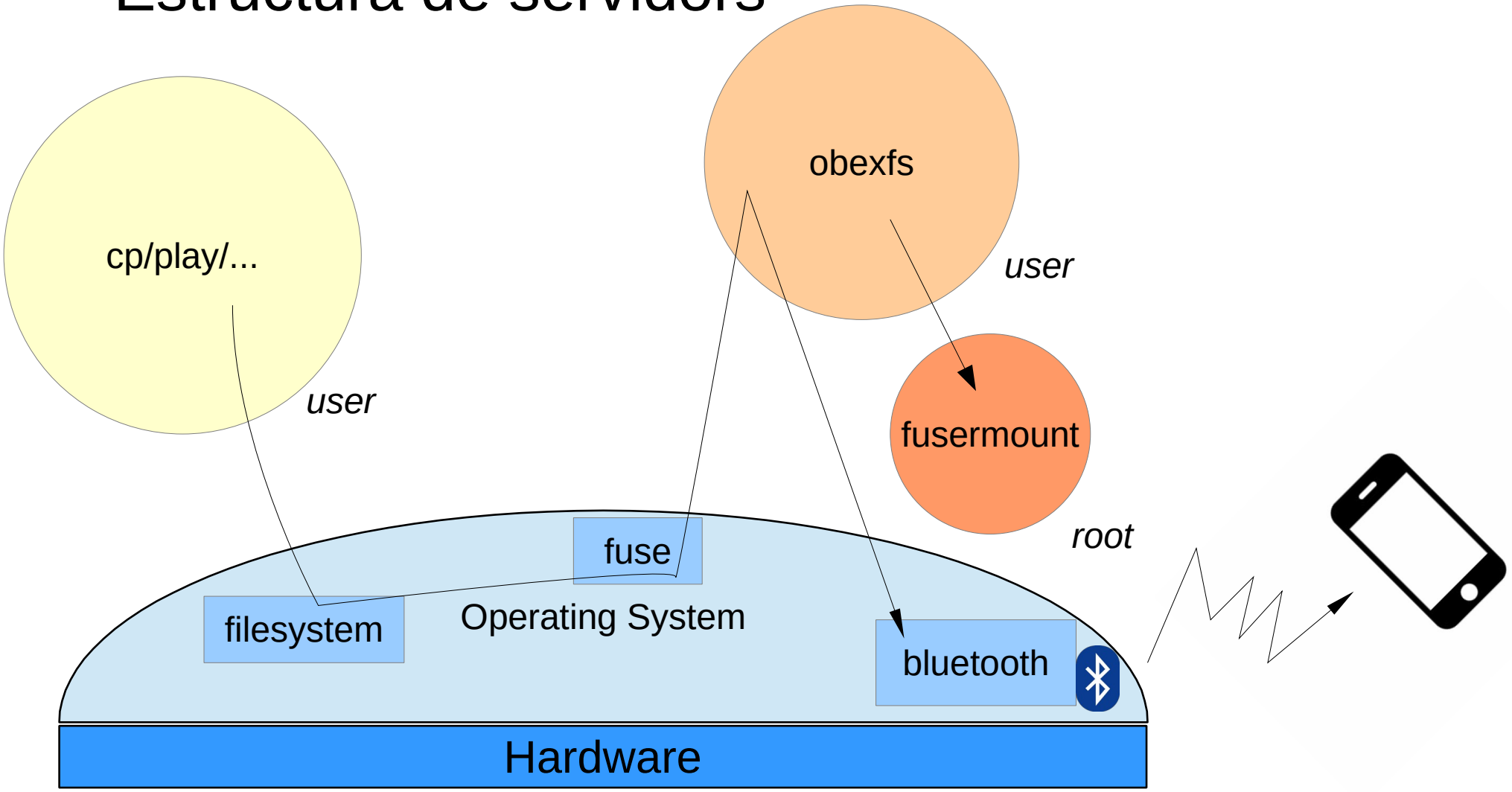
FUSE

- Interfície
  - open / create / read / write / fsync / flush / release
  - get file attributes / stat / statfs / access
  - symlink / readlink / link / rename
  - mknod / mkdir
  - unlink / rmdir
  - opendir / readdir / releasedir / fsyncdir
  - chmod / chown
  - truncate
  - +++

