Sistemas de Operação / Fundamentos de Sistemas Operativos

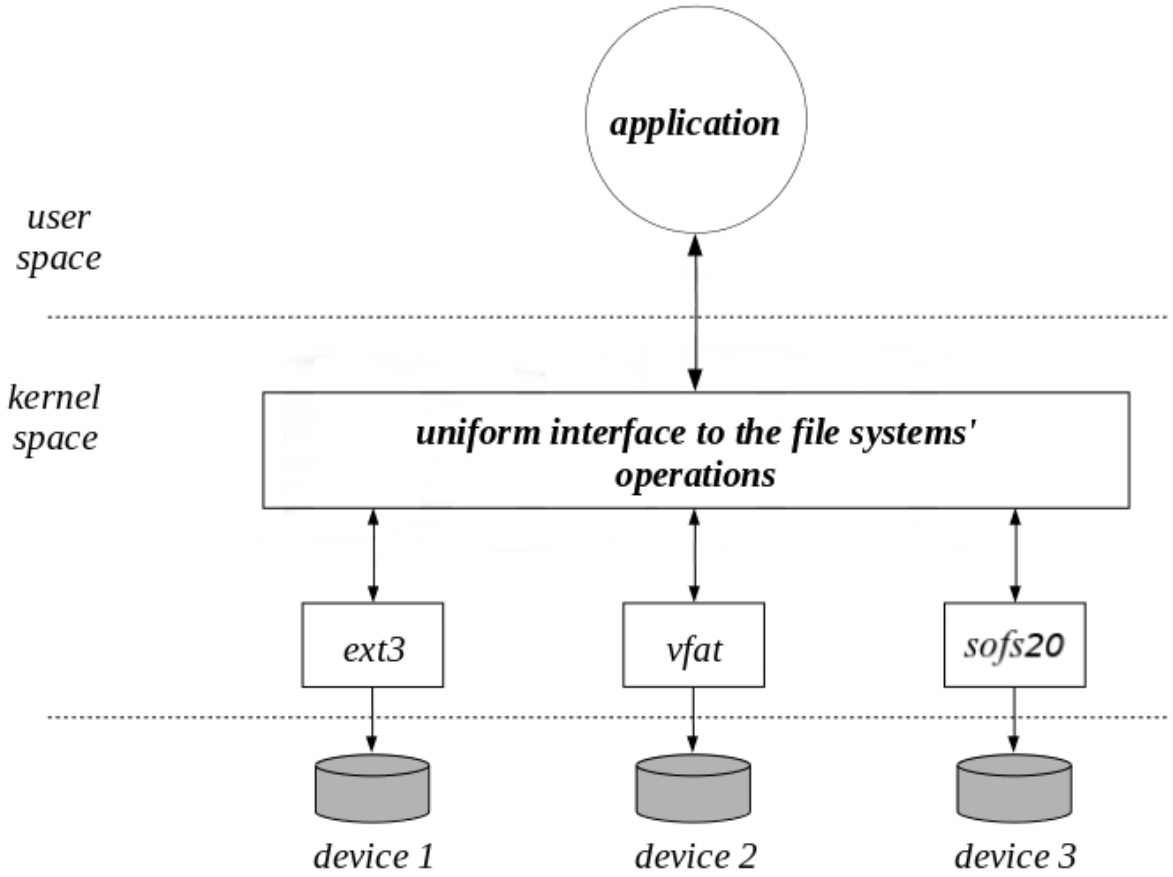
The *sofs20* file system

Artur Pereira <artur@ua.pt>

DETI / Universidade de Aveiro

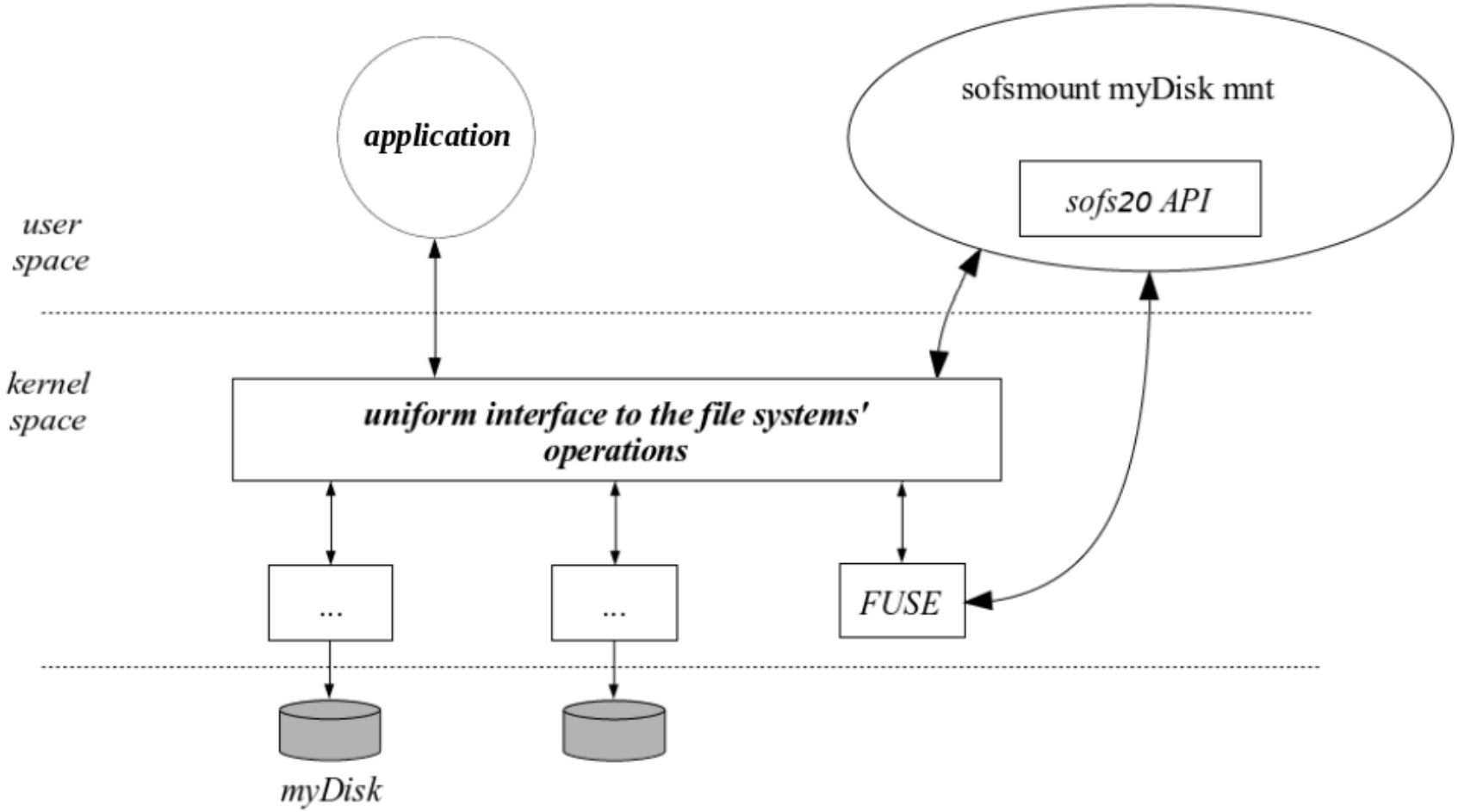
Sumário

The FUSE file system
sofs20 as a kernel module



- Safety issue: running in kernel space
  - Malicious or erroneous code com damage the system

The FUSE file system
sofs20 as a FUSE module



- Safe: running in user space
  - Malicious or erroneous code only affects the user

The sofs20 architecture
Block partitioning

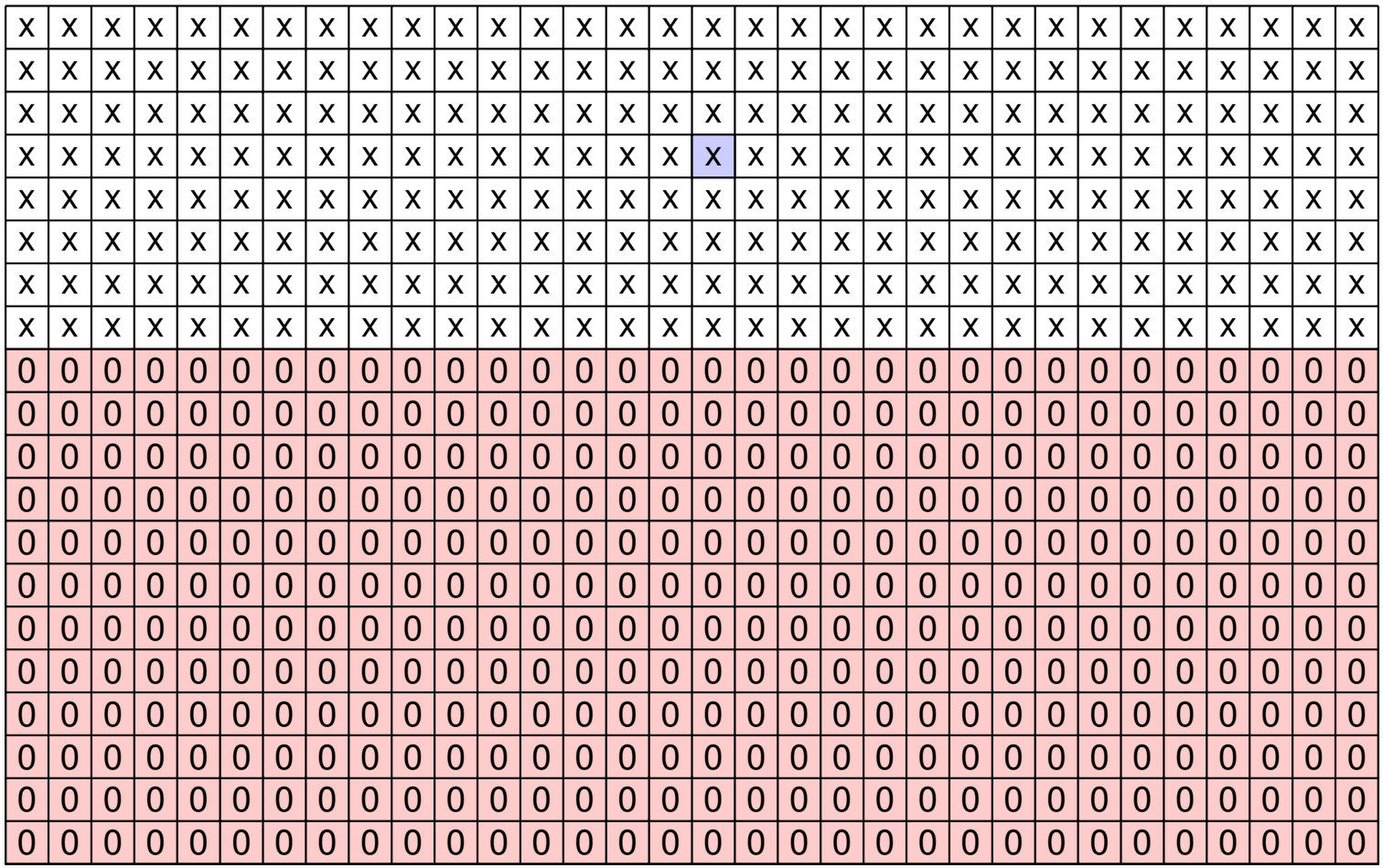
- A sofs20 disk is partitioned/structured as follows:
  - Inodes are stored in a fized-size dedicated set of blocks (`inode table`)
  - Data blocks are also stored in a fized-size dedicated set of blocks (`data block pool`)
  - A block, named `superblock`, is used for general metadata
  - List of free inodes is stored in the superblock
  - List of free data blocks is stored in the superblock and in a set of dedicated blocks (`reference table`)
  - Sequences of blocks used by inodes are stored in the inodes themselves and in data blocks allocated for that purpose

