Structuring data for efficient I/O

Sébastien Ponce
sebastien.ponce@cern.ch

CERN

Thematic CERN School of Computing 2015

Overall Course Structure

Structuring Data for efficient I/O

- Data formats, data compression
- Data addressing

Many ways to Store Data

- Storage devices and their specificities
- Distributing and parallelizing storage

Preserving data

- Data consistency
- Data safety

Key ingredients to achieve efficient I/O

- Synchronous vs asynchronous I/O
- I/O optimizations and caching

Outline

Data format

What is the best data format ?

What is the best data format ?

Obviously depends on your use case...

What is the best data format ?

Obviously depends on your use case...

Several criteria

- Human readability
- Machine readability
  - aka parsing easiness
- Compacity
  - very close to I/O efficiency
- Computing efficiency

What is the best data format ?

Obviously depends on your use case...

Several criteria

- Human readability
- Machine readability
  - aka parsing easiness
- Compacity
  - very close to I/O efficiency
- Computing efficiency

Guess what ? You won't get everything !

Main handles to play with

The syntax level

- which format do you use ?
- will influence human and machine readibility
- will define compacity / I/O efficiency

The structure

- how do you group / hierarchize your data ?
- will mainly influence computing efficiency

Human readable formats

XML

```
<Detector name="Atlas">
  <SubDetector name="InnerTracker">
    <Tile name="ABC" x=123 y=456 z=789/>
    <Tile name="DEF" x=987 y=345 z=111/>
  </SubDetector>
</Detector>
```

Human readable formats

JSON

```
{name : atlas,
 subdetectors : [
   {name : InnerTracker,
    tiles: [
     {name : ABC, x : 123, y : 456, z : 789},
     {name : DEF, x : 987, y : 345, z : 111}]}]}
```

Human readable formats - Usage

Pros

- nice for small amounts of data, read by humans for real
  - Typically config files
- very easy to parse
  - many standard parsers available for all languages

Contras

- expensive to parse
- extremely verbose (bad compacity)

Actual data usually use binary formats

Binary formats

Root

```
00000000  72 6f 6f 74 00 00 d0 b2  00 00 00 64 00 00 01 3a  |root.......d...:|
00000010  00 00 01 05 00 00 00 35  00 00 00 01 00 00 00 36  |.......5.......6|
00000020  04 00 00 00 01 00 00 00  00 00 00 00 00 00 01 5a  |...............Z|
00000030  c6 b3 34 ec 47 11 e4 97  17 aa 84 8e 80 be ef 00  |..4.G...........|
00000040  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00000050  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00000060  00 00 00 00 00 00 00 72  00 04 00 00 00 47 51 35  |.......r.....GQ5|
00000070  51 a1 00 2b 00 01 00 00  00 64 00 00 00 00 05 54  |Q..+.....d.....T|
00000080  46 69 6c 65 09 66 69 6c  65 2e 72 6f 6f 74 00 09  |File.file.root..|
00000090  66 69 6c 65 2e 72 6f 6f  74 00 00 05 51 35 51 a1  |file.root...Q5Q.|
000000a0  51 35 51 b4 00 00 00 2f  00 00 00 36 00 00 00 64  |Q5Q..../...6...d|
000000b0  00 00 00 00 00 00 00 d6  00 01 5a c6 b3 34 ec 47  |..........Z..4.G|
000000c0  11 e4 97 17 aa 84 8e 80  be ef 00 00 00 00 00 00  |................|
000000d0  00 00 00 00 00 00 00 00  00 2f 00 04 00 00 00 04  |........./......|
000000e0  51 35 51 b4 00 2b 00 01  00 00 00 d6 00 00 00 64  |Q5Q..+.........d|
000000f0  05 54 46 69 6c 65 09 66  69 6c 65 2e 72 6f 6f 74  |.TFile.file.root|
00000100  00 00 00 00 00 00 00 00  35 00 04 00 00 00 0a 51  |........5......Q|
00000110  35 51 b4 00 2b 00 01 00  00 01 05 00 00 00 64 05  |5Q..+.........d.|
00000120  54 46 69 6c 65 09 66 69  6c 65 2e 72 6f 6f 74 00  |TFile.file.root.|
00000130  00 01 00 00 01 3a 77 35  94 00                    |.....:w5..|
```

Binary formats

Root

```
00000000  72 6f 6f 74 00 00 d0 b2  00 00 00 64 00 00 01 3a  |root.......d...:|
00000010  00 00 01 05 00 00 00 35  00 00 00 01 00 00 00 36  |.......5.......6|
00000020  04 00 00 00 01 00 00 00  00 00 00 00 00 00 01 5a  |...............Z|
00000030  c6 b3 34 ec 47 11 e4 97  17 aa 84 8e 80 be ef 00  |..4.G...........|
00000040  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00000050  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00000060  00 00 00 00 00 00 00 72  00 04 00 00 00 47 51 35  |.......r.....GQ5|
00000070  51 a1 00 2b 00 01 00 00  00 64 00 00 00 00 05 54  |Q..+.....d.....T|
00000080  46 69 6c 65 09 66 69 6c  65 2e 72 6f 6f 74 00 09  |File.file.root..|
00000090  66 69 6c 65 2e 72 6f 6f  74 00 00 05 51 35 51 a1  |file.root...Q5Q.|
000000a0  51 35 51 b4 00 00 00 2f  00 00 00 36 00 00 00 64  |Q5Q..../...6...d|
000000b0  00 00 00 00 00 00 00 d6  00 01 5a c6 b3 34 ec 47  |..........Z..4.G|
000000c0  11 e4 97 17 aa 84 8e 80  be ef 00 00 00 00 00 00  |................|
000000d0  00 00 00 00 00 00 00 00  00 2f 00 04 00 00 00 04  |........./......|
000000e0  51 35 51 b4 00 2b 00 01  00 00 00 d6 00 00 00 64  |Q5Q..+.........d|
000000f0  05 54 46 69 6c 65 09 66  69 6c 65 2e 72 6f 6f 74  |.TFile.file.root|
00000100  00 00 00 00 00 00 00 00  35 00 04 00 00 00 0a 51  |........5......Q|
00000110  35 51 b4 00 2b 00 01 00  00 01 05 00 00 00 64 05  |5Q..+.........d.|
00000120  54 46 69 6c 65 09 66 69  6c 65 2e 72 6f 6f 74 00  |TFile.file.root.|
00000130  00 01 00 00 01 3a 77 35  94 00                    |.....:w5..|
```

it's actually an empty file !

Root format

Header Decoded

726f6f74	'root', root file identifier
0000d0b2	version identifier
00000064	pointer to first object
0000013a	pointer to last+1 object
00000105	pointer to free data record
00000035	nb bytes free
00000001	nb bytes in name/title
00000036	nb bytes for pointers
04	compression level
...	...

Binary formats - Usage

Pros

- very compact, so I/O efficient
- seems easy to parse
  - you have to write your own parser
  - but you just use memory representation

Contras

- portability may be a nightmare
  - little indian vs big indian
  - 32 vs 64 bits architectures
- you cannot easily "see" /"use" your data directly
  - you need special tools, e.g. the ROOT framework

Data structure by example - scenario

Scenario

- You are measuring temperatures within a piece of detector
- You have 10K captors and you take one measure every minute
- After a month, you got 432M measures
- That is 1.6GB if you take single precision floats (32bits)

Data structure by example - row storage

Naive structure

- You arrange your captors in a sequential order according to the detector geometry
- Each minute, you create a new "row" of data, with 10K floats representing temperatures given by the captors, in that order

Data structure by example - row storage

Naive structure

- You arrange your captors in a sequential order according to the detector geometry
- Each minute, you create a new "row" of data, with 10K floats representing temperatures given by the captors, in that order

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0

Data structure by example - row storage

Naive structure

- You arrange your captors in a sequential order according to the detector geometry
- Each minute, you create a new "row" of data, with 10K floats representing temperatures given by the captors, in that order

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1

Data structure by example - row storage

Naive structure

- You arrange your captors in a sequential order according to the detector geometry
- Each minute, you create a new "row" of data, with 10K floats representing temperatures given by the captors, in that order

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1
...	...	...	...	...
n	a_n	b_n	...	z_n

Data structure by example - row storage

Naive structure

- You arrange your captors in a sequential order according to the detector geometry
- Each minute, you create a new "row" of data, with 10K floats representing temperatures given by the captors, in that order

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1
...	...	...	...	...
n	a_n	b_n	...	z_n