Exemple
https://github.com/zuckschwerdt/obexfs/blob/master/fuse/obexfs.c

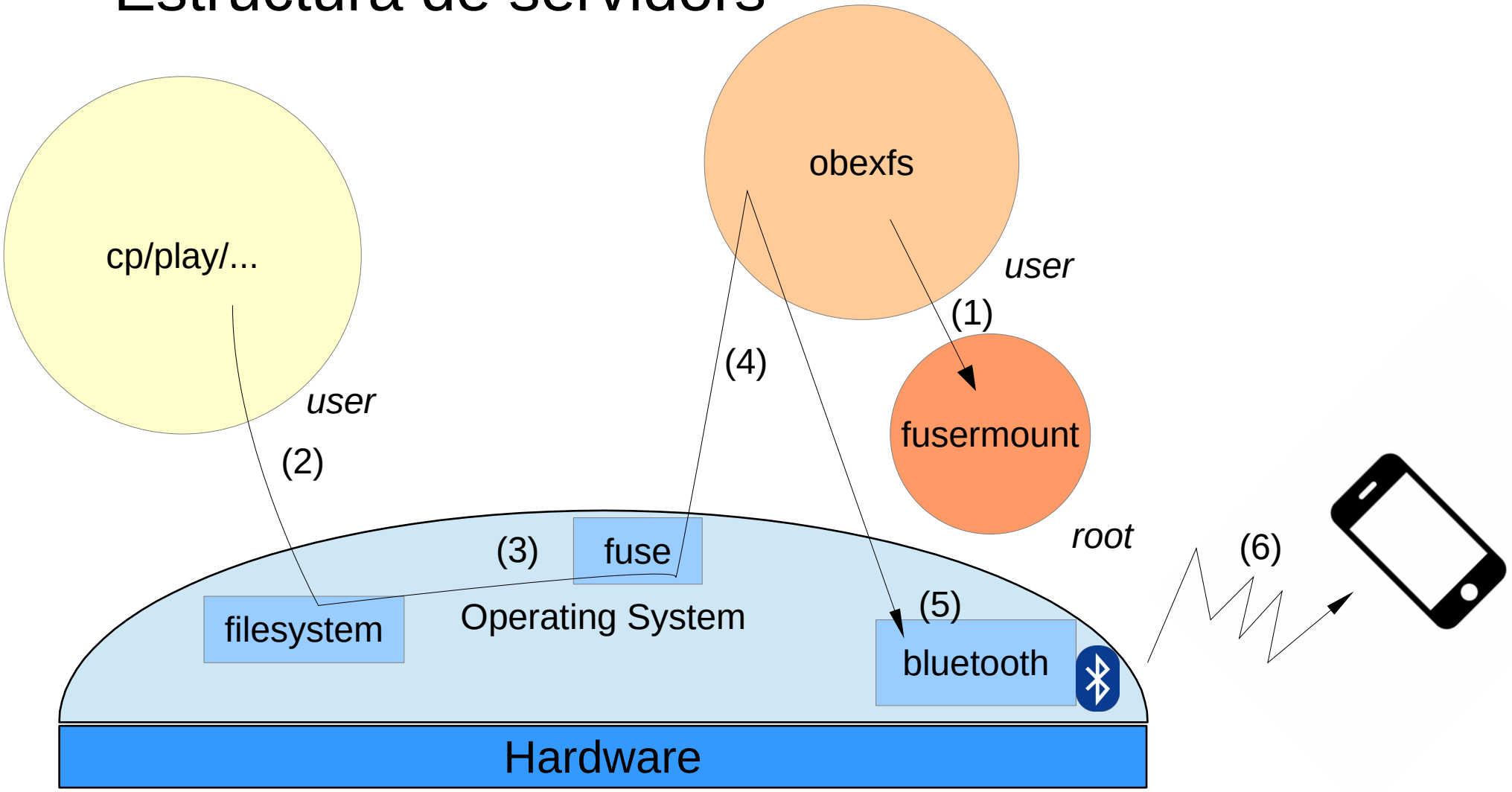
OBEX + FUSE

- Estructura de servidors



OBEX + FUSE

- Estructura de servidors



OBEX + FUSE (exemples)