The sofs20 architecture
List of free inodes

- Based on a bit map
  - there is a one-to-one correspondence between bits in the map and inodes in the inode table, including inode 0
  - 0 $\Rightarrow$ inode is free; 1 $\Rightarrow$ inode is in-use
- The bitmap is stored in the superblock (`ibitmap` field)
  - seen as an array of 32-bit words, with fixed size
  - inode 0 is represented by bit 0 of word 0, and so on
  - unused bits are kept at 0
- freeing operation:
  - clean the inode and put the corresponding bit at 0
- allocating operation:
  - search for a bit at 0, put it at 1, and initialize the corresponding inode
  - the search must start in the position circularly next to the last allocated inode (`iidx` field)

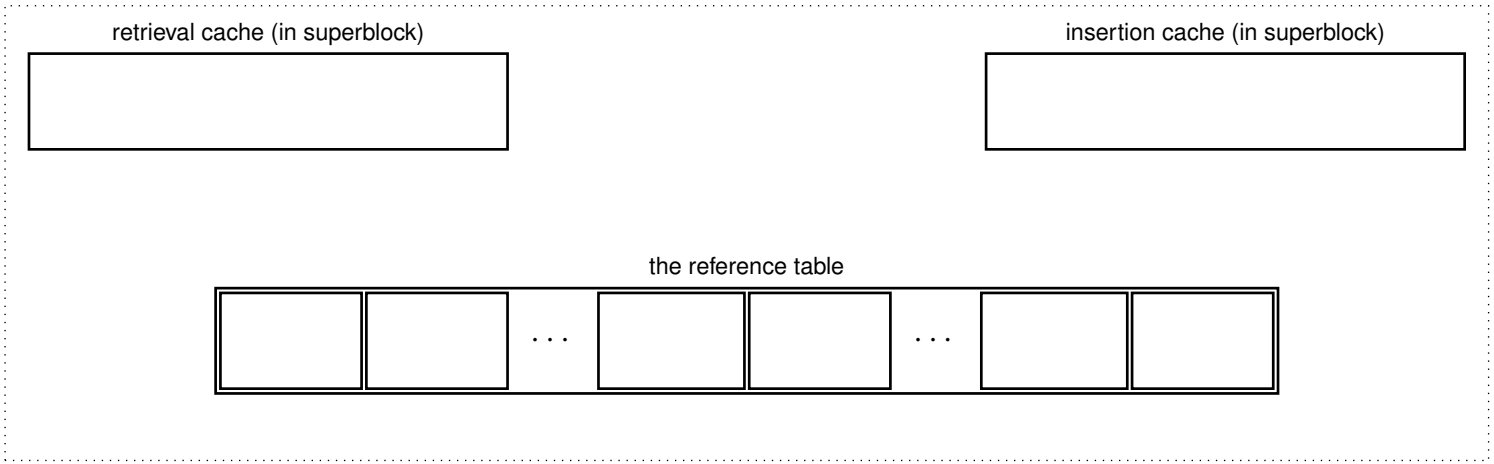
The sofs20 architecture
List of free inodes (2)

- A possible state of the bit map



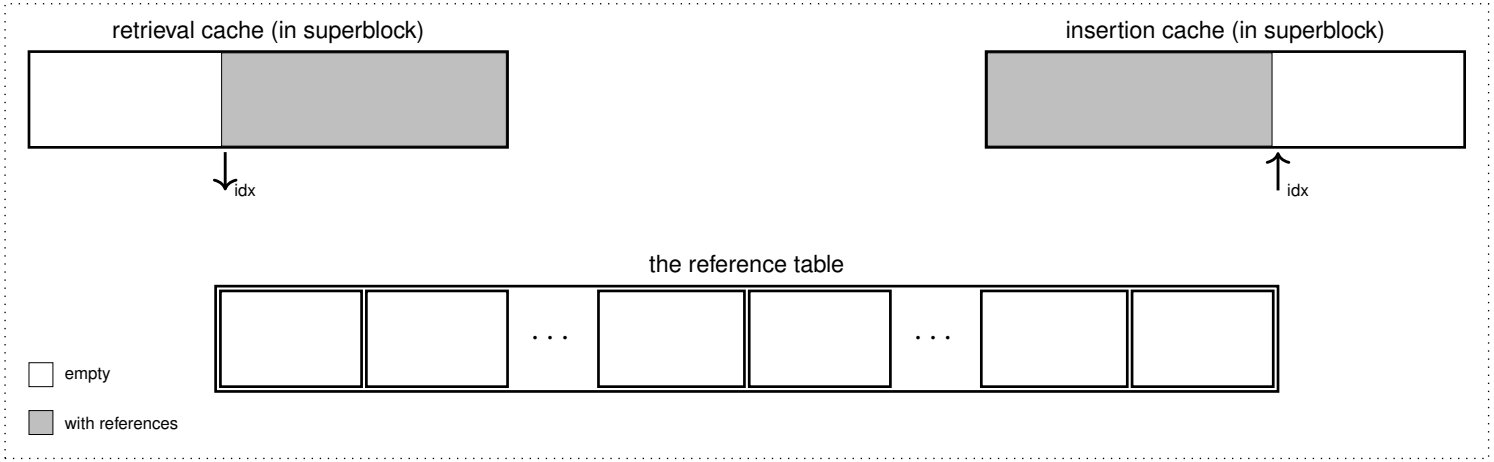
The sofs20 architecture
List of free data blocks

- Based on a FIFO
  - first free data block to be used is the oldest one in the list
  - a reference is a 32-bit word, being 0xFFFFFFFF the null reference
- The list is stored in the superblock and dedicated blocks
  - the first ordered sub-sequence is stored in the retrieval cache, representing the oldest references in the list
  - the last ordered sub-sequence is stored in the insertion cache, representing the most recent references in the list
  - the remaining ordered references are stored in the reference table



