- http://elle.epfl.ch/wiki
- Logiciels-Libres@groupes.epfl.ch
- epfl.comp.logiciels-libres

Ce que vous ne voulez pas savoir sur les systèmes de fichiers

Vittoria Rezzonico, SB-IT

7 octobre 2009

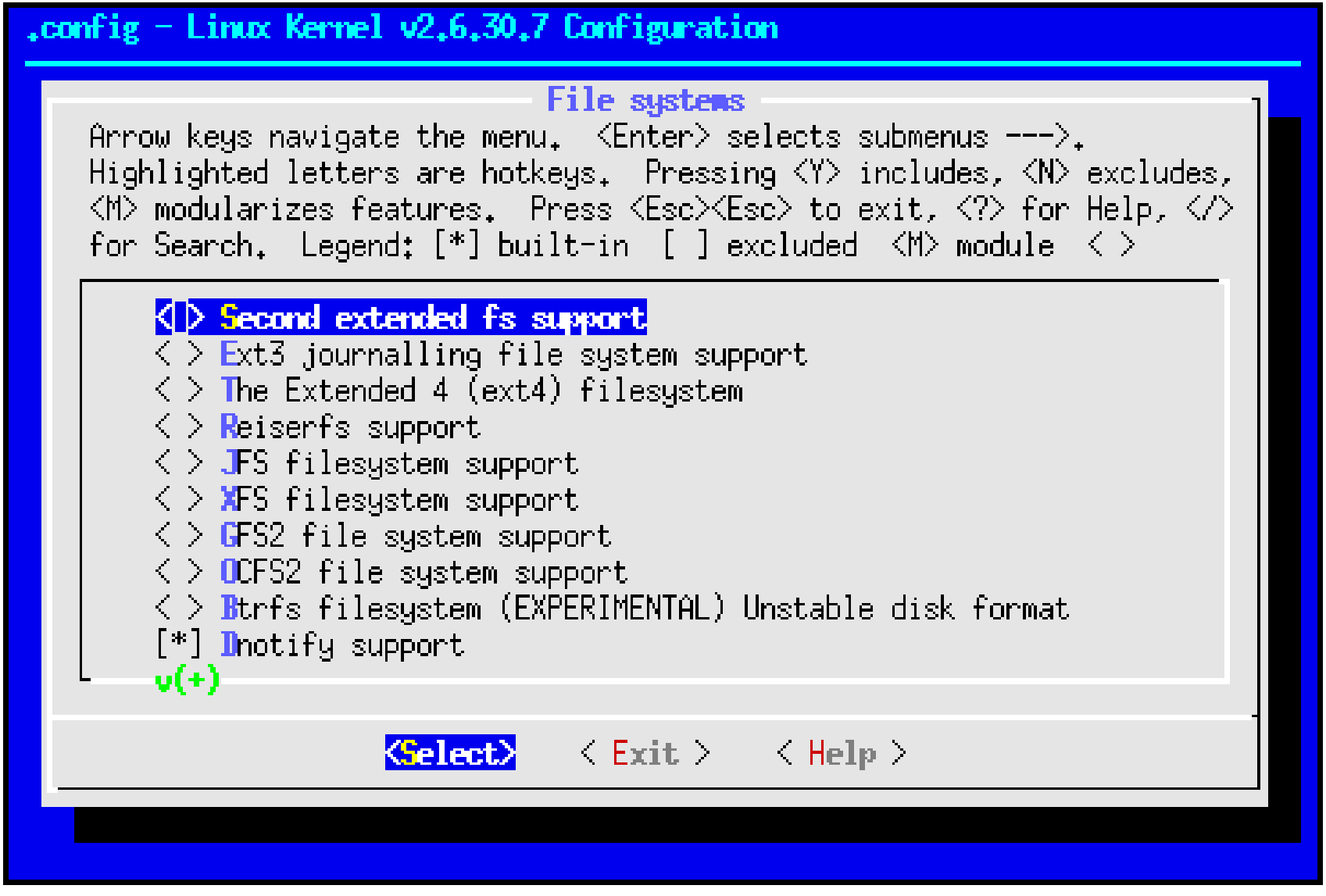
Nagios®

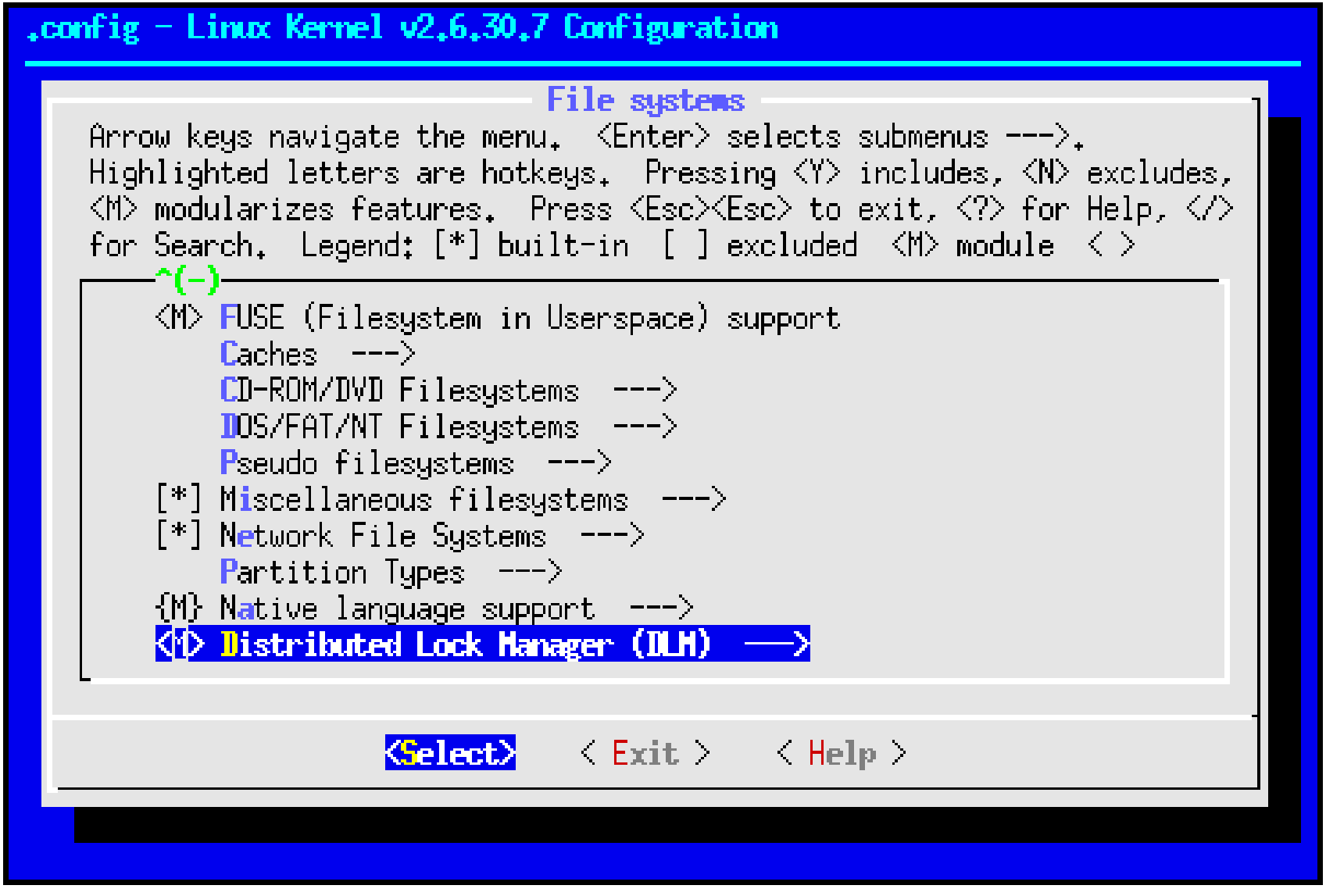
Séminaires passés

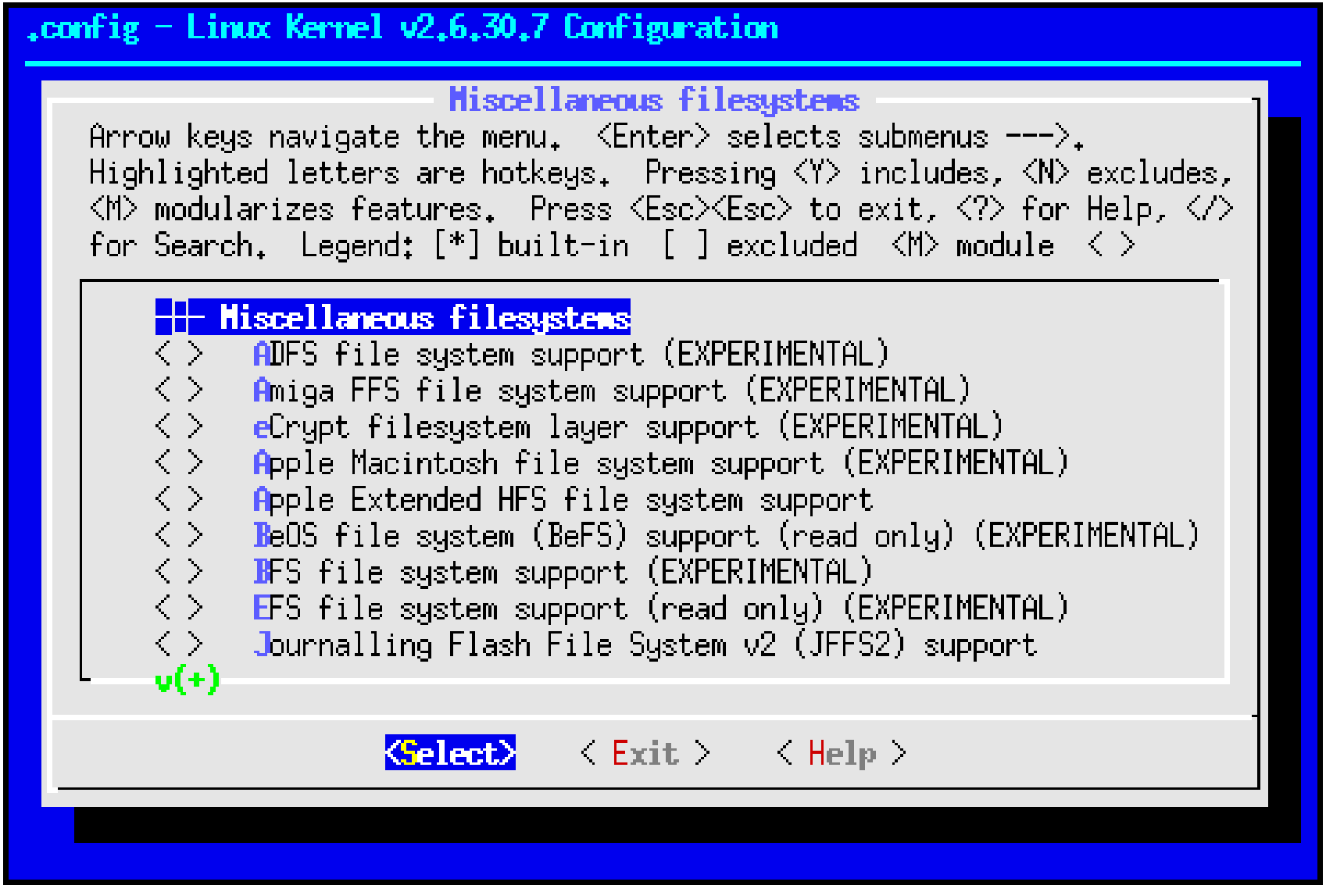
Extensions Firefox	Michela	Tout public	3 novembre 2008
IPv6	Laurent	Avancé	28 novembre 2008
Tequila	Claude	Tout public	19 mars 2009
PXE	Vicky et Ale	Avancé	10 juin 2009

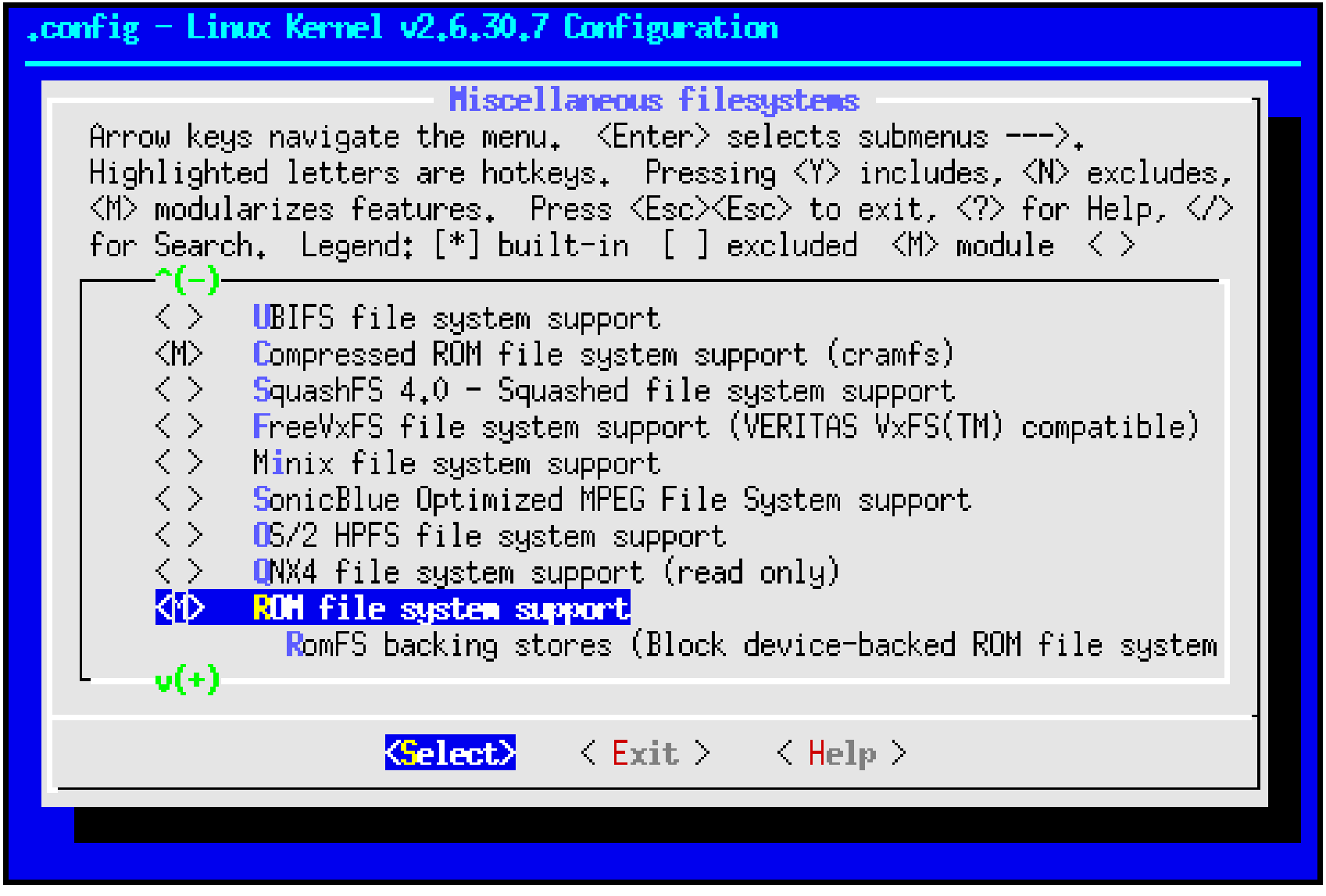
Bientôt ?

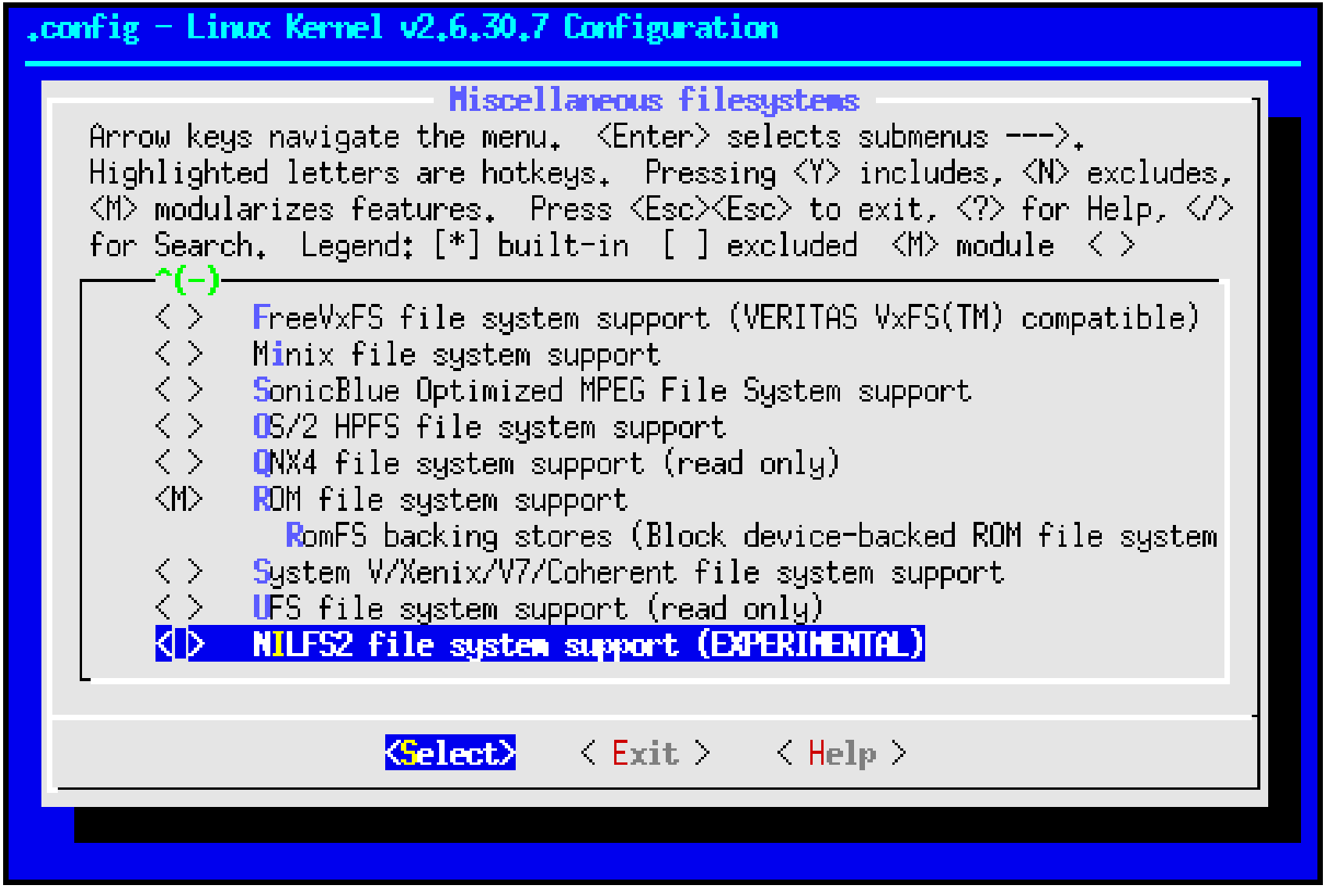
- GPG
- Extensions Thunderbird
- Extensions Firefox, II
- PAM
- vi
- Nagios