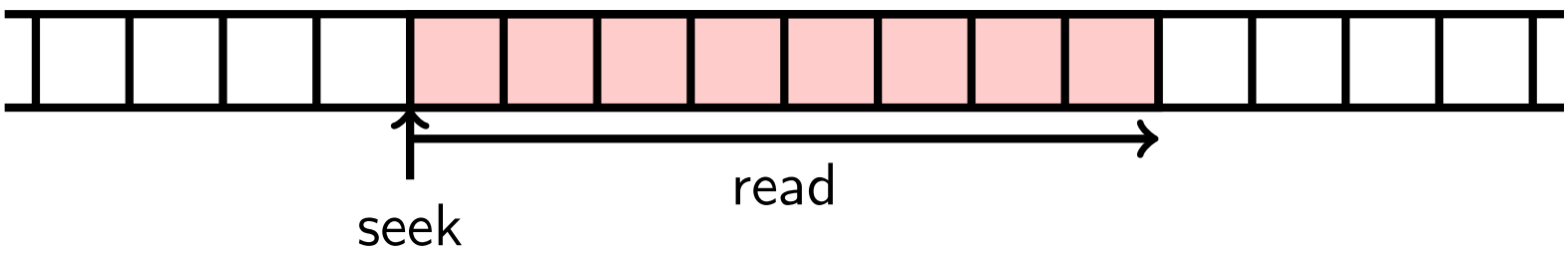
File content

a_0 b_0 ... z_0 a_1 b_1 ... z_1 ... a_n b_n ... z_n

Data structure by example - access

Find out overheated devices at a given time

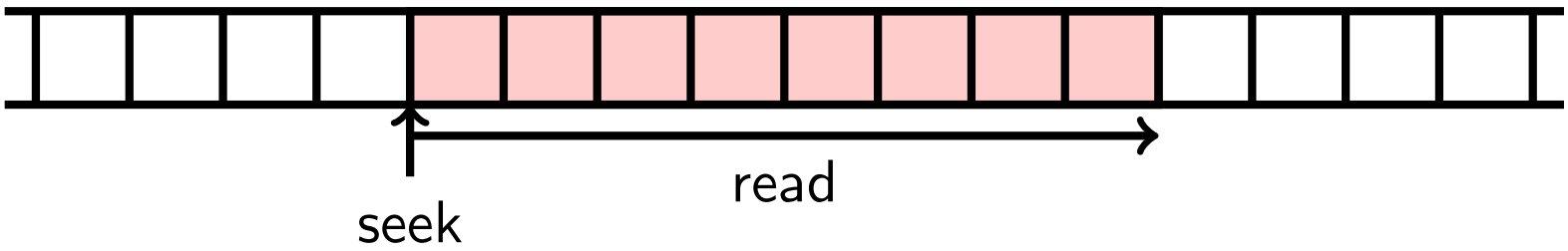
- find the offset of that time in the file
- read 10K numbers
- apply simple filter



Data structure by example - access

Find out overheated devices at a given time

- find the offset of that time in the file
- read 10K numbers
- apply simple filter



Cost

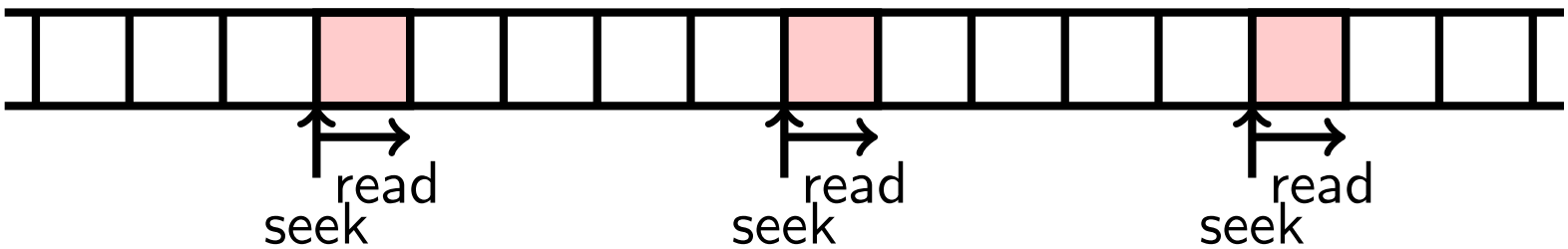
- one seek
- one read of 10K ints

This is efficient !

Data structure by example - access (2)

Graph the temperature evolution of a given device

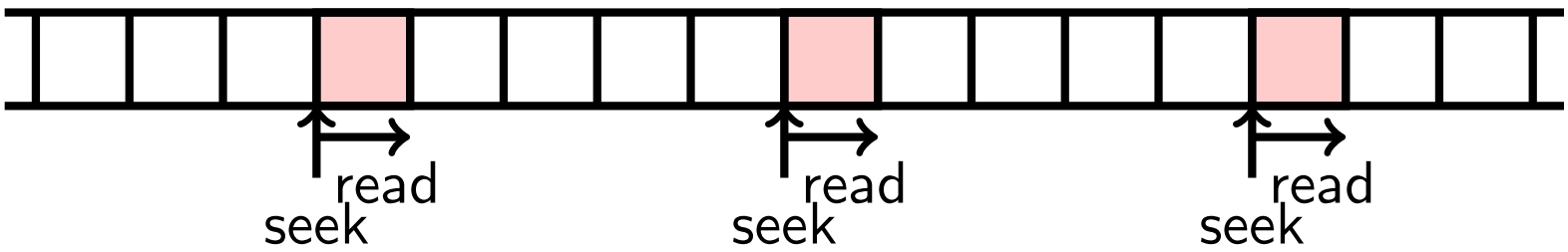
- read 43.2K numbers from the file, every 40K bytes
- graph them



Data structure by example - access (2)

Graph the temperature evolution of a given device

- read 43.2K numbers from the file, every 40K bytes
- graph them



Cost

- 43.2K reads of 4 bytes and 43.2K seeks !
- on top typical block size in a filesystem is 8k
- you will probably read effectively 20% of the file !
- actually reading the whole file will be more efficient

Here the structure of our data is a killer

Column storage

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1
...	...	...	...	...
n	a_n	b_n	...	z_n

Column storage

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1
...	...	...	...	...
n	a_n	b_n	...	z_n

File content

a_0 a_1 ... a_n b_0 b_1 ... b_n ... z_0 z_1 ... z_n

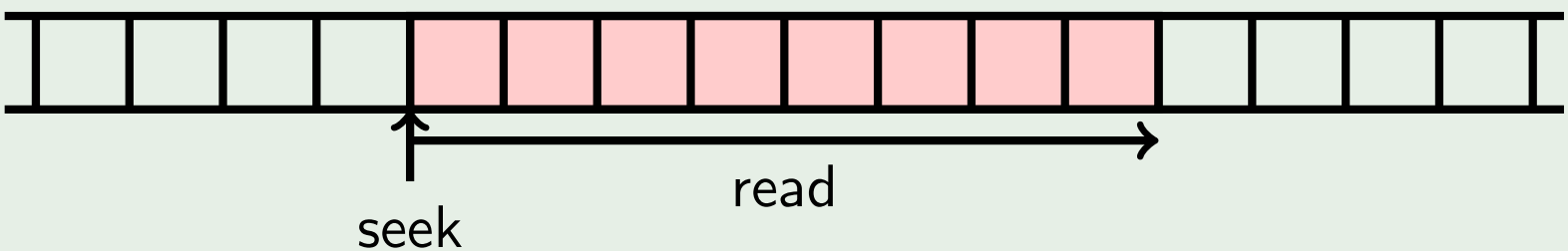
Column storage

Time (mn)	Captor 1	Captor 2	...	Captor c
0	a_0	b_0	...	z_0
1	a_1	b_1	...	z_1
...	...	...	...	...
n	a_n	b_n	...	z_n

File content

a_0 a_1 ... a_n b_0 b_1 ... b_n ... z_0 z_1 ... z_n

Back to efficient read



Row vs column storage

Definition

Row storage respects internal structure of the data and puts the different items one next in a sequence

Column storage breaks the internal structure of the data to collate similar pieces

Row vs column storage

Definition

Row storage respects internal structure of the data and puts the different items one next in a sequence

Column storage breaks the internal structure of the data to collate similar pieces

Why to use column ?

- to avoid scattered reads in the second example
- because the first example is naive. You want overheating devices for a range of time and will be in the second case.

Row vs column storage

Definition

Row storage respects internal structure of the data and puts the different items one next in a sequence

Column storage breaks the internal structure of the data to collate similar pieces

Why to use column ?

- to avoid scattered reads in the second example
- because the first example is naive. You want overheating devices for a range of time and will be in the second case.

Drawback of column storage

- a column organized file cannot be updated easily
- column storage is usually created from row storage in a postprocessing phase.

Compressing data

Data compression

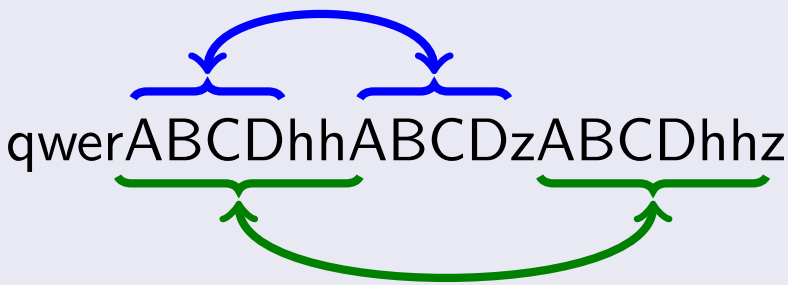
Main idea

- eliminate redundancy (e.g. LZ77)
- optimize encoding (Huffman coding)
- to squeeze more information into less bytes

Eliminate redundancy

LZ77

Replace redundancy with pointers



qwerABCDhh(6,4)z(11,6)z

Optimize encoding

Huffman coding idea

- use short codes for symbols repeated often
- build an optimal code set depending on symbols' frequencies

Optimize encoding

Encoding “faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa”