- Llistar el contingut d'un directori
  - obexftp -b DC:tt:zz:aa:xx:yy -c <path> -l
- Transferir fitxers al dispositiu
  - obexftp -b DC:tt:zz:aa:xx:yy -c <path> --put <file>
- Transferir fitxers des del dispositiu
  - obexftp -b DC:tt:zz:aa:xx:yy -c <path> --get <file>
- Muntar el dispositiu a <mountpoint>
  - obexfs -b DC:tt:zz:aa:xx:yy – – <mountpoint>

Activitat de Bluetooth

- tcpdump permet veure l'activitat del Bluetooth
  - Cal determinar el dispositiu
    - lsusb -v | grep -i bluetooth →
      - Bus 004 Device 004: ID 0a5c:2145 Broadcom Corp. Bluetooth …
    - tcpdump -D →

      1.eth0

      ...

      7.usbmon4 (USB bus number 4)

      …

      12.any (Pseudo-device that captures on all interfaces)

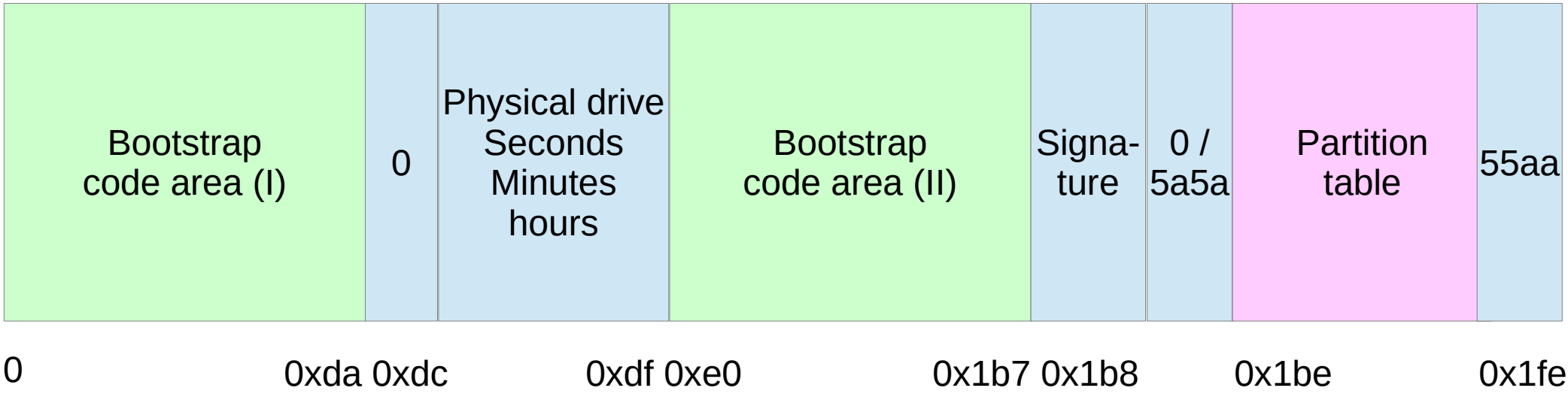
      ...
  - tcpdump -i 7 -w - | od -c

Índex

- Gestió de quotes
- Journaling
- Sistemes de fitxers en xarxa
- Mecanisme de *boot*

BIOS/DOS

- Inici del disc, Master Boot Record (MBR)