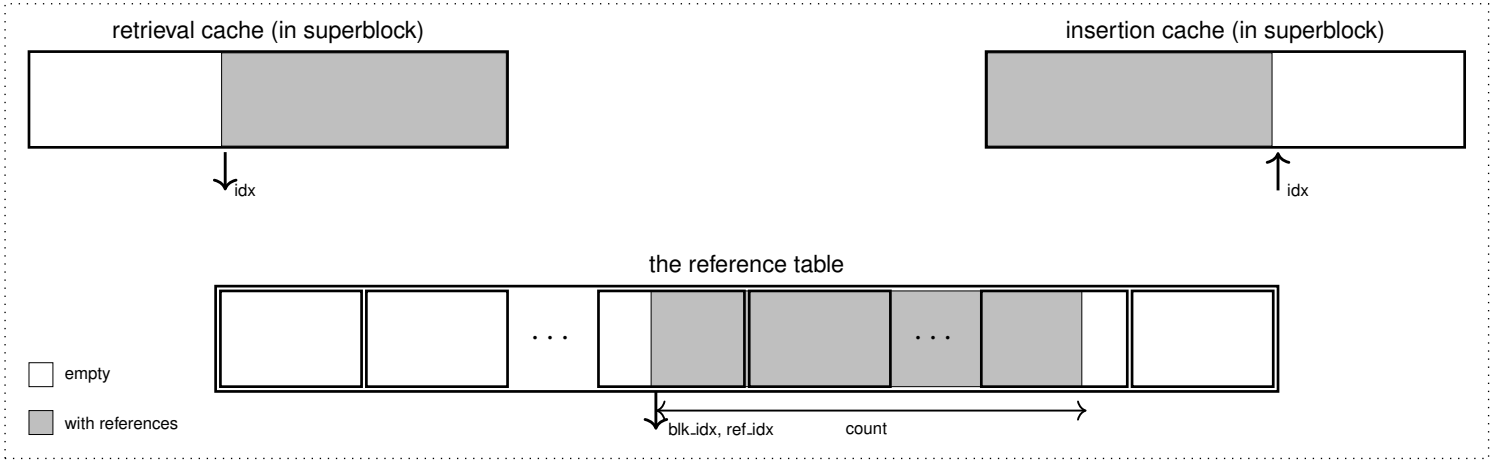
The sofs20 architecture
List of free data blocks

- retrieval cache
  - this cache may be partially empty, meaning that some references (necessarily at the begining) were already retrieved
  - an index (idx in the figure) points to the first cell with a reference
- insertion cache
  - this cache may be partially filled, meaning that some references (at the begining) were already inserted
  - an index (idx in the figure) points to the first empty cell



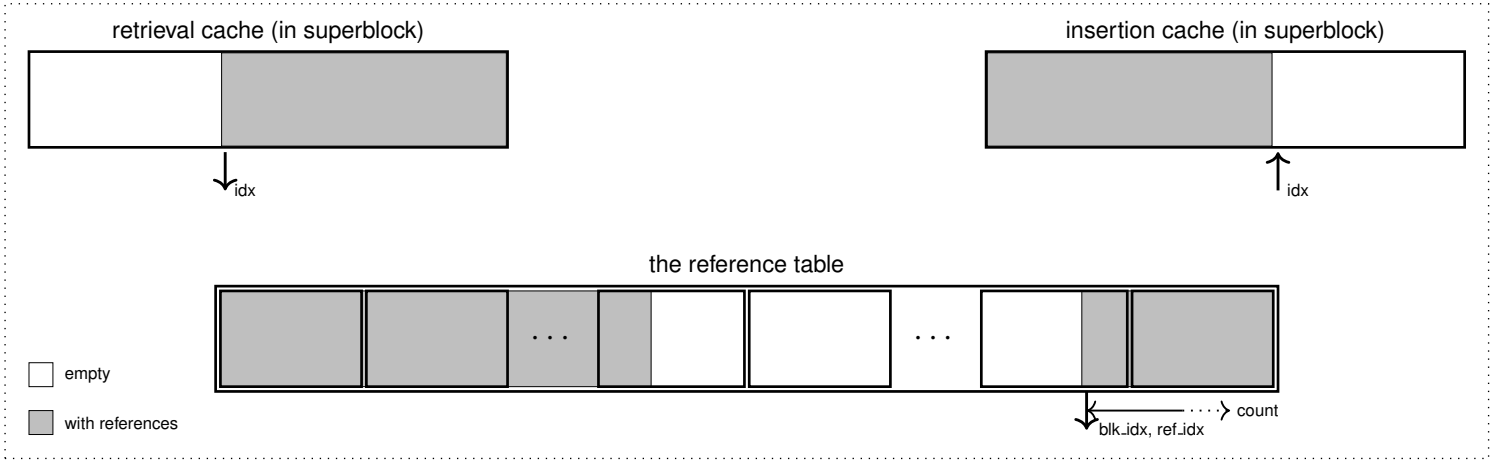
The sofs20 architecture
List of free data blocks

- reference table
  - Two fields in the superblock (`rt_start` and `rt_size`) delimit the region of the disk with the reference table
  - In general, only part of this table will be filled, as illustrated in the figure (gray area)
  - Another field (`reftable`), composed of three subfields (`blk_idx`, `ref_idx`, and `count`), stores the state of the reference table



The sofs20 architecture
List of free data blocks

- reference table
  - Two fields in the superblock (`rt_start` and `rt_size`) delimit the region of the disk with the reference table
  - In general, only part of this table will be filled, as illustrated in the figure (gray area)
  - Another field (`reftable`), composed of three subfields (`blk_idx`, `ref_idx`, and `count`), stores the state of the reference table
  - It is managed in a circular way, meaning the position after the last one is index 0.
    - thus, the occupied region can resemble that of the next figure (gray area).