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

Frequencies

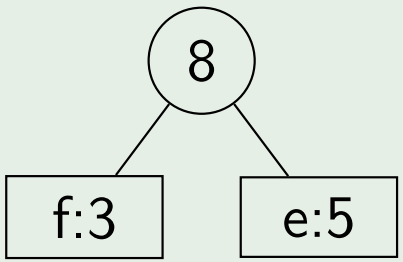
a	17	d	8
b	7	e	5
c	5	f	3

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

Frequencies

a	17	d	8
b	7	e	5
c	5	f	3

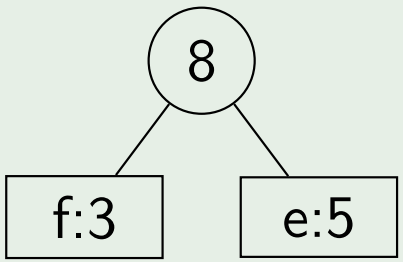


Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

Frequencies

a	17	d	8
b	7	ef	8
c	5		

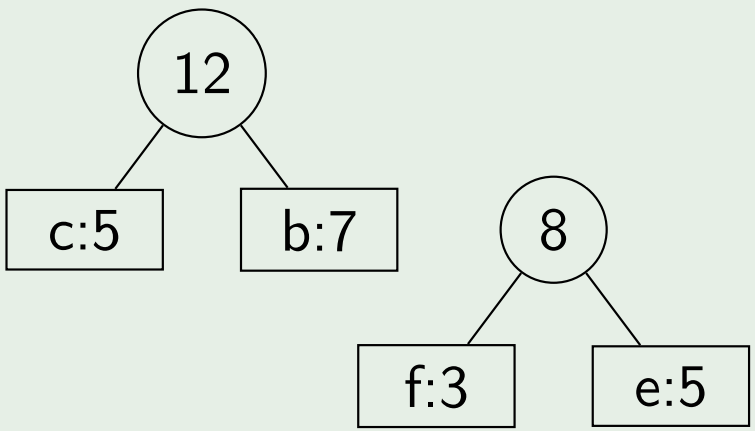


Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

Frequencies

a	17	d	8
bc	12	ef	8



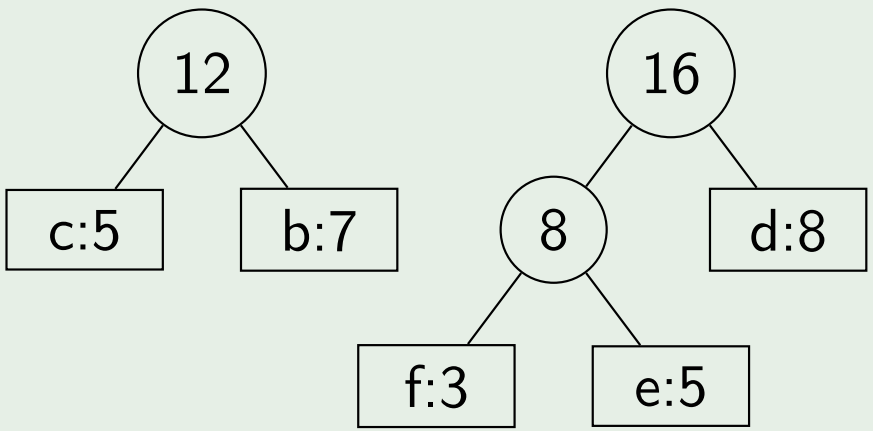
Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

Frequencies

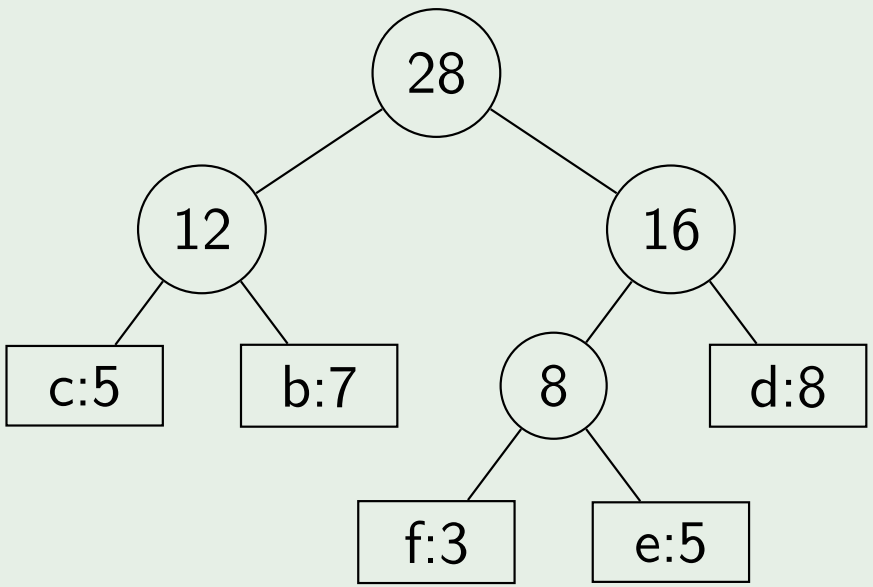
a 17 def 16

bc 12



Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

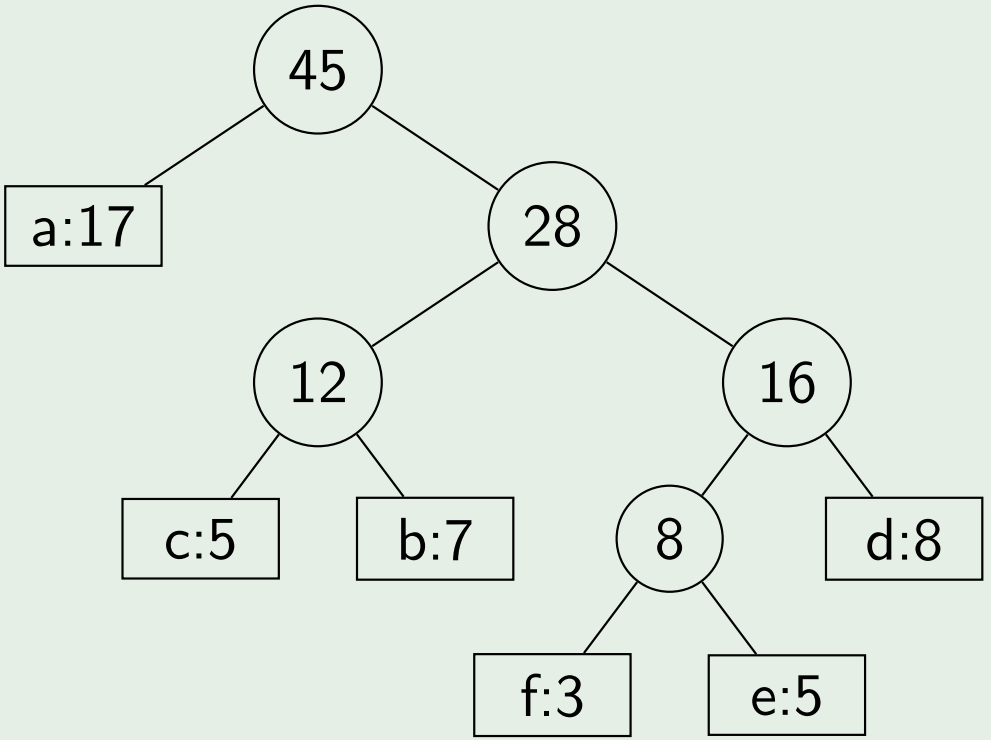


Frequencies

a 17 bcdef 28

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

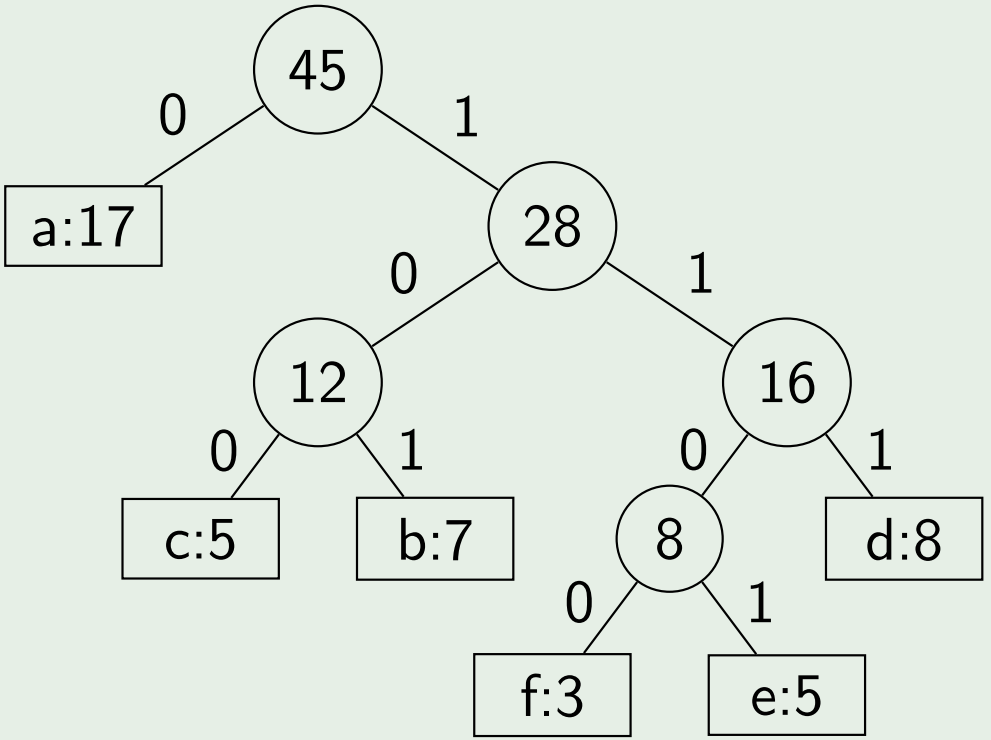


Frequencies

all 45

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"