BIOS/DOS

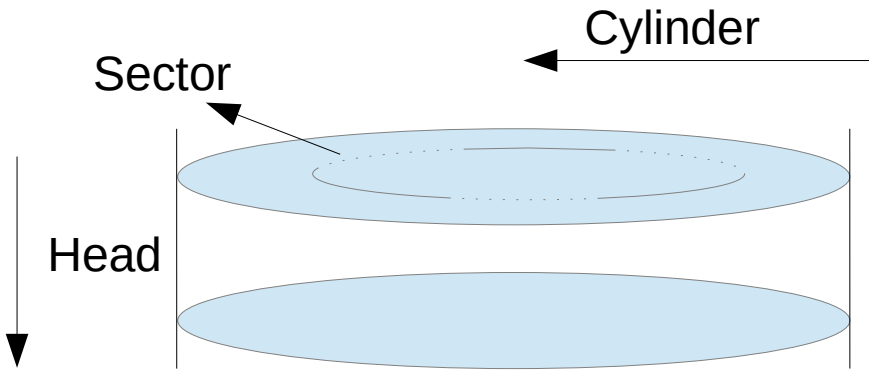
- Inici del disc, Master Boot Record
  - Codi de boot

```
0000000    0eb4    20b0    5850    c0fe    cd50    b410    b986    000f
         ? 016   ?       P   X   ?   ?   P   ? 020   ? 206   ? 017  \0
0000020    40ba    cd42    f415    edeb    ebf4    00fd    0000    0000
         ?   @   B   ? 025   ?   ?   ?   ?   ?   ?  \0  \0  \0  \0  \0
0000040    0000    0000    0000    0000    0000    0000    0000    0000
        \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
*
0000700    0002    0aee    0208    0001    0000    7fff    0000    0000
       002  \0   ?  \n  \b 002 001  \0  \0  \0   ? 177  \0  \0  \0  \0
0000720    0000    0000    0000    0000    0000    0000    0000    0000
        \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
*
0000760    0000    0000    0000    0000    0000    0000    0000    aa55
```

  - Taula de particions

BIOS/DOS

- Taula de particions



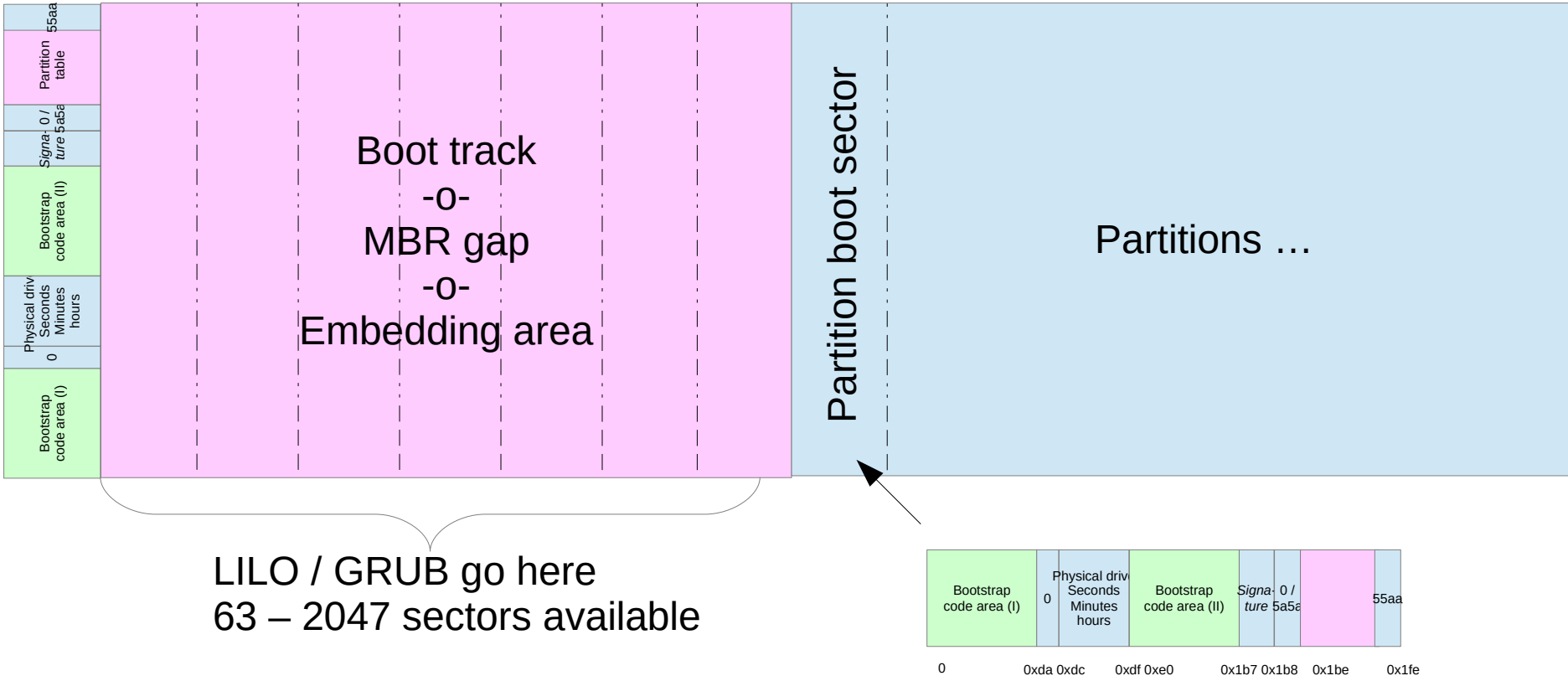
Estructura d'una entrada de la taula de particions