The sofs20 architecture
Sequence of blocks of a file (1)

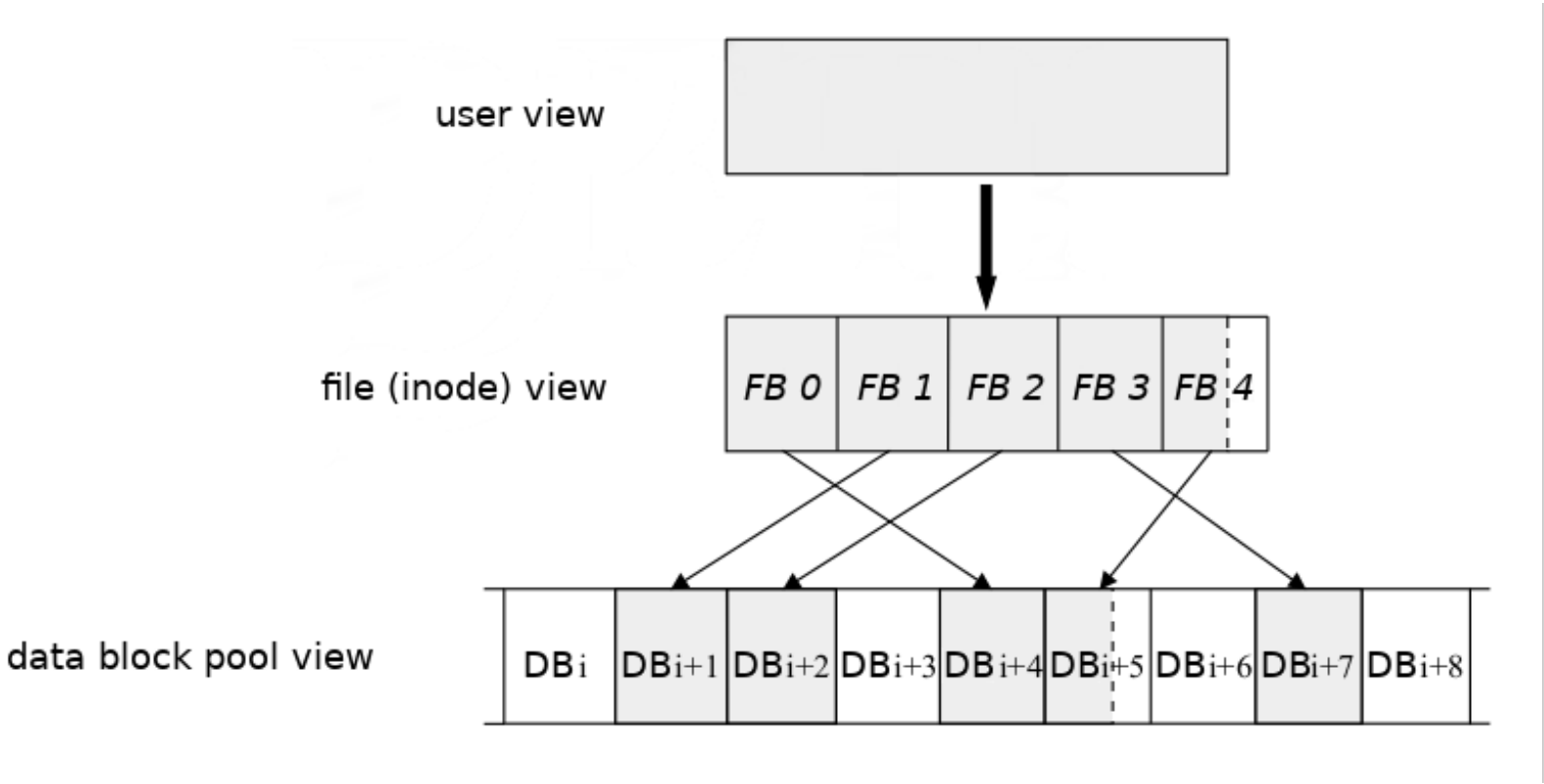
- Blocks are not shareable among files
  - an in-use block belongs to a single file
- The number of blocks required by a file to store its information is given by

$$N_b = \text{roundup}\left(\frac{\text{size}}{\texttt{BlockSize}}\right)$$

- N_b can be very big
  - if block size is 1024 bytes, a 2 GByte file needs 2 MBlocks
- N_b can be very small
  - a 0 bytes file needs no blocks for data
- It is impractical that all the blocks used by a file are contiguous in disk
- The access to the file data is in general not sequencial, but instead random
- Thus a flexible data structure, both in size and location, is required

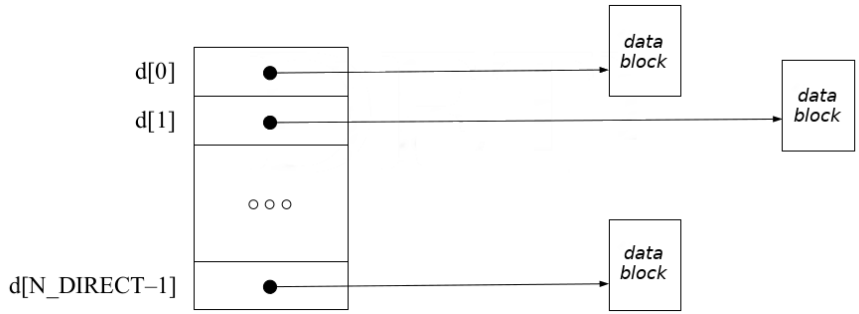
The sofs20 architecture
Sequence of blocks of a file (2)

- The programmer views a file as a continuum of bytes
- The inode views a sequence of blocks (`file block`)
- The data blocks are, in general, scattered along the data block pool