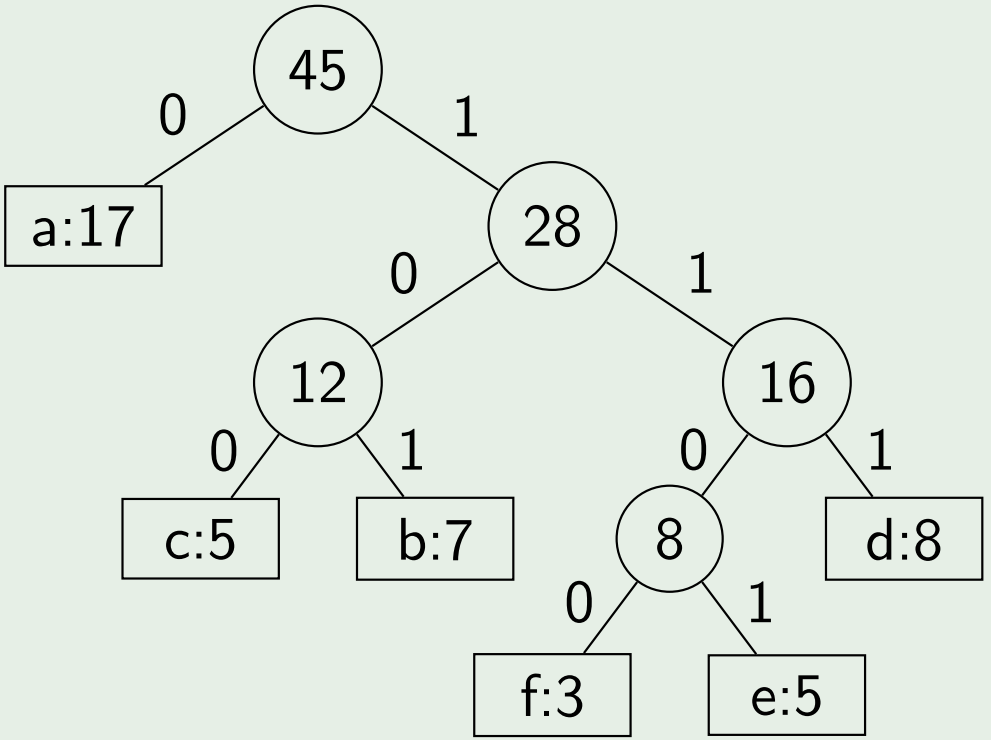


Frequencies

a	17	d	8
b	7	e	5
c	5	f	3

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"



Frequencies

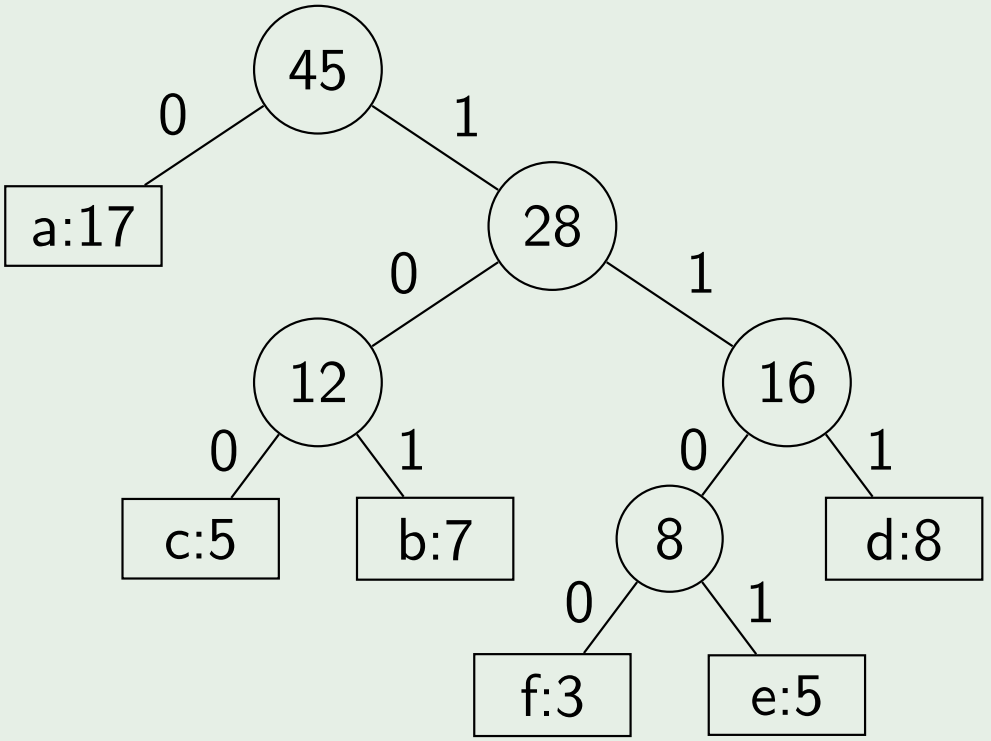
a	17	d	8
b	7	e	5
c	5	f	3

Codes

a	0	d	111
b	101	e	1101
c	100	f	1100

Optimize encoding

Encoding "faeaebacdeeabadeacdfbdaddadcafdacbaababbaaccaa"



Frequencies

a	17	d	8
b	7	e	5
c	5	f	3

Codes

a	0	d	111
b	101	e	1101
c	100	f	1100

*f*6*b*6*a*7*dd*57*d*4*ff*7*bf*78*fe*94*ab*490

28+6 bytes instead of 46

Classical compression algorithms

Best compression first

bzip2 high quality, size within 10% to 15% to best compressors and twice faster to compress, six times faster to decompress

gzip classical ziping with *deflate* algorithm, combination of LZ77 and Huffman Coding

LZO fast - twice faster than gzip - but less good - 50% bigger result. Composed of small blocks of 256 kB of compressed data

Snappy even faster - 250MB/s encoding, 500MB/s decoding on single core i7 - but even less good. Use internally by Google

A word on lossy compression

What does it mean, lossy ?

- that you will lose part of your data !
- but that you may not see the difference
- and it may allow to compress MUCH better

A word on lossy compression

What does it mean, lossy ?

- that you will lose part of your data !
- but that you may not see the difference
- and it may allow to compress MUCH better

Classical examples

jpg did you notice the images may be slightly blured ?

mp3 removes all the frequencies you cannot hear

A word on lossy compression

The main ideas

1. tranformation of the data, usually based on fourier transforms
2. sensibility (nb bits) is reduced on less important parts
   - e.g. high frequencies
3. thanks to the previous cuts, a lot of similar values (0) appear
4. use standard compression

A word on lossy compression

The main ideas

1. tranformation of the data, usually based on fourier transforms
2. sensibility (nb bits) is reduced on less important parts
   - e.g. high frequencies
3. thanks to the previous cuts, a lot of similar values (0) appear
4. use standard compression

Usage in scientific computing

- mostly when encoding the data, that is within a subdetector
- very seldom at the level of complete data files

Compression efficiency factors

Algorithm - a small factor

- always a tradeoff between speed and compression
- best algorithms will gain maximum a factor 2 in size compared to fast ones

Data - the main factor

content some contents are most compressable than others - zeros compress very well, while already compressed data don't

format human readable formats (XML) tend to compress very well as they are full of redundant markers

structure column storage compresses better than raw storage, as identical data types are close to each other

Compression and file splitting

Why to split a file ?

- for striping purposes
- in Map/Reduce approach
- see later talk

Compression and file splitting

Why to split a file ?

- for striping purposes
- in Map/Reduce approach
- see later talk

How to split a file ?

- splitting the compressed data blindly will not work
- in the best case, the compressing algorithm works with blocks and you can split at block boundaries
  - this is the case for LZO and bzip2
- for gzip and Snappy, you cannot split without decompressing

Compression and partial reads

Not working well...

- you can not decompress a random subpart of a file
- again in best case, you can decompress by block
  - this will work for LZO and bzip2
- thus a few bytes read will force a complete read and decompression of a block or not the whole file

Globally, is compression interesting ?

