Garbage Collection in LogFS

Jörn Engel, Dirk Bolte & Robert Mertens

Lazybastard.org, IBM & University of Osnabrück

January 16, 2007

Flash properties

For all flashes:

- Medium split into eraseblocks (16-128 KiB)
- Eraseblocks can be written in any order
- Eraseblocks must be erased before being written
- Eraseblocks can be partially written

For some flashes:

- Eraseblock split into pages (8-2048 Bytes)
- Partial pages cannot get written
- Writes within eraseblocks must happen in order

LogFS device abstraction

For all media:

- Medium split into segments
- Segments can be written in any order
- Segments must be erased before being written
- Segments can be partially written

For all media:

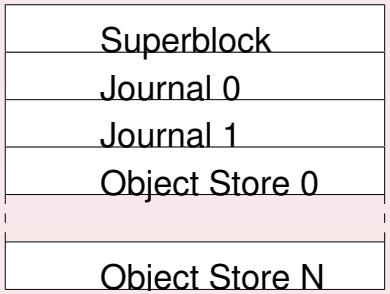
- Eraseblock split into blocks
- Partial blocks cannot get written
- Writes within segments must happen in order

Mapping flash to LogFS

- One or more eraseblocks form a segment
- One or more flash pages form a block

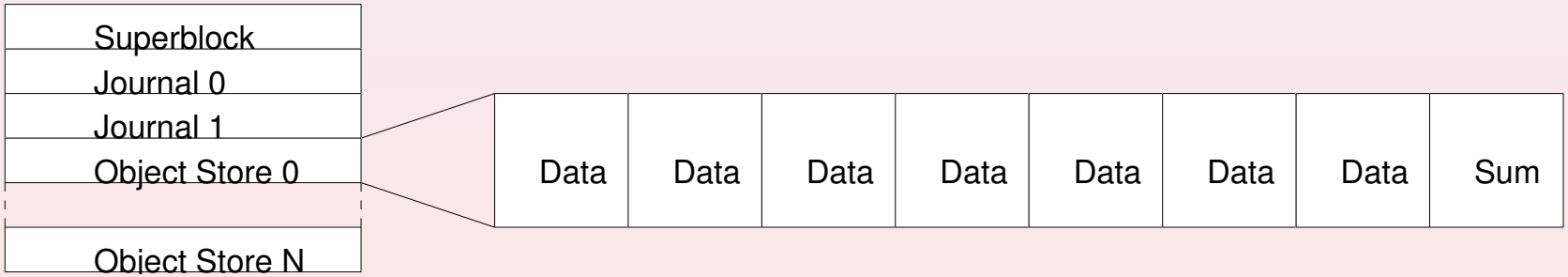
Three storage areas

- Superblock (1 Segment)
- Journal (2-8 Segments)
- Object store



Object store

- Object store split into segments
- Segments split into blocks
- Last block of each segment contains summary



Segment summary

Summary contains for each block:

- Inode number (ino)
- Logical position in a file (pos)
- Physical offset of the block (ofs)

Plus some segment-global information:

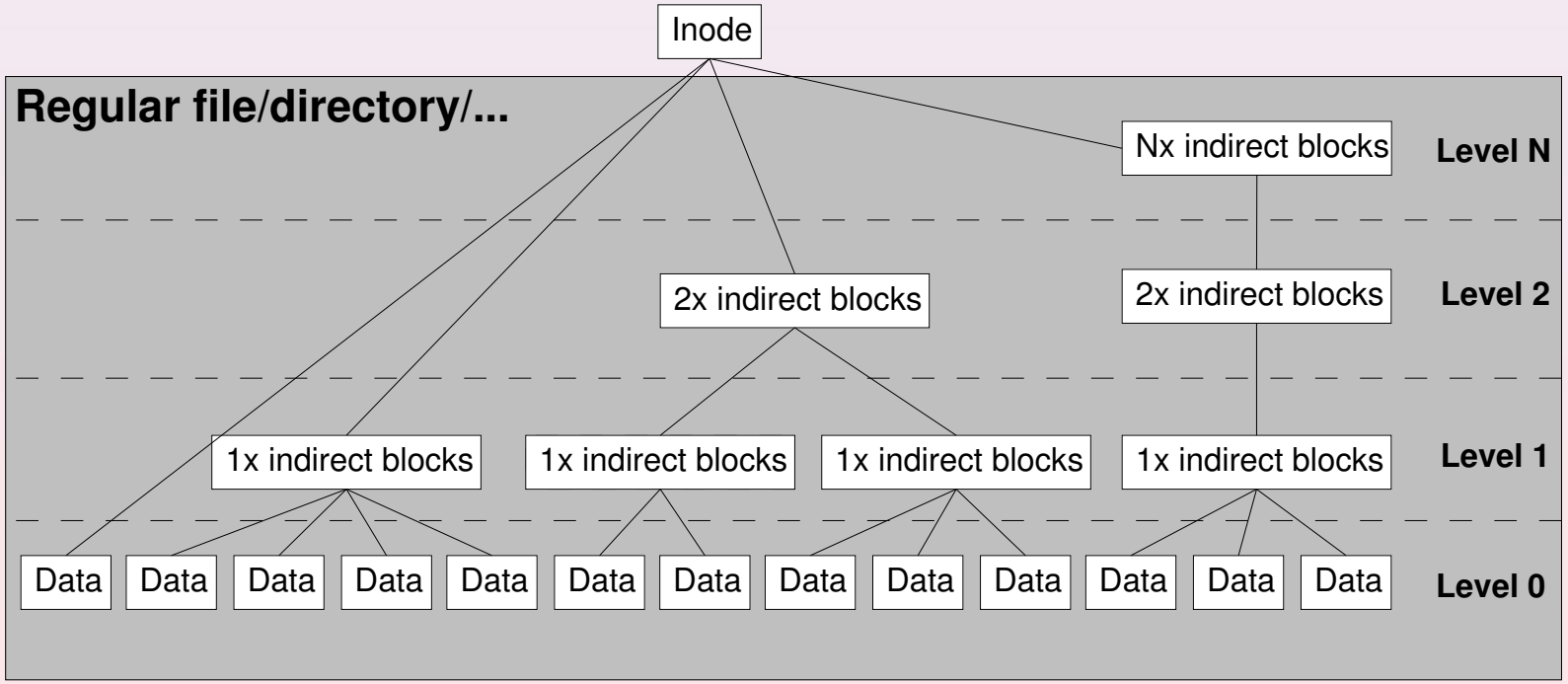
- Erase count
- Write time
- ...

Block verification

- No bitfields to track block usage
- A block is used iff an inode points to it
- Summary contains (ino,pos) pair as a back pointer

File format

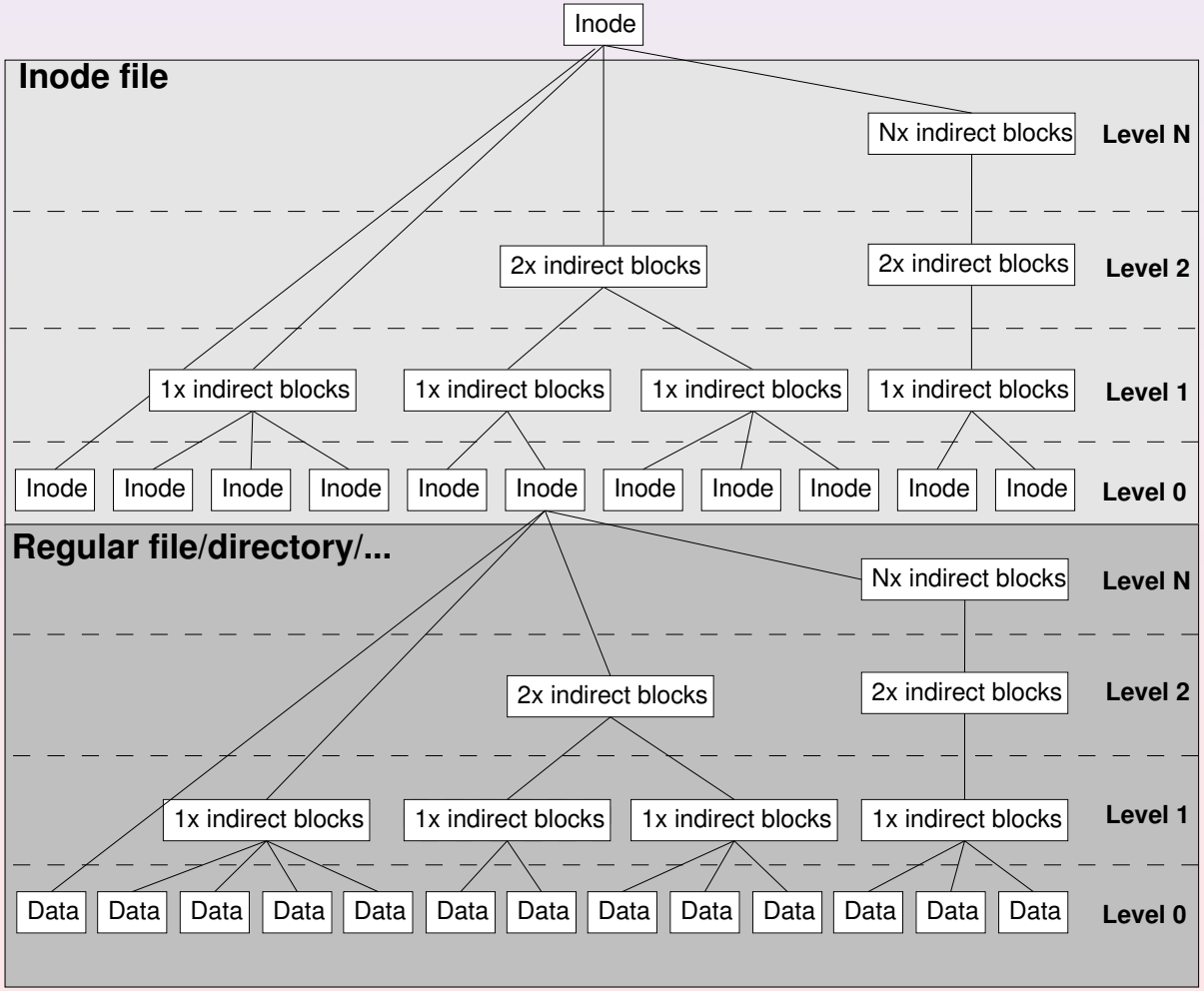
Regular Unix format with inode, indirect blocks and data blocks:



Inode file

- No reserved areas for inodes
- Inodes stored in inode file (ifile)
- Ifile's inode stored in journal

Inode file

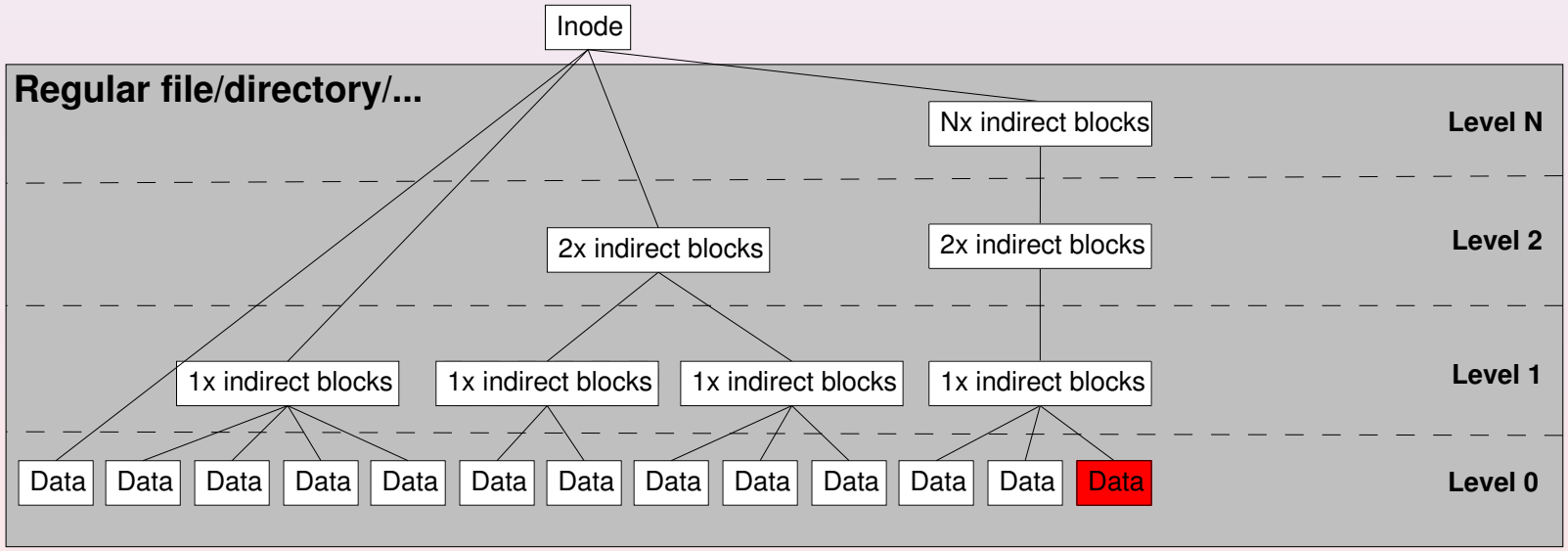

Inode
Inode file
Nx indirect blocks
Level N
2x indirect blocks
2x indirect blocks
Level 2
1x indirect blocks
1x indirect blocks
1x indirect blocks
1x indirect blocks
Level 1
Inode
Level 0
Regular file/directory/...
Nx indirect blocks
Level N
2x indirect blocks
2x indirect blocks
Level 2
1x indirect blocks
1x indirect blocks
1x indirect blocks
1x indirect blocks
Level 1
Data
Level 0

Writes

- Flash must be erased before being written
- Eraseblocks too large to be practical
- Writes cannot happen in-place
- Solution: Wandering tree

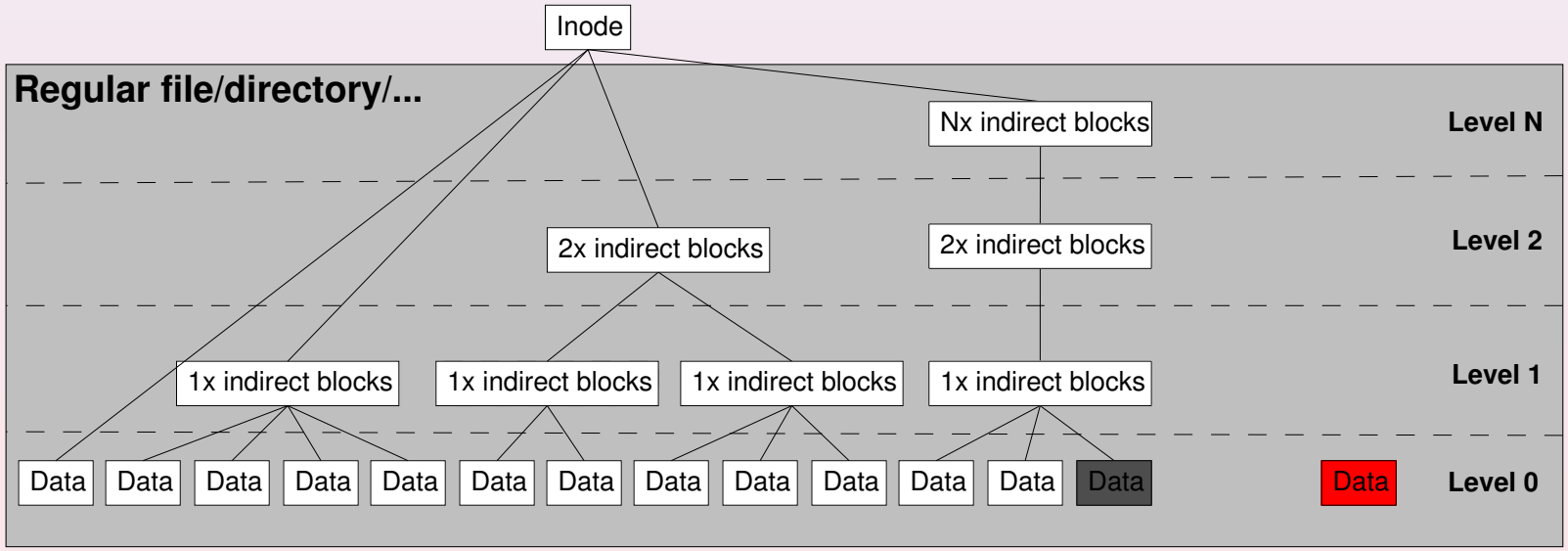
Writes

The red block is overwritten with new data ...



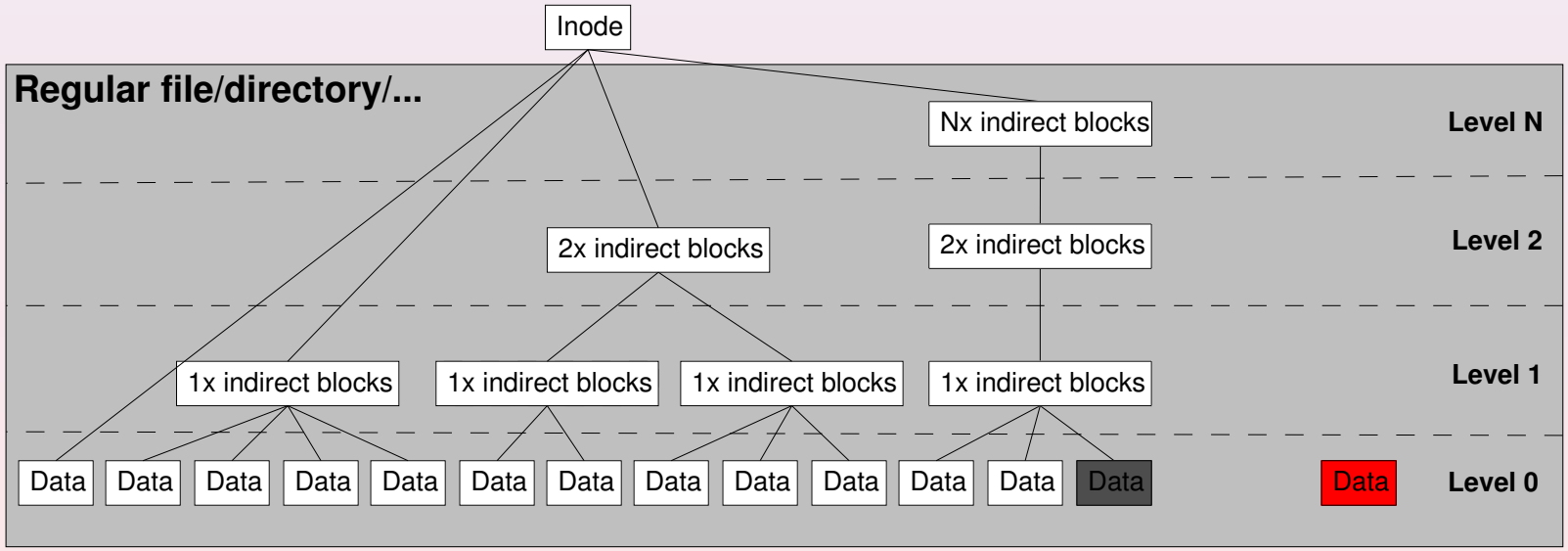
Writes

... and has to be written to a different location.



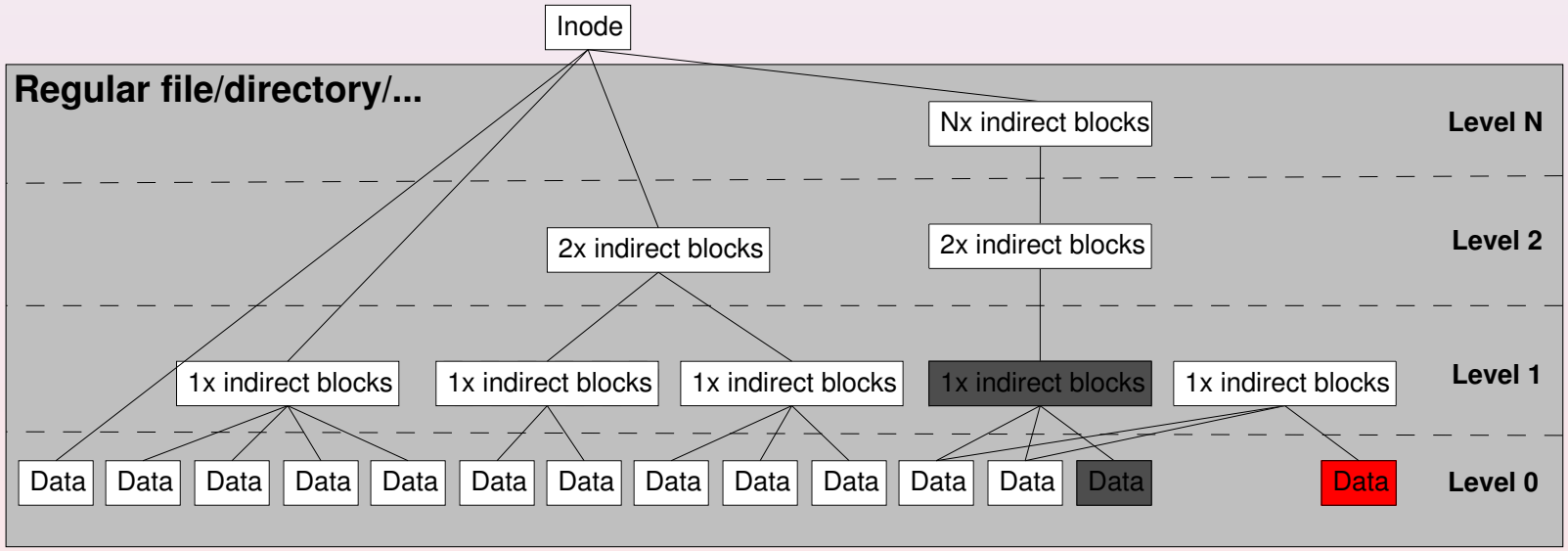
Writes

Since the indirect block still points to the old location ...



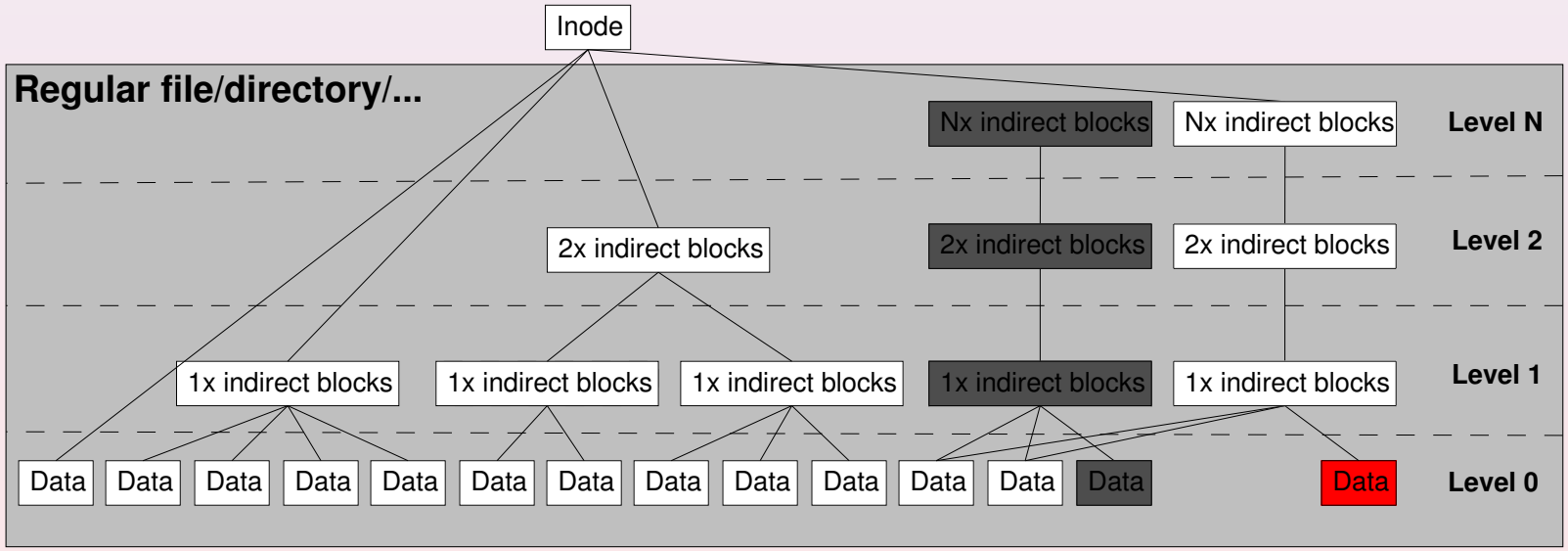
Writes

... it has to be changed as well ...



Writes

... along with all other indirect blocks further up.



Why Garbage Collection?

- Writes happen out-of-place
- Blocks obsoleted by writes
- Near-empty segments are useless

Why Garbage Collection?

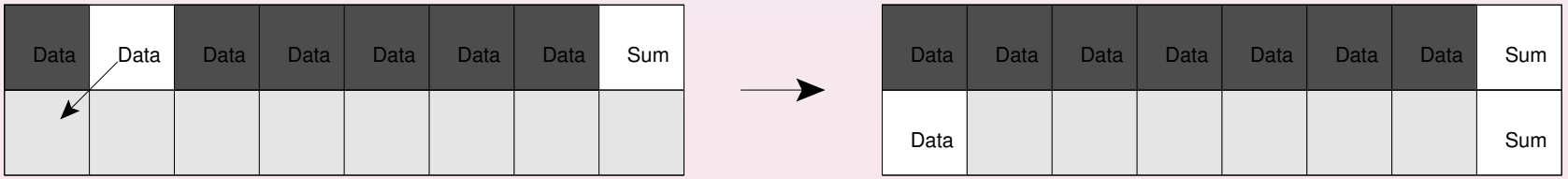
- A single valid block in a segment prevents reuse

Simple Garbage Collection

- GC has to move each valid block away ...

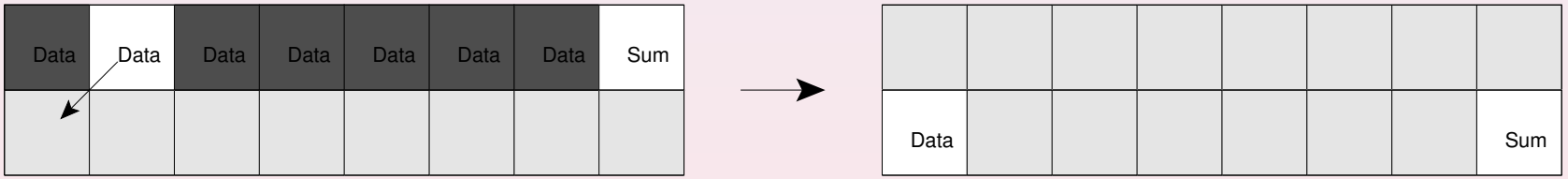
Simple Garbage Collection

- ... thereby obsoleting the old block ...



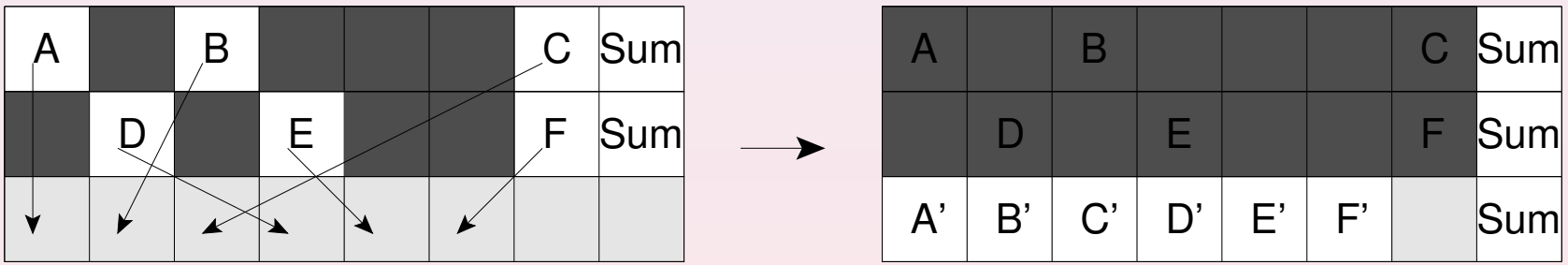
Simple Garbage Collection

- ... so it can be deleted.



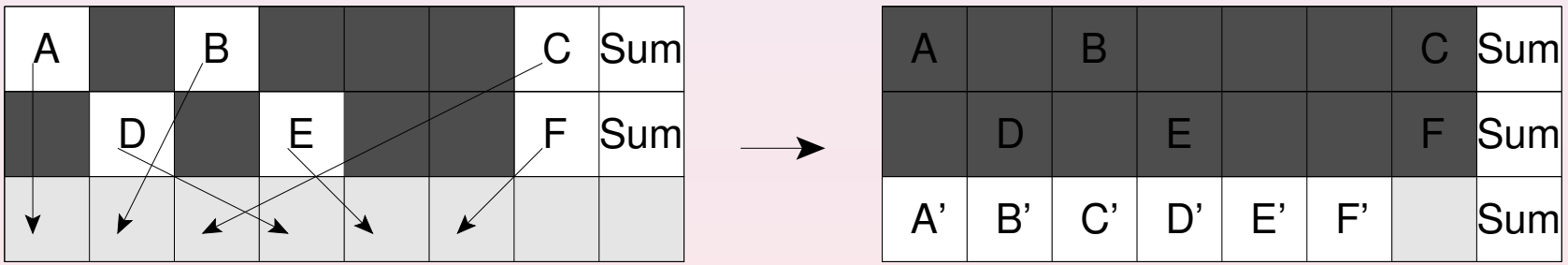
Simple Garbage Collection

- GC should free more segments than it uses.



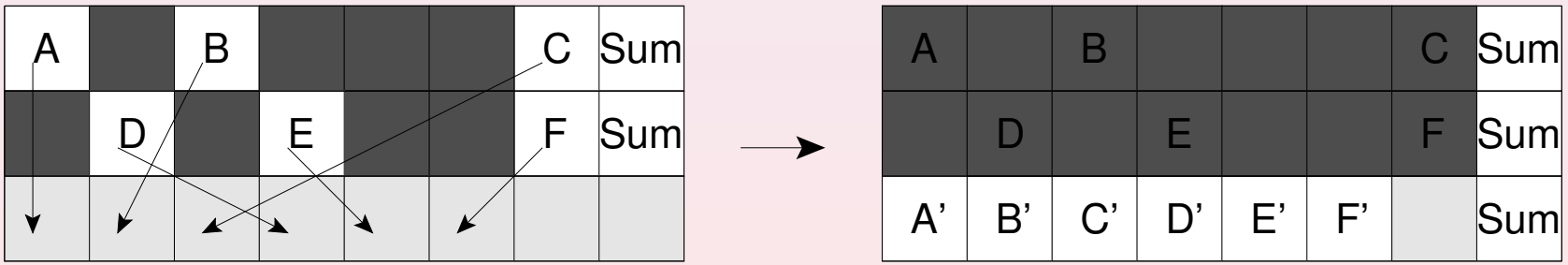
Simple Garbage Collection

- GC *must* free more segments than it uses!



Simple Garbage Collection

- GC *MUST* free more segments than it uses!!!
- GC *MUST* free more segments than it uses!!!
- GC *MUST* free more segments than it uses!!!

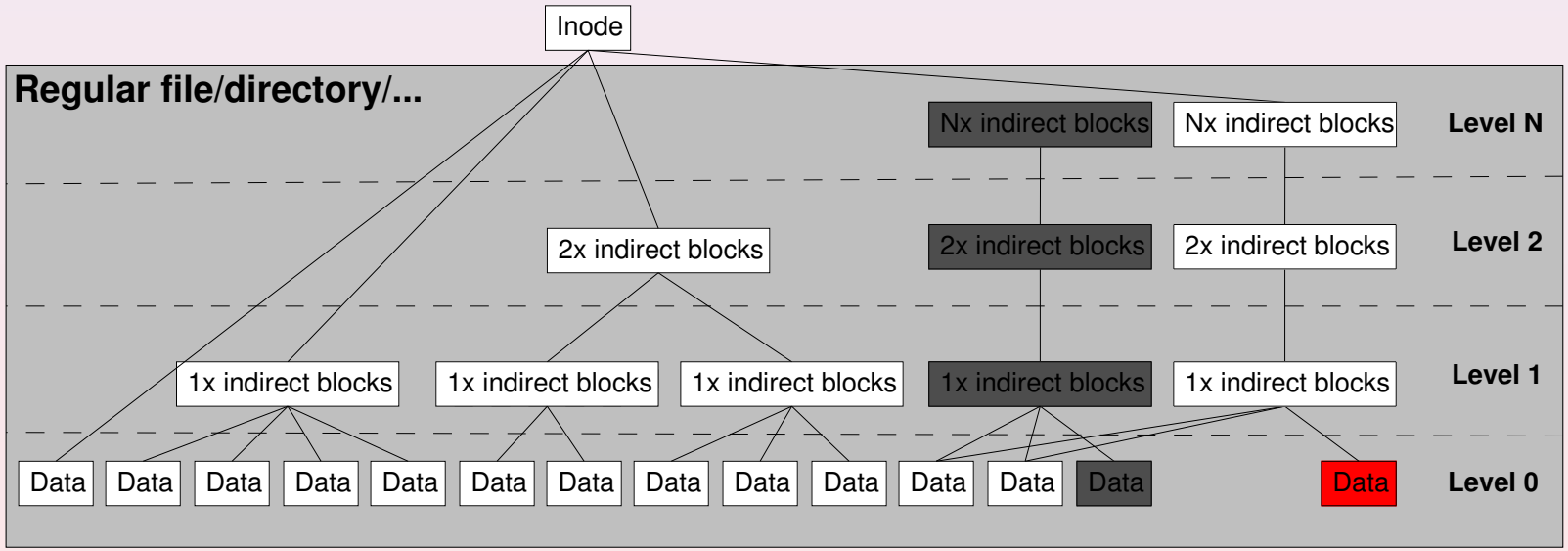


Garbage Collection with a Wandering Tree

Wandering tree?

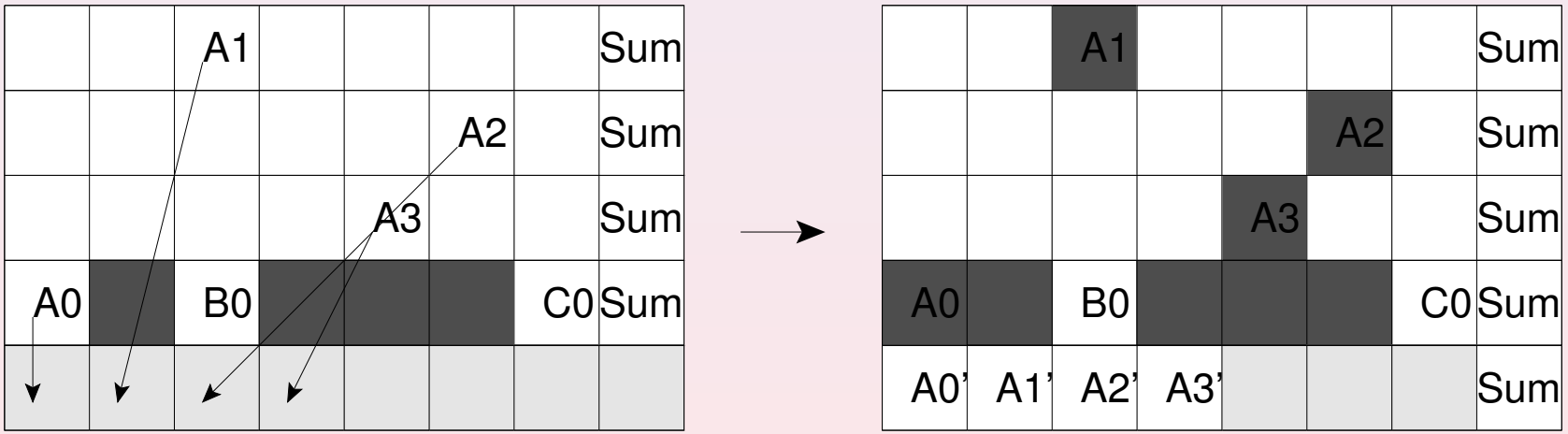
Garbage Collection with a Wandering Tree

- Didn't we have a wandering tree?



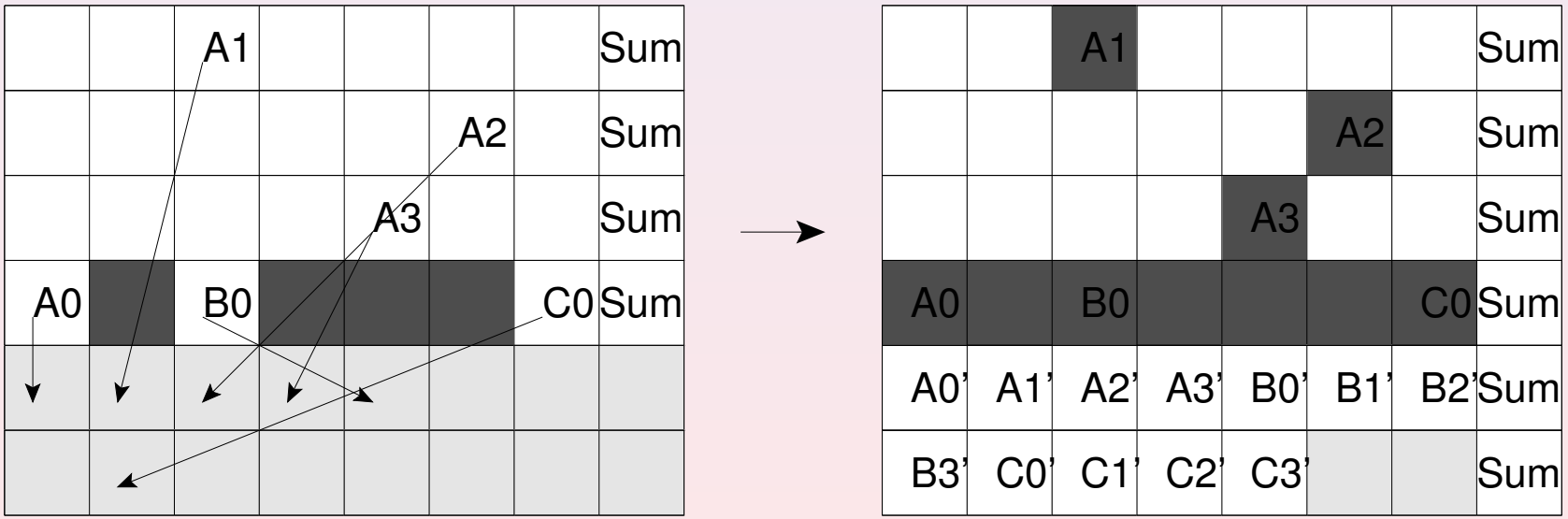
Garbage Collection with a Wandering Tree

- Oh dear!



Garbage Collection with a Wandering Tree

- Mustn't GC free more segments than it uses?

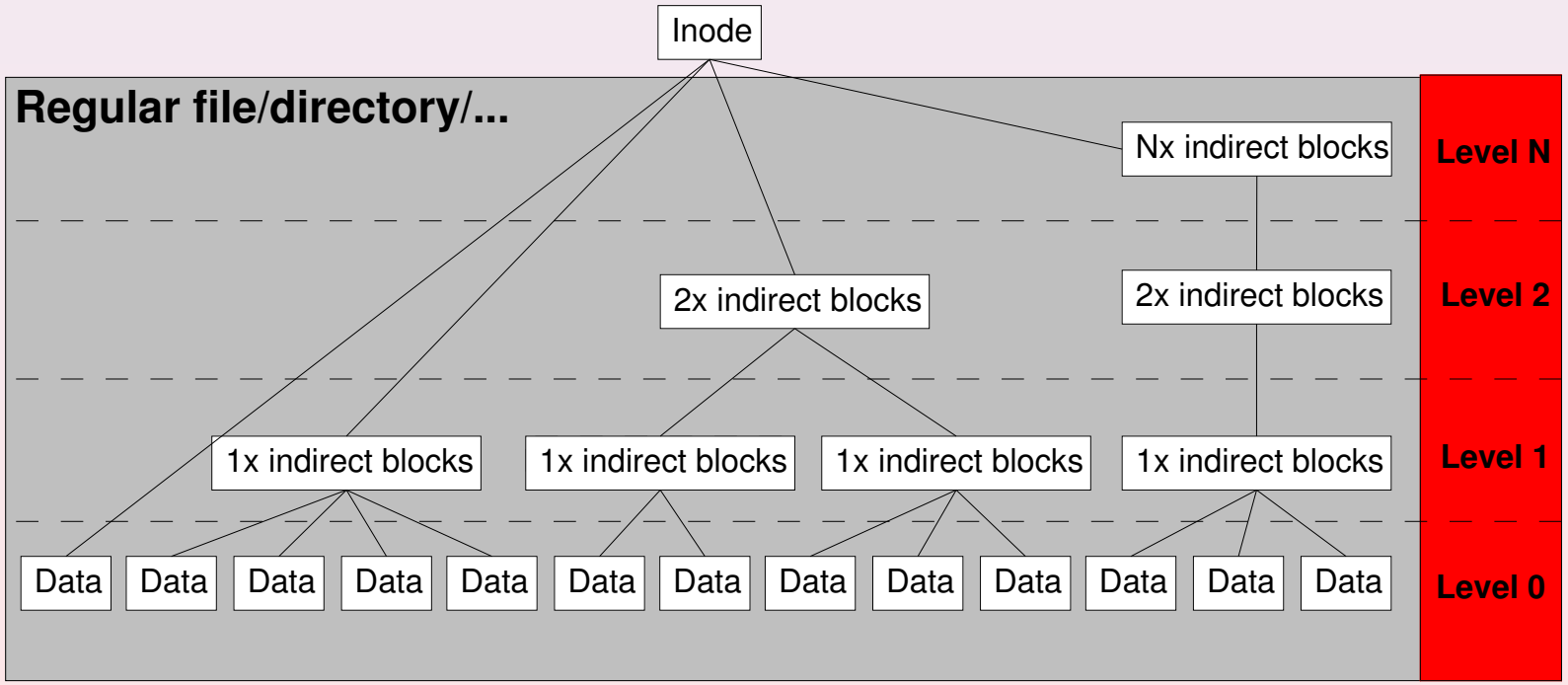


Garbage Collection with a Wandering Tree

Ooooh!

Introducing Levels

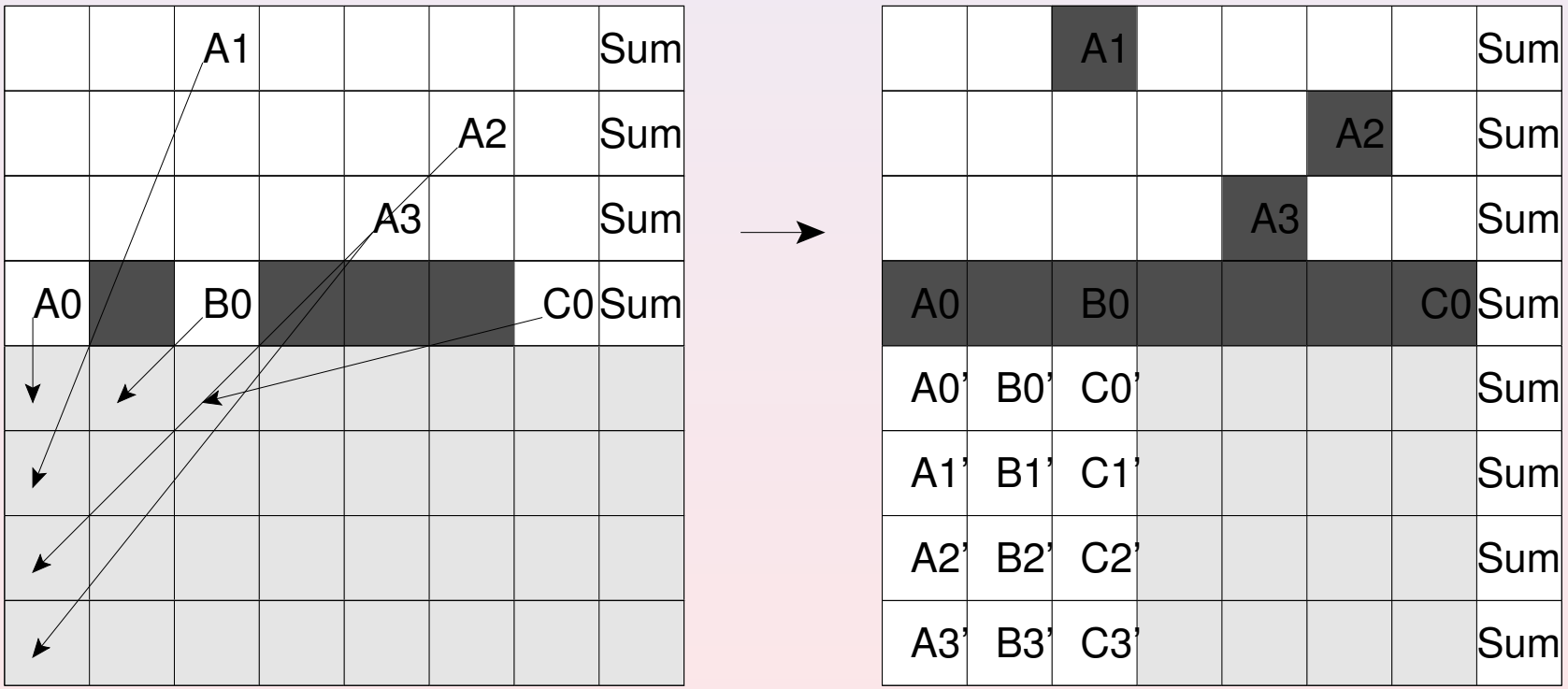
Definition: Two nodes are on the same level, if their distance from the root node is identical



Garbage Collection with Levels

- Blocks on the same level are written to the same segments
- Blocks on different levels are written to different segments

Garbage Collection with Levels



Garbage Collection with Levels

Aaaah!

Thanks!

Arnd Bergmann
Martin Schwidefsky
Dirk Bolte
Robert Mertens
You, the audience

Wrapup

Questions?
Suggestions?
Sponsors?