	Status / Physical drive
First sector	Head
	Sector[0-5] / cylinder[8-9]
	Cylinder[0-7]
	Partition type (b, 82, 83...)
Last sector	Head
	Sector[0-5] / cylinder[8-9]
	Cylinder[0-7]
	LBA of first sector in the partition
	Number of sectors in the partition

- LBA (Logical block addressing)
  - Substitueix head/cyl/sector
  - Permet adreçar discos més grans
  - … i amb geometria irregular

BIOS/DOS

- Bootloader installation



http://www.gnu.org/software/grub/manual/grub.html#BIOS-installation

http://www.uefi.org

Nou sistema de *boot*

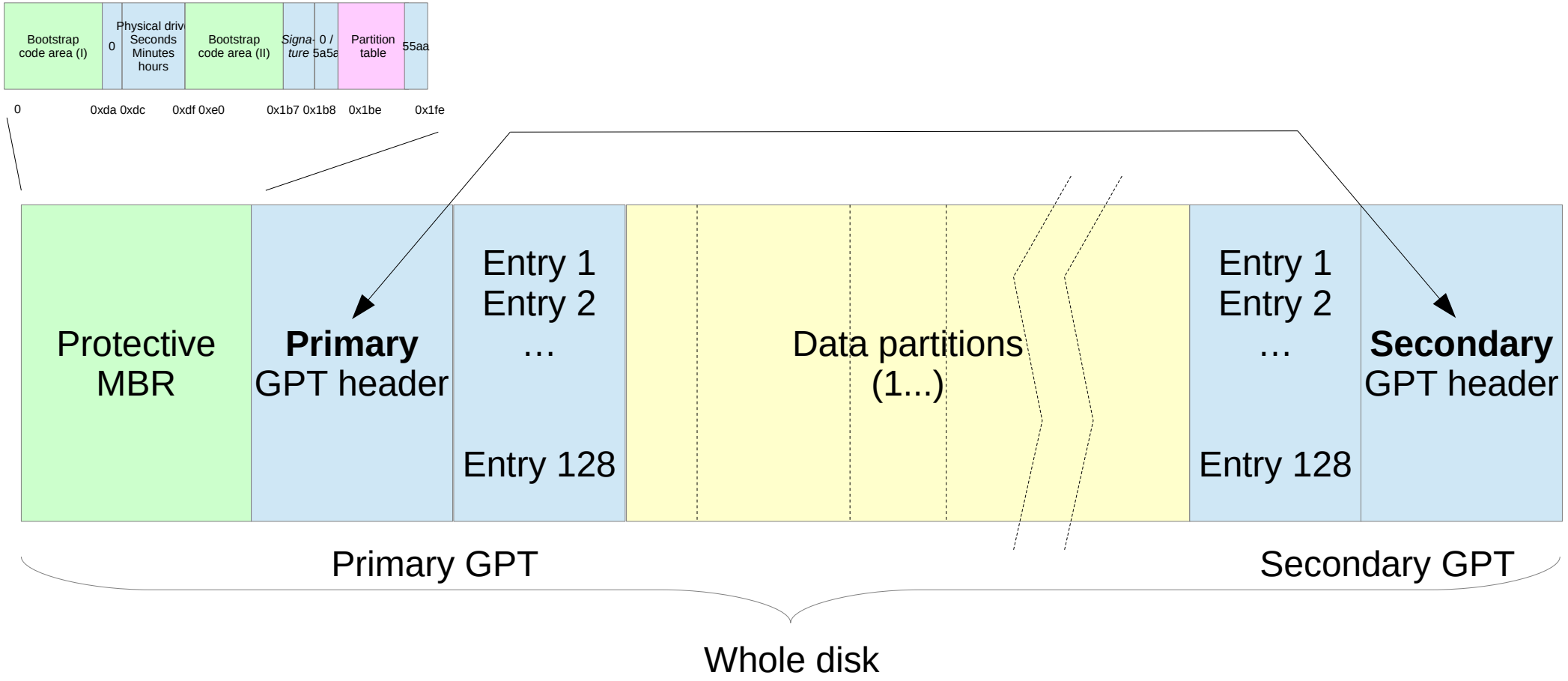
- UEFI
  - Open firmware
  - Unified Extensible Firmware Interface
- Nova taula de particions
  - GPT – GUID Partition Table
    - GUID – Global Unified IDentifiers
- Base
  - Partició FAT32/vfat
  - Binaris de Windows