Yes if

- you data are very cold, that is not frequently accessed
  - and the gain of space is worth the annoyance
  - tapes are compressing systematically
- if you are really I/O bound and compression factor is high
- you use a fast compression algorithm (snappy) and gain space for little price

No if

- you become CPU bound due to (de)compression
  - typical with bzip2
- you do frequent partial reads
- you need to split and cannot use LZO or bzip2

Data addressing

Data addressing

The most important part of your data is your metadata

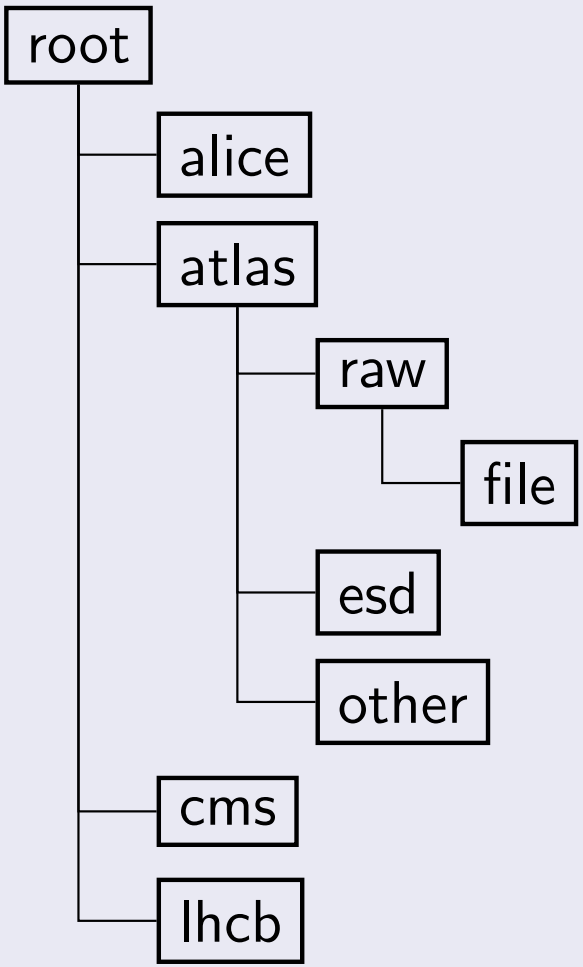
Addressing options

- good old hierarchical namespace
- databases
- object store approach

Traditionnal, hierarchical data addressing

Directory tree

- structured via "directories"
- containing "files" and subdirectories
- filesystem like

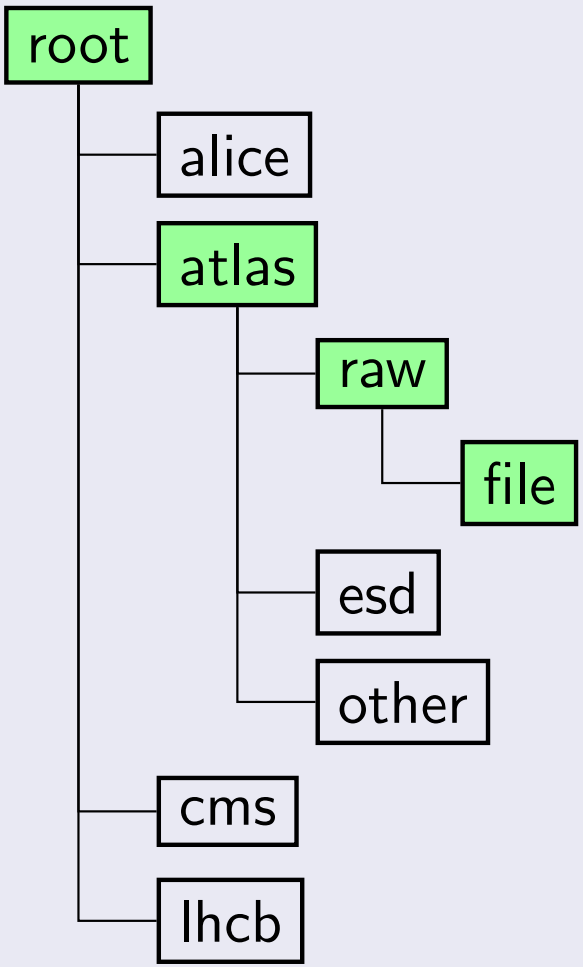


Traditionnal, hierarchical data addressing

Directory tree

- structured via "directories"
- containing "files" and subdirectories
- filesystem like

/atlas/raw/file



Hierarchy pros and contras

Pros

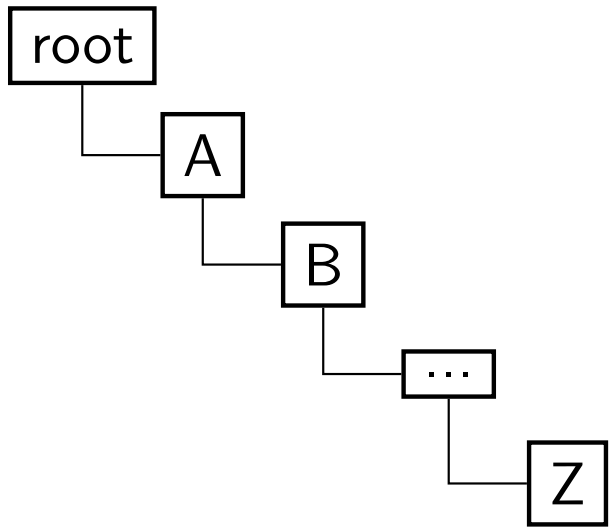
- easiness of use (for humans)
- helps to structure the data sets
  - easy to list/erase/deny access to a whole subtree
  - allows to implement quotas per directory
- atomic operations
  - in particular moves, removals
  - change of permissions
- always consistent

Contras

- becomes slow for deep hierarchies
- in general scales badly

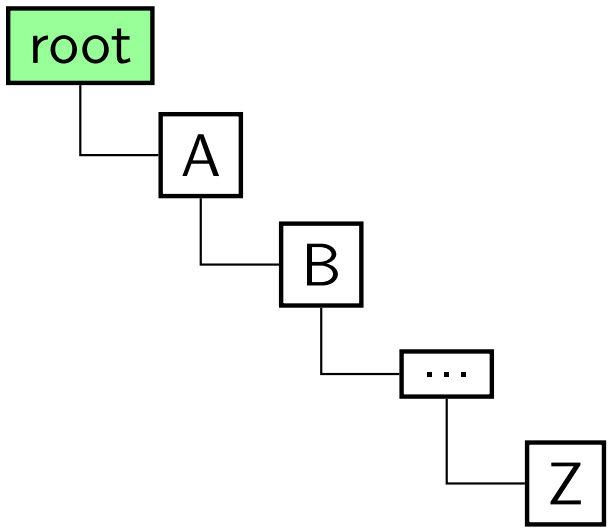
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :



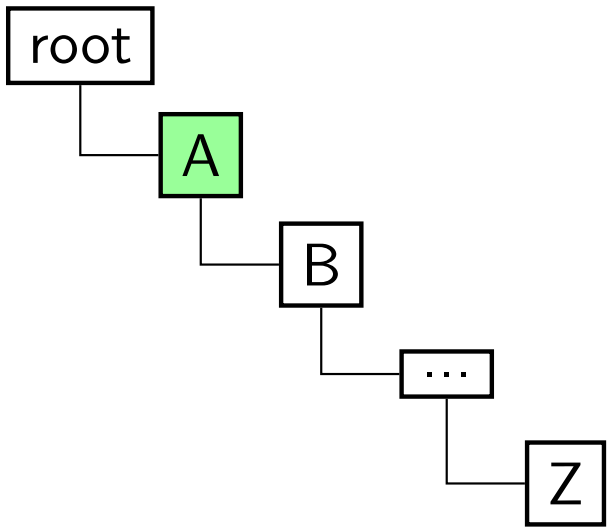
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right



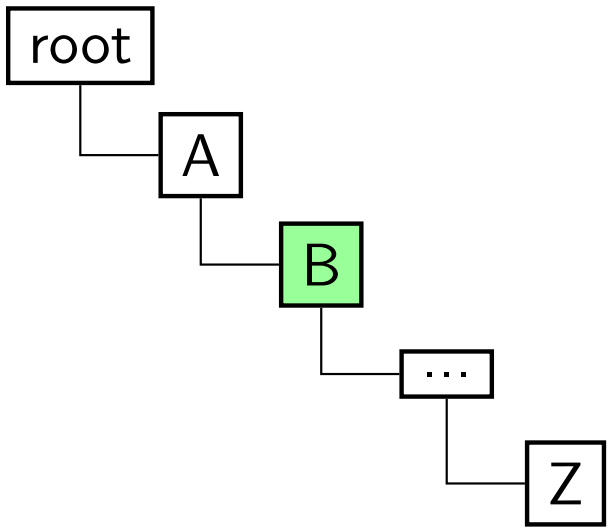
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right
  - find *A* within *root*
  - check you have execution right



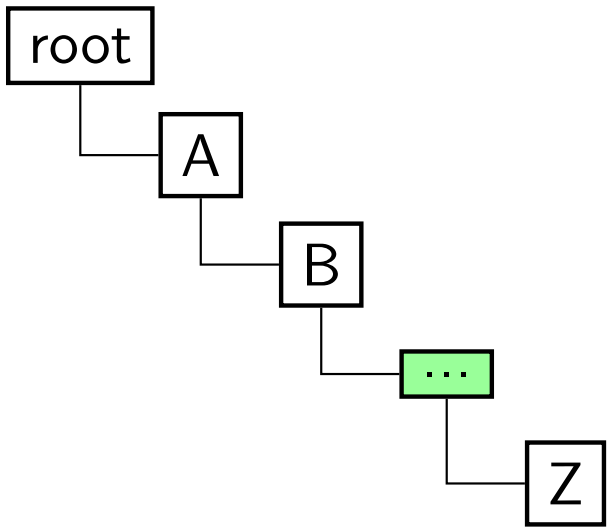
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right
  - find *A* within *root*
  - check you have execution right
  - find *B* within *A*
  - check you have execution right



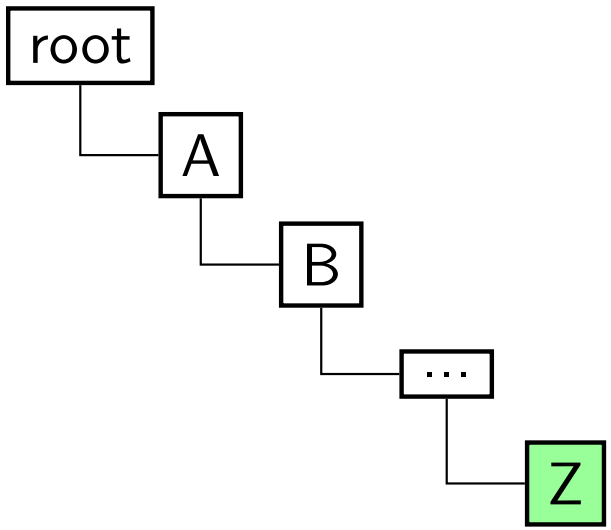
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right
  - find *A* within *root*
  - check you have execution right
  - find *B* within *A*
  - check you have execution right
  - ... repeat until *Y* ...