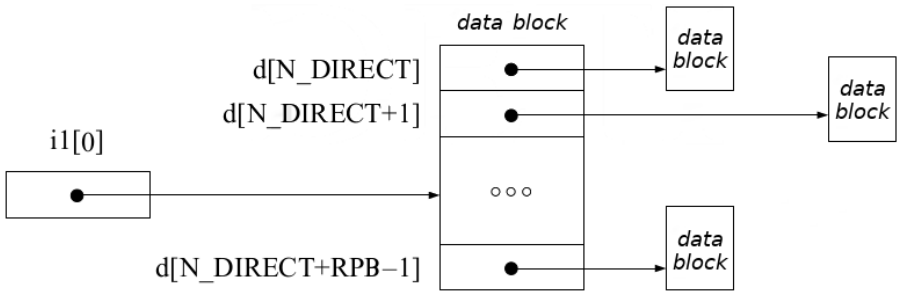


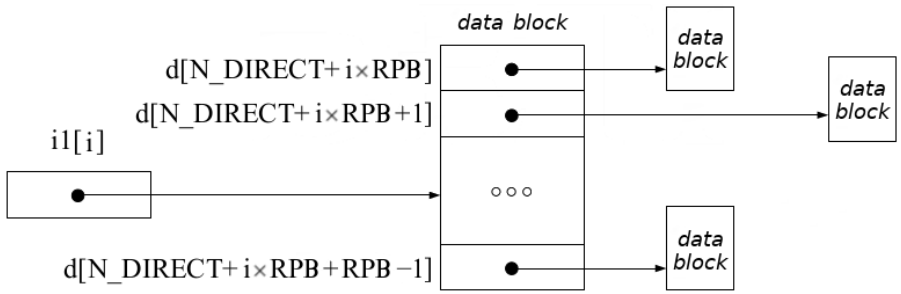
The sofs20 architecture
Sequence of blocks of a file (3)

- How is the sequence of (references to) blocks stored?
- The first references are directly stored in the inode



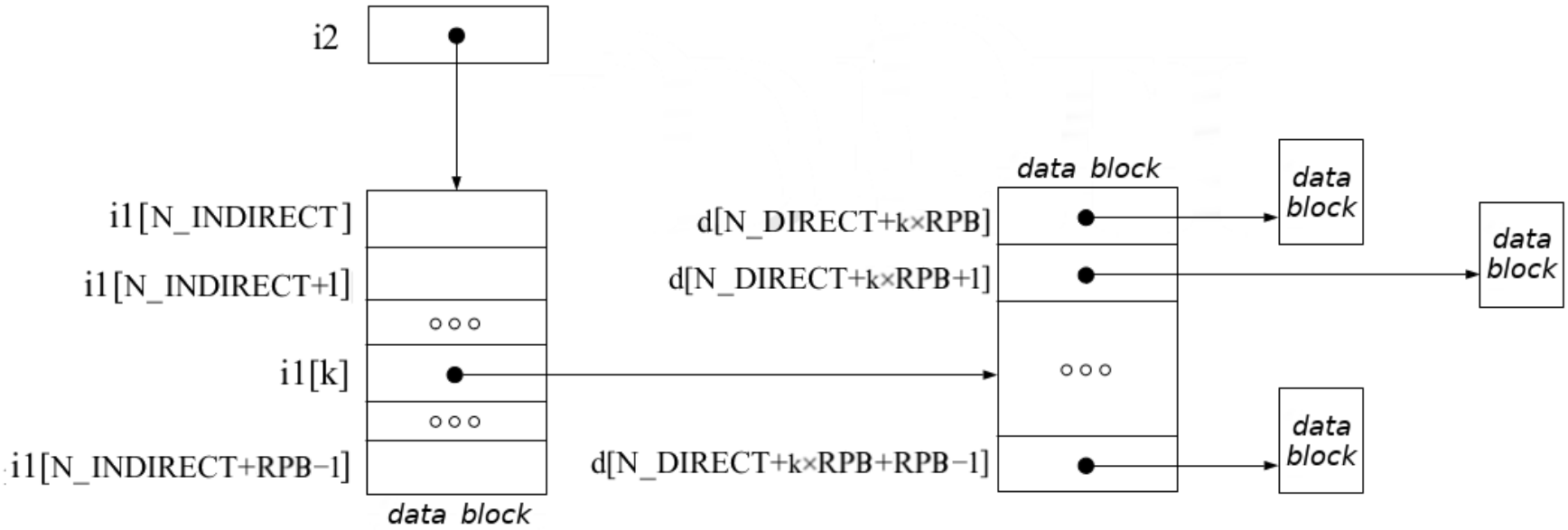
- Then, inode field `i1[.]` points to data blocks with references





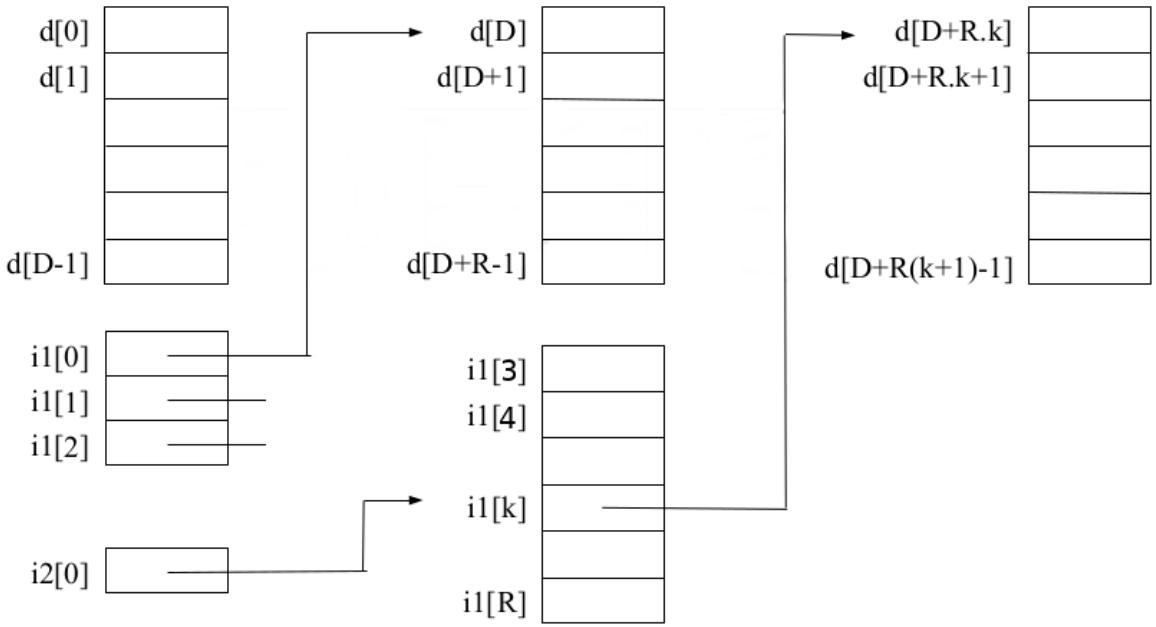
The sofs20 architecture
Sequence of blocks of a file (4)