elilo.conf: ASCII text
elilo.efi: PE32+ executable (EFI application) x86-64 (stripped to external PDB), for MS Windows
vmlinuz: x86 boot sector

http://en.wikipedia.org/wiki/GUID_Partition_Table

GUID Partition Table

- Protective MBR: ajuda a detectar la GPT en sistemes antics
- Disposem de còpia de seguretat al final del disc



EFI

```
root@pcxavim5:~# dd if=/dev/sda bs=1024 count=1 2>/dev/null| od -xc
0000000    0000    0000    0000    0000    0000    0000    0000    0000    0000
         \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
*
0000660    0000    0000    0000    0000    a240    4916    0000    0000
         \0  \0  \0  \0  \0  \0  \0  \0   @   ? 026   I  \0  \0  \0  \0
0000700    0001    feee    ffff    0001    0000    602f    3a38    0000
        001  \0   ?   ?   ?   ? 001  \0  \0  \0   /   `   8   :  \0  \0
0000720    0000    0000    0000    0000    0000    0000    0000    0000
         \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
*
0000760    0000    0000    0000    0000    0000    0000    0000    aa55
         \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0   U   ?
0001000    4645    2049    4150    5452    0000    0001    005c    0000
          E   F   I       P   A   R   T  \0  \0 001  \0   \  \0  \0  \0
0001020    8858    90de    0000    0000    0001    0000    0000    0000
          X 210   ? 220  \0  \0  \0  \0 001  \0  \0  \0  \0  \0  \0  \0
0001040    602f    3a38    0000    0000    0022    0000    0000    0000
          /   `   8   :  \0  \0  \0  \0   "  \0  \0  \0  \0  \0  \0  \0
0001060    600e    3a38    0000    0000    3c03    ed01    048b    4f51
        016   `   8   :  \0  \0  \0  \0 003   <  001   ? 213 004   Q   O
0001100    8f9d    76a1    cb04    6527    0002    0000    0000    0000
        235 217   ?   v 004   ?   '   e 002  \0  \0  \0  \0  \0  \0  \0
0001120    0080    0000    0080    0000    03c4    8291    0000    0000
        200  \0  \0  \0 200  \0  \0  \0   ? 003 221 202  \0  \0  \0  \0
0001140    0000    0000    0000    0000    0000    0000    0000    0000
         \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
*
0002000
```

UEFI

- Pot usar el MBR per protegir una taula de particions que d'altra manera seria fàcil esborrar

```
root@pcxavim5:~# fdisk -l

WARNING: GPT (GUID Partition Table) detected on '/dev/sda'! The util fdisk doesn't support GPT.
Use GNU Parted.


Disk /dev/sda: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 4096 bytes
Disk identifier: 0x4916a240

   Device Boot      Start         End      Blocks   Id  System
/dev/sda1               1   976773167   488386583+  ee  GPT
Partition 1 does not start on physical sector boundary.
```

UEFI

- Fent servir l'eina corresponent:

```
bash-4.2$ sudo /sbin/gdisk -l /dev/sda
GPT fdisk (gdisk) version 0.8.7
Partition table scan:     MBR: protective   BSD: not present     APM: not present  GPT: present

Found valid GPT with protective MBR; using GPT.
Disk /dev/sda: 976773168 sectors, 465.8 GiB       Logical sector size: 512 bytes
Disk identifier (GUID): ED013C03-048B-4F51-9D8F-A17604CB2765
Partition table holds up to 128 entries. First usable sector is 34, last usable sector is 976773134
Partitions will be aligned on 2048-sector boundaries.    Total free space is 4077 sectors (2.0 MiB)

Number  Start (sector)    End (sector)  Size       Code  Name
   1            2048         2050047   1000.0 MiB  2700
   2         2050048         2582527   260.0 MiB   EF00  EFI system partition
   3         2582528         2844671   128.0 MiB   0C01  Microsoft reserved part
   4         2844672       484110335   229.5 GiB   0700  Basic data partition
   5       484110336       584773631   48.0 GiB    8300  Basic data partition
   6       584773632       685436927   48.0 GiB    8300  Basic data partition
   7       685436928       786100223   48.0 GiB    8300  Basic data partition
   8       786100224       786116607   8.0 MiB     8200  Basic data partition
   9       786116608       853225471   32.0 GiB    8300  Basic data partition
  10       853225472       952451071   47.3 GiB    8300  Basic data partition
  11       952453120       976773119   11.6 GiB    2700
bash-4.2$
```

UEFI

- Directoris de Boot
  - Boot loader (.efi)
  - Fitxers de configuració (Elilo, Grub2)
    - elilo.conf, grub.cfg
  - Kernel: vmlinuz...
  - Initrd: initrd.gz...
- Permet també que cada bootloader (e.g. GRUB) tingui un directori per a ell sol
- Linux: efibootmgr

```
.
|-- BOOT
|   `-- boot.sdi
`-- EFI
    |-- Boot
    |   |-- LenovoBT.EFI
    |   |-- License.txt
    |   |-- ReadMe.txt
    |   `-- bootx64.efi
    |-- Lenovo
    |   `-- Boot
    |       |-- boot.stl
    |       |-- bootmgfw.efi
    |       |-- bootmgr.efi
    |-- Microsoft
    |   `-- Boot
    |       |-- boot.stl
    |       |-- bootmgfw.efi
    |       |-- bootmgr.efi
    |-- Slackware
    |   |-- elilo.conf
    |   |-- elilo.efi
    |   `-- vmlinuz
    `-- Android
        |-- elilo.conf
        |-- elilo.efi
        `-- kernel
```

efibootmgr

```
# efibootmgr -v
BootCurrent: 0013
Timeout: 10 seconds
BootOrder:
0013,0015,0007,0008,0000,0001,0002,0003,000C,0014,0009,000A,000B,000D,0012
Boot0000  Setup
Boot0001  Boot Menu
Boot0002  Diagnostic Splash Screen
Boot0003  Lenovo Diagnostics
Boot0004  Startup Interrupt Menu
Boot0005  Rescue and Recovery
Boot0006  MEBx Hot Key
Boot0007* USB CD
Boot0008* USB FDD
Boot0009* ATA HDD0
Boot000A* ATA HDD1
Boot000B* ATA HDD2
Boot000C* USB HDD
Boot000D* PCI LAN
Boot000E* IDER BOOT CDROM
Boot000F* IDER BOOT Floppy
Boot0010* ATA HDD
Boot0011* ATAPI CD
Boot0012* PCI LAN

Boot0013* Slackware     HD(2,1f4800,82000,5862ef46-25d6-4653-9a47-7d3b1b620acf)
                        File(\EFI\Slackware\elilo.efi)
Boot0014* Windows Boot Manager  HD(2,1f4800,82000,5862ef46-25d6-4653-9a47-7d3b1b620acf)
                        File(\EFI\Microsoft\Boot\bootmgfw.efi)
                        WINDOWS.....x...B.C.D.O.B.J.E.C.T.=.
                        {.9.d.e.a.8.6.2.c.-.5.c.d.d.-.4.e.7.0.-.a.c.c.1.-.f.3.2.b.3.4.4.d.4.7.9.5.}.....
Boot0015* Android       HD(2,1f4800,82000,5862ef46-25d6-4653-9a47-7d3b1b620acf)
                        File(\EFI\Android\elilo.efi)
```