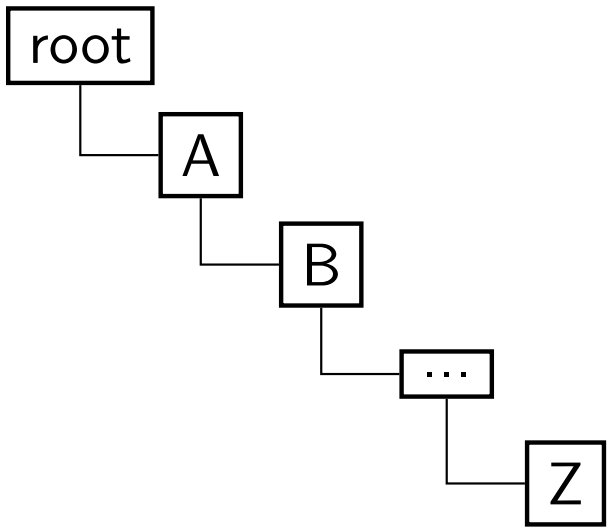
Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right
  - find *A* within *root*
  - check you have execution right
  - find *B* within *A*
  - check you have execution right
  - ... repeat until *Y* ...
  - find *Z* within *Y*
  - check you have read/write rights



Influence of depth

- Suppose you access '/A/B/C/.../Z'
- The following will happen :
  - find *root*
  - check you have execution right
  - find *A* within *root*
  - check you have execution right
  - find *B* within *A*
  - check you have execution right
  - ... repeat until *Y* ...
  - find *Z* within *Y*
  - check you have read/write rights



→ $2n + 1$ operations for depth n

depth seen at CERN (AFS, CASTOR) : 20-25

Atomicity and consistency consequences

root is a bottle neck

- every single requests starts at root
- no easy parallelization of this part because of atomicity

Atomicity and consistency consequences

root is a bottle neck

- every single requests starts at root
- no easy parallelization of this part because of atomicity

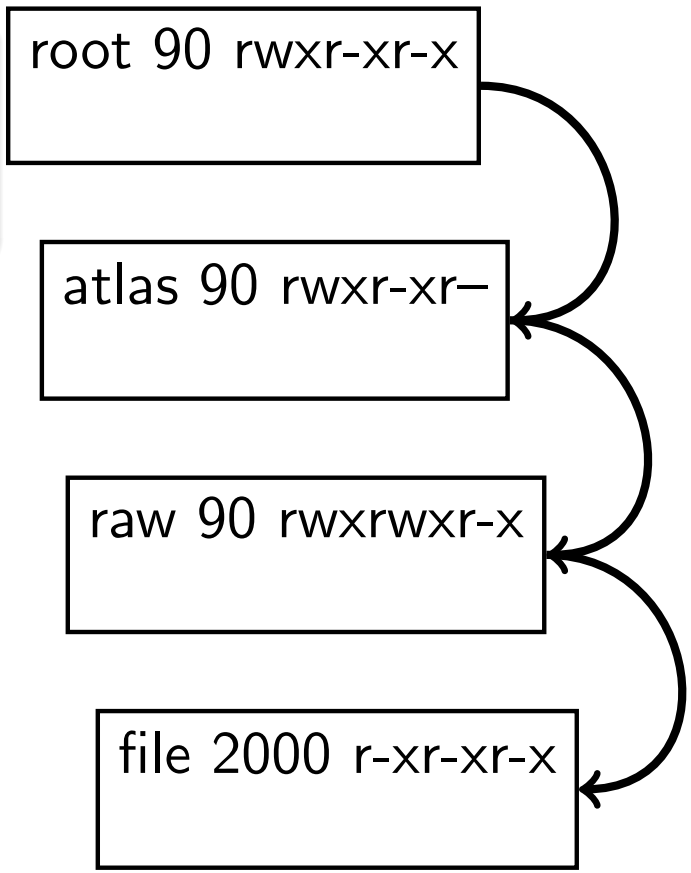
consistent recursive listing/searching is a nightmare

- you would need to lock all subtree for the whole listing time
- makes full listing impossible

Some partial solutions - rights

Precomputation

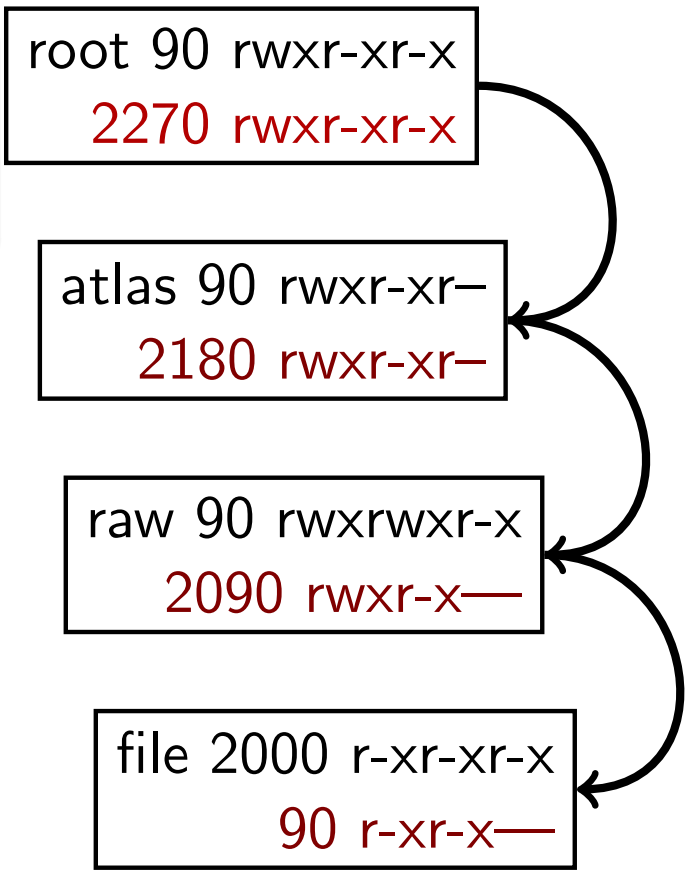
- precompute effective rights/size/... at every level



Some partial solutions - rights

Precomputation

- precompute effective rights/size/... at every level



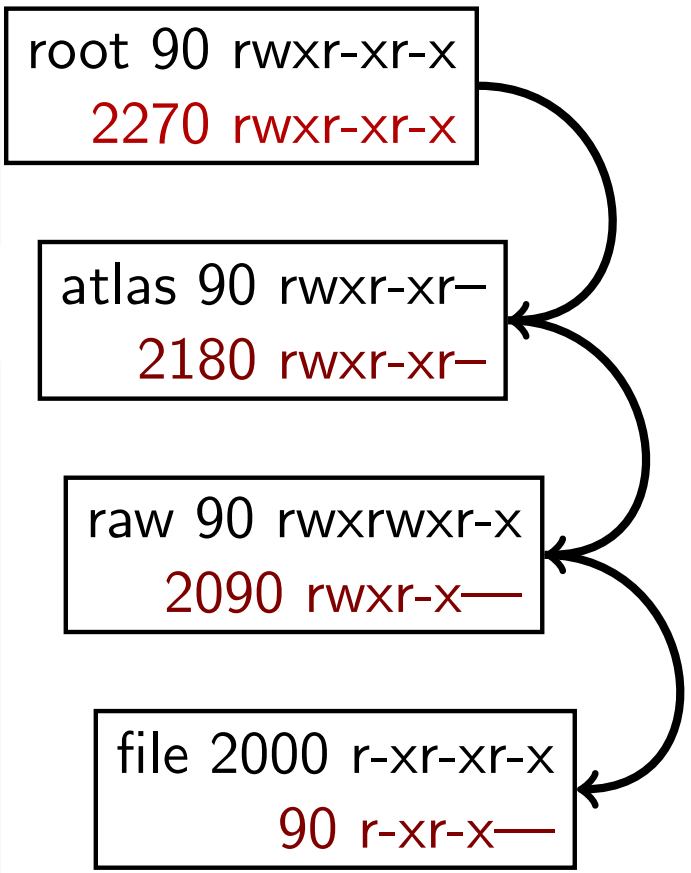
Some partial solutions - rights

Precomputation

- precompute effective rights/size/... at every level

But...