- Finally, inode field `i2` point to a data block that extends `i1`



The sofs20 architecture
Sequence of blocks of a file (5)

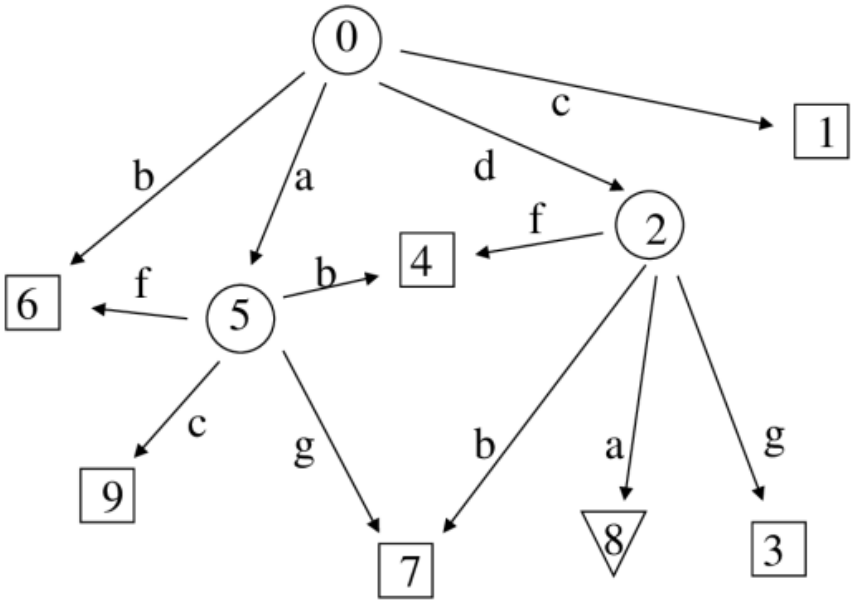
- Puting all together



- A file can contain “holes”
  - corresponding to null references covered by the size
  - and representing blocks of zeros

The sofs20 architecture
Directories and directory entries

- A directory is just a list of directory entries
- A directory entry is a pair that associates a name to an inode

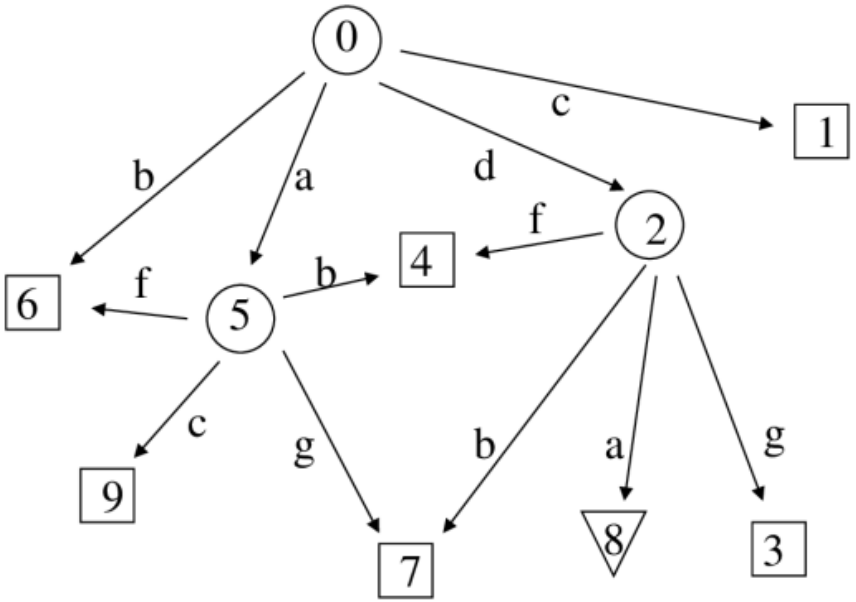


- The contents of directory "/" (inode 0) is:

name	inode
.	0
..	0
c	1
d	2
a	5
b	6

The sofs20 architecture
Directories and directory entries

- A directory is just a list of directory entries
- A directory entry is a pair that associates a name to an inode