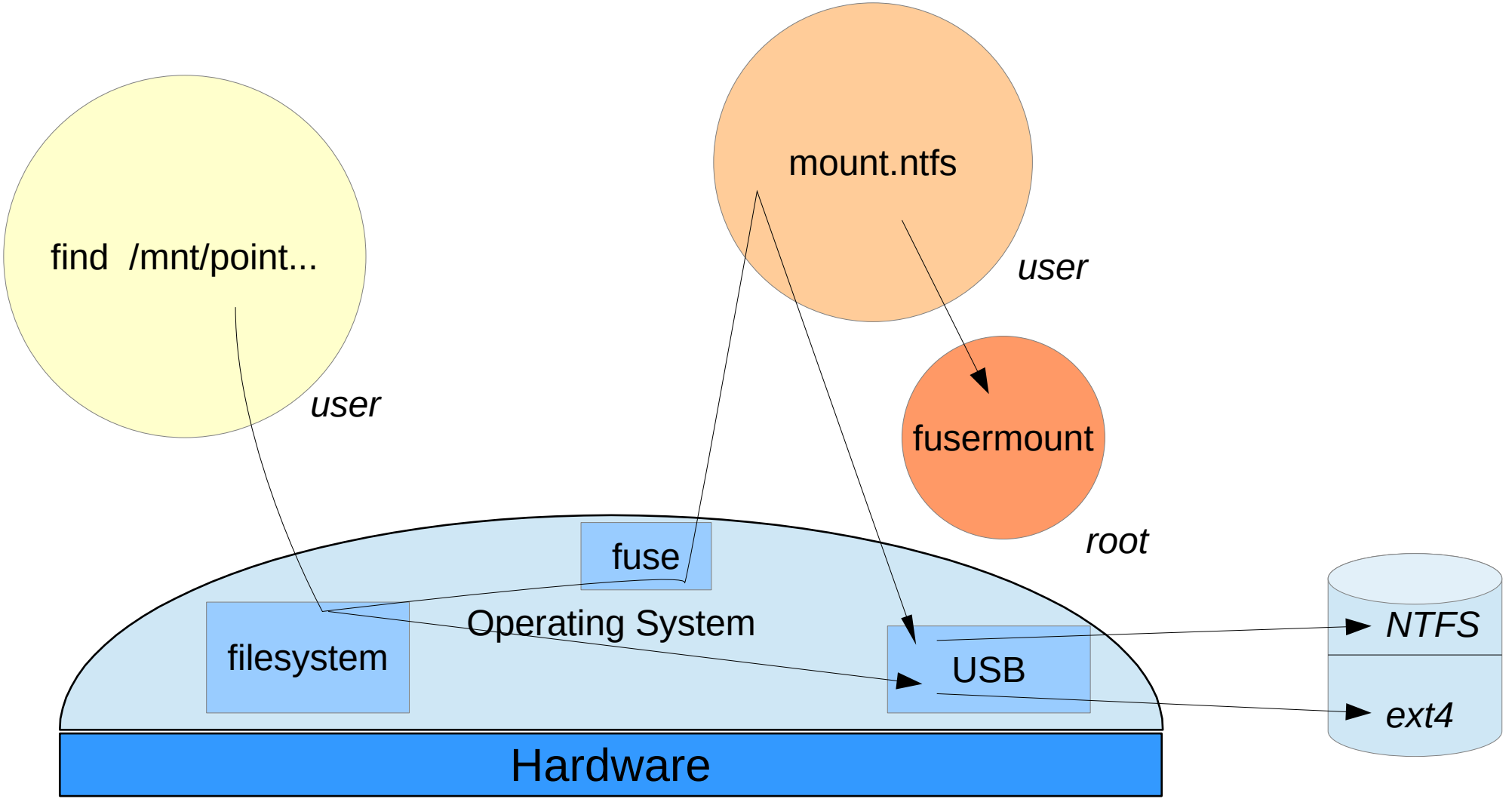
Activitat

- Exclusió mútua i variables de condició
  - En l'exemple
    - condvar-exercise.c

    s'ha omès l'ús de la variable de condició:
    - workers_ready_condition

    la qual cosa fa que el programa no funcioni
  - Determineu com s'han de fer les crides a:
    - pthread_cond_wait
    - pthread_cond_broadcast
    - pthread_cond_signal

    per a que el programa torni a funcionar correctament
  - Proveu també que el programa va més ràpid en fer la feina si definiu:
    - NWORKERS 2 o 4


find /mnt/point...
user
mount.ntfs
user
fusermount
root
fuse
Operating System
filesystem
USB
NTFS
ext4
Hardware