- updating a top level right means full recomputation !
- the complete expansion for each file may be extremely heavy
- especially if complex ACLs are used at directory levels



Some partial solutions - rights

Avoiding full recomputation

- one can mark a subtree as "dirty"
- can then be recomputed offline
- in case we hit a dirty tree, we go back to tree scanning

Some partial solutions - rights

Avoiding full recomputation

- one can mark a subtree as "dirty"
- can then be recomputed offline
- in case we hit a dirty tree, we go back to tree scanning

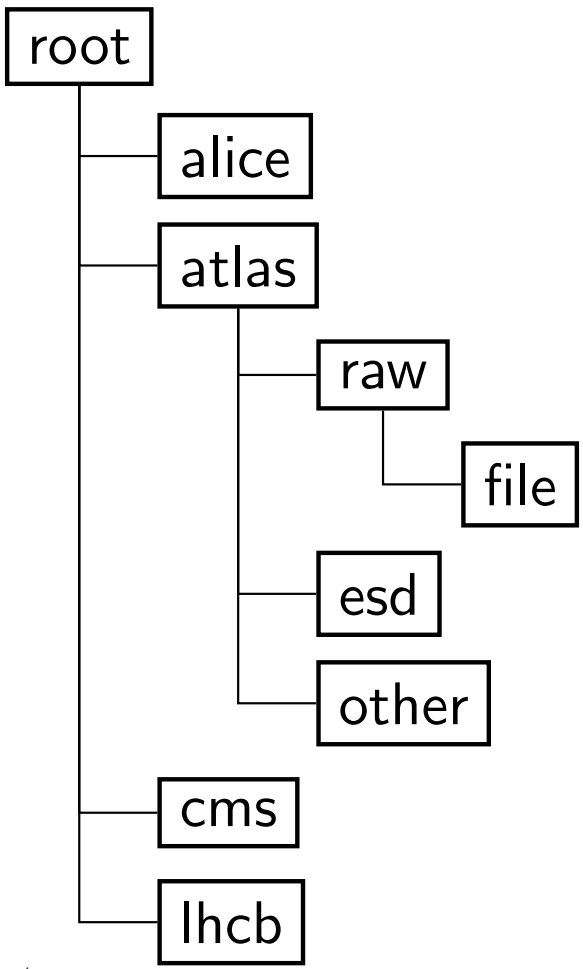
But...

- tree scanning is needed to check for the dirty flag !
- actually, special structures like bit map indexes can help speeding it up

Some partial solutions - root bottleneck

Path indexing

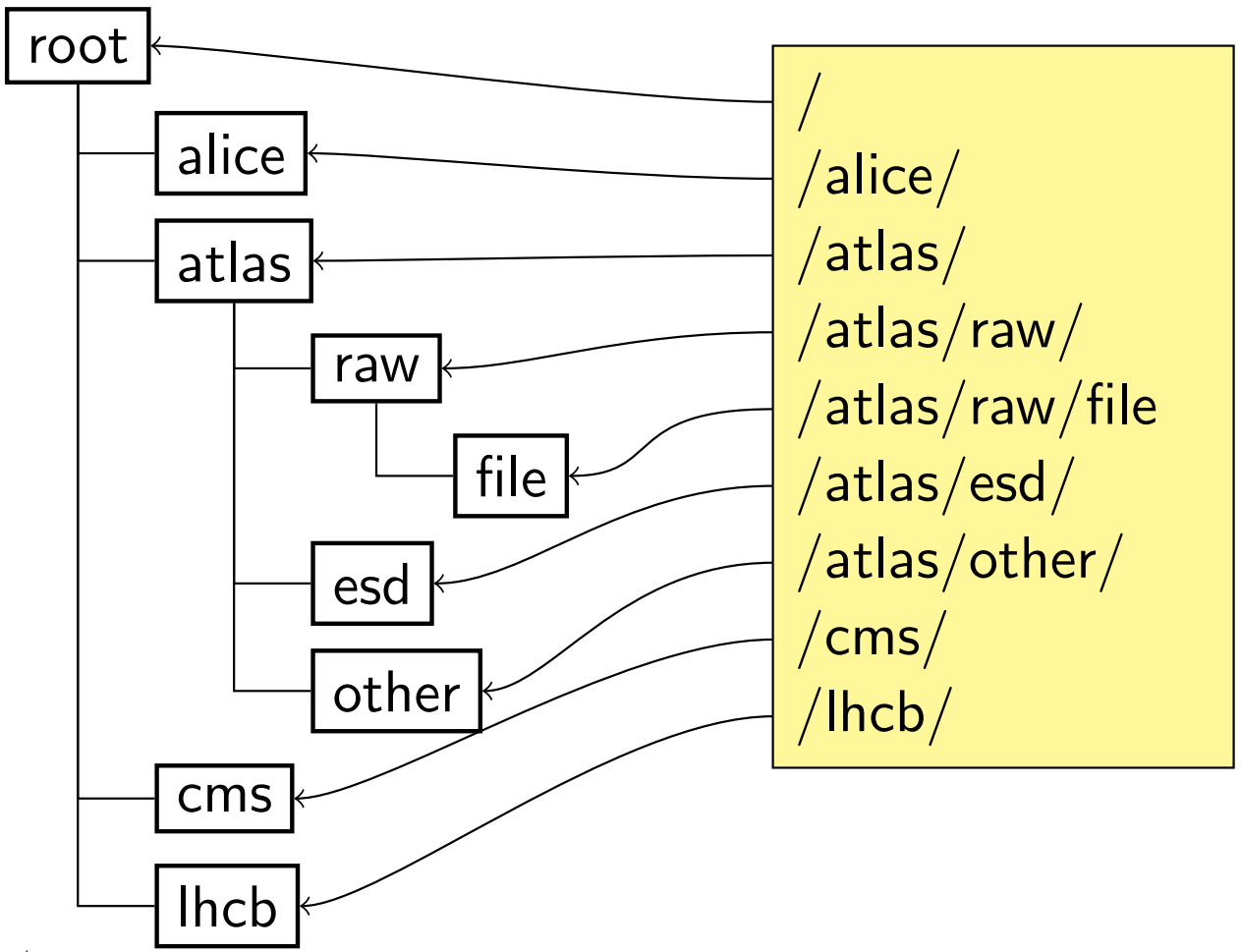
- keep an index of full paths and their locations



Some partial solutions - root bottleneck

Path indexing

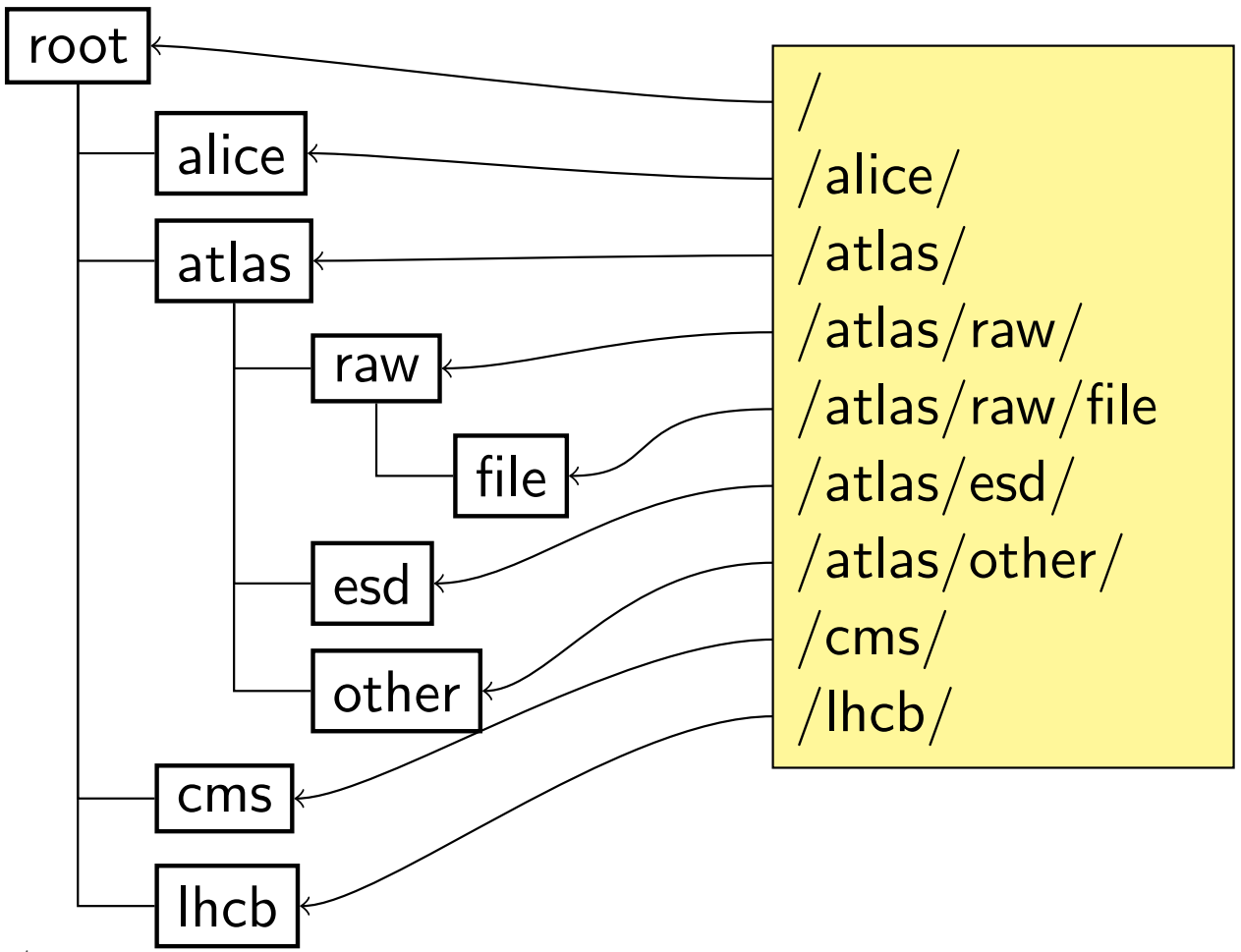
- keep an index of full paths and their locations