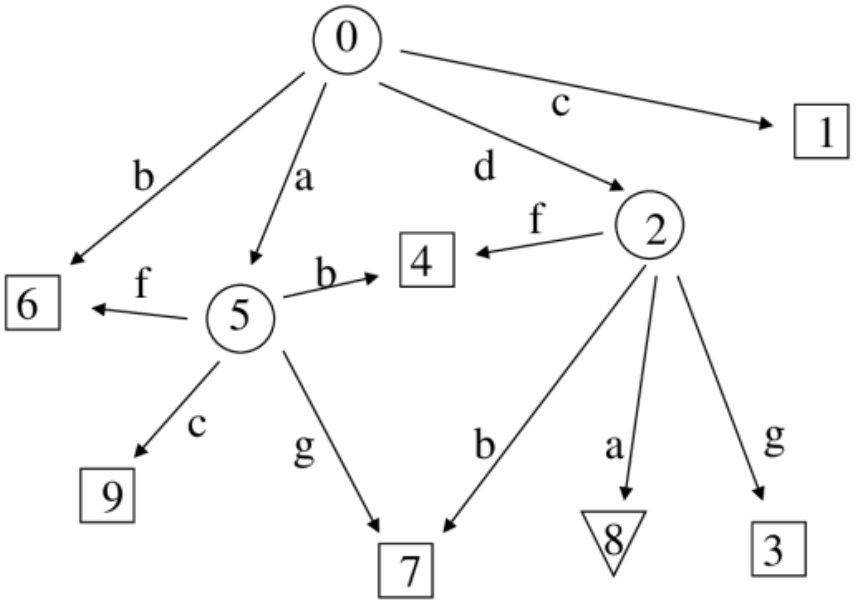


- The contents of directory "/a/" (inode 5) is:

name	inode
.	5
..	0
c	9
b	4
f	6
g	7

The sofs20 architecture
Directories and directory entries

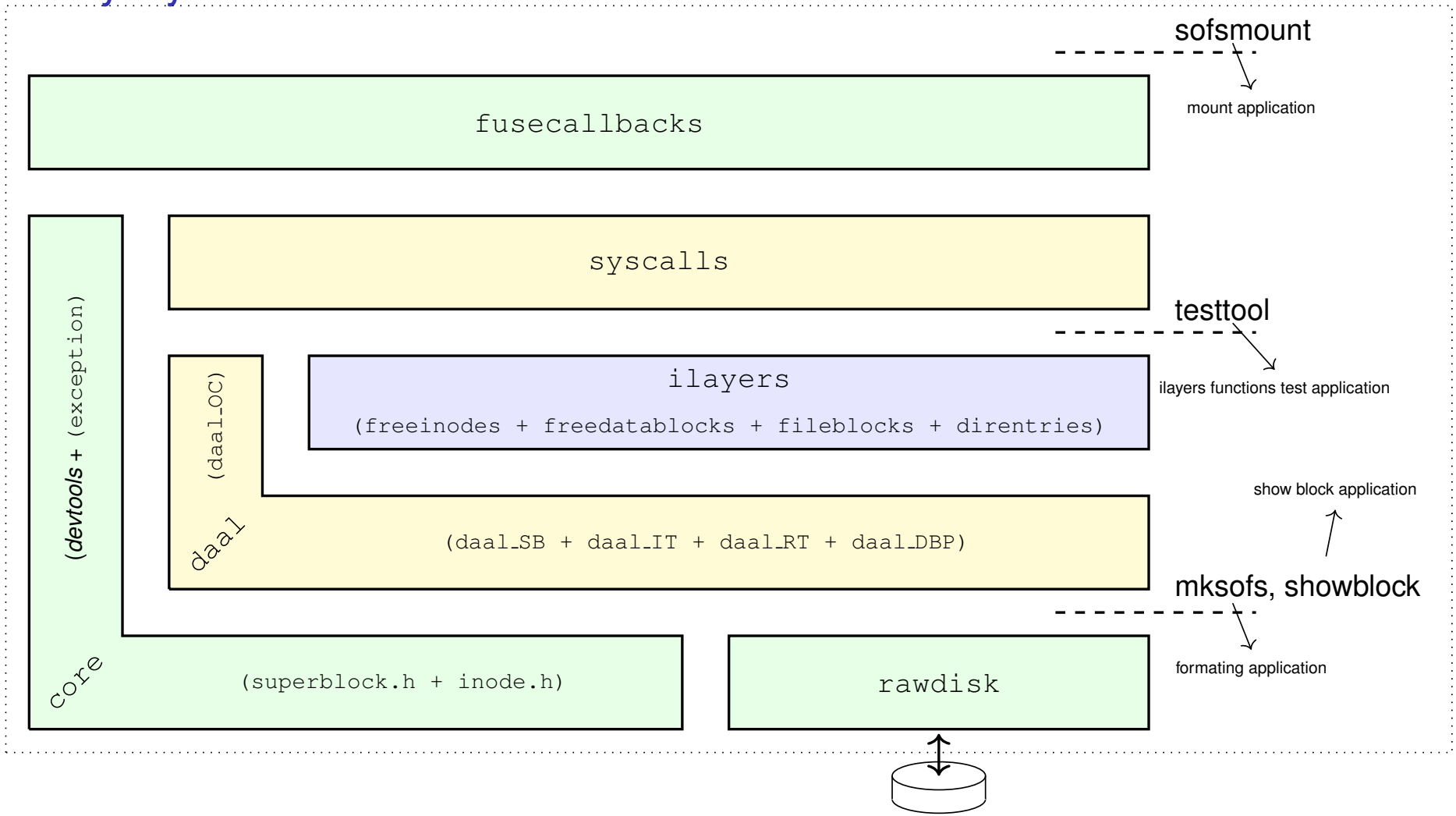
- A directory is just a list of directory entries
- A directory entry is a pair that associates a name to an inode



- The contents of directory "/d/" (inode 2) is:

name	inode
.	2
..	0
a	8
g	3
f	4
b	7

The sofs20 code structure
Library layers and tools



The sofs20 code structure
Tools

- Code prepared to use the building tool cmake
  - Need to prepare cmake
  - Can choose between make and ninja

- Code prepared to use the documentation tool doxygen
  - Configured to use only `.h` files
  - Configured to generate only html pages

- sofs20 tools:
  - showblock – show one or more blocks of a sofs20 disk
  - testtool – call functions of the intermediate layers

The formating tool

`mksofs`

- Purpose:
  - Fill in the blocks of a raw disk to make it be a sofs20 file system

- State of a newly formatted disk:
  - Inode 0 is used by the root directory
  - Data of the root directory is stored in data block number 0
  - A set of other rules have also to be observed
    - they are stated in the documentation

- Approach:
  - Code was decomposed in 6 auxiliary functions
  - Source of the main code is given