Some partial solutions - root bottleneck

Path indexing

- keep an index of full paths and their locations



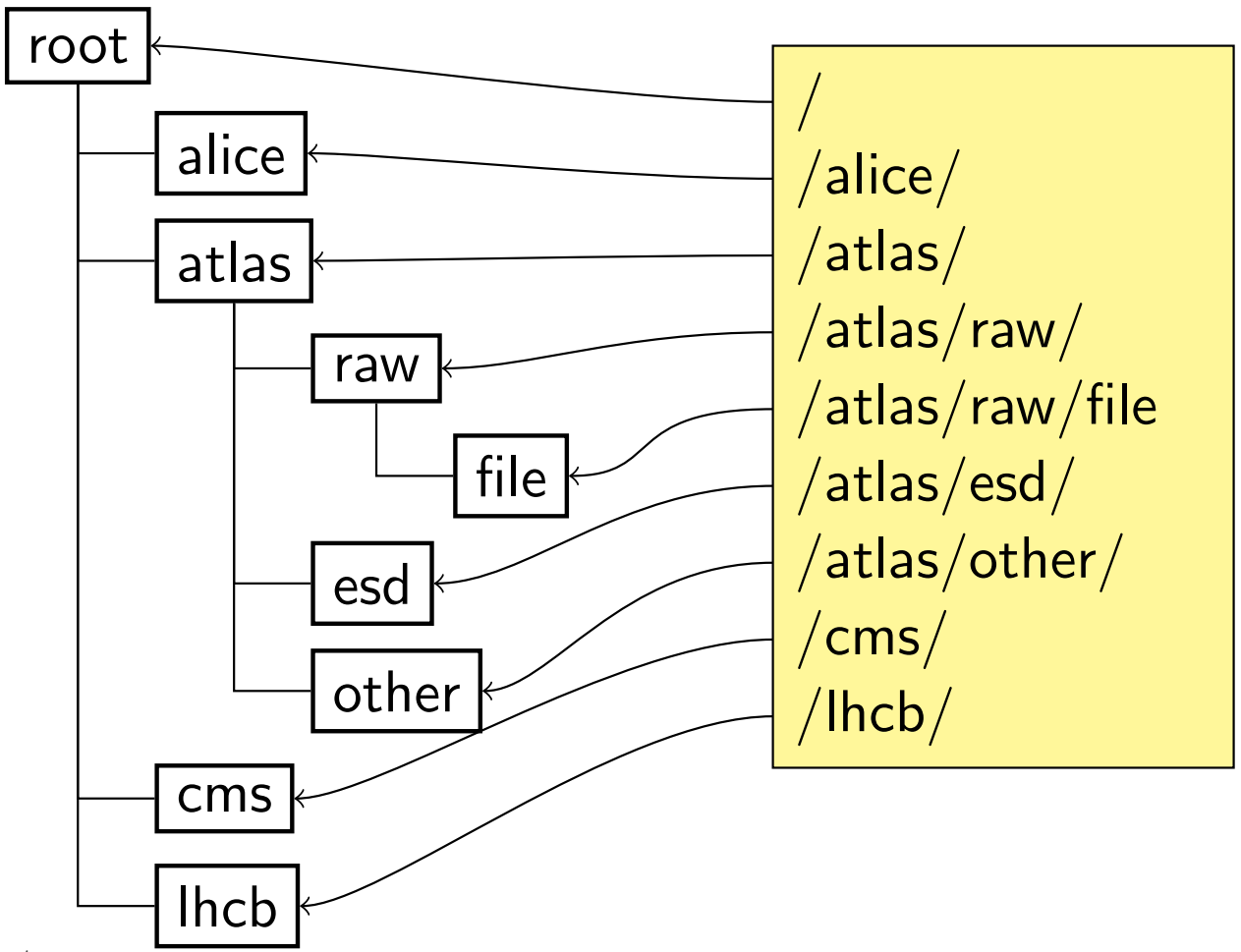
But...

- just thing of a rename of atlas into titan...

Some partial solutions - root bottleneck

Path indexing

- keep an index of full paths and their locations



But...

- just thing of a rename of atlas into titan...

Ideas

- use redirections, dirty flags, ...

Some partial solutions - conclusion

→ very high complexity

no scalable solution on the market so far

Radical change : flat namespaces

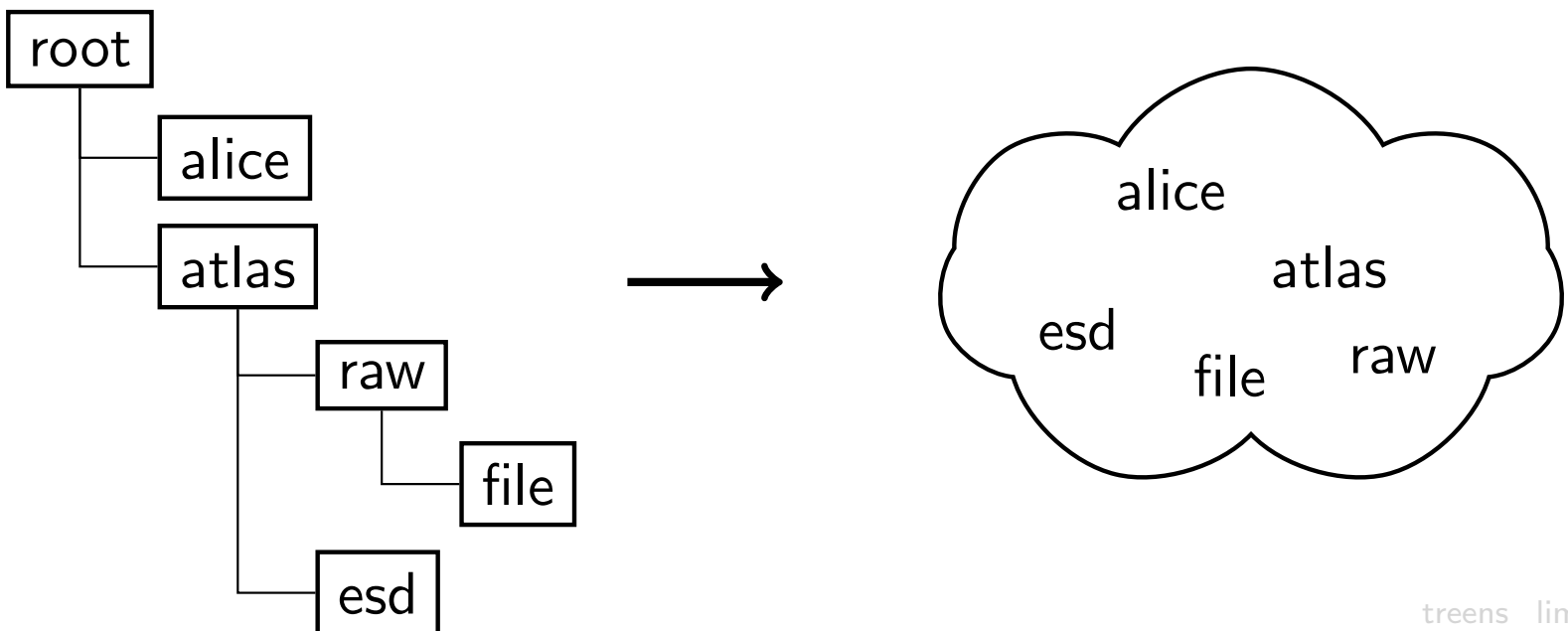
The idea

- drop the hierarchy
- only have a sea of objects with a unique identifier
- forbid object modifications (only appends)
- forbid renames
  - or implement them as copy + drop

Radical change : flat namespaces

The idea

- drop the hierarchy
- only have a sea of objects with a unique identifier
- forbid object modifications (only appends)
- forbid renames
  - or implement them as copy + drop



Radical change : flat namespaces

Infinitely scalable namespace

- client can hash the entry path
- and connect accordingly to the right server

Radical change : flat namespaces

Infinitely scalable namespace

- client can hash the entry path
- and connect accordingly to the right server

You lose

- data management aspect, left to user
- listing facilities
  - can be implemented in a Map/Reduce way

About hashing on the client

Large systems are dynamic

- i.e. they change over time
  - adding/removing hardware
  - changing network topology
- data may need to be rebalanced (see later)
- the client hashing algorithm needs to deal with that

About hashing on the client

Large systems are dynamic

- i.e. they change over time
  - adding/removing hardware
  - changing network topology
- data may need to be rebalanced (see later)
- the client hashing algorithm needs to deal with that

CRUSH : the ceph's hash

- achieves pseudo-random but deterministic data distribution
- supports replication and erasure coding (see later)
- taking into account storage structure (machines, racks, rows, ...)
- minimizing the data movements in case of cluster modifications

Conclusion

Conclusion

Key messages of the day

- Choose well your data format
- Think twice before compressing
- Value your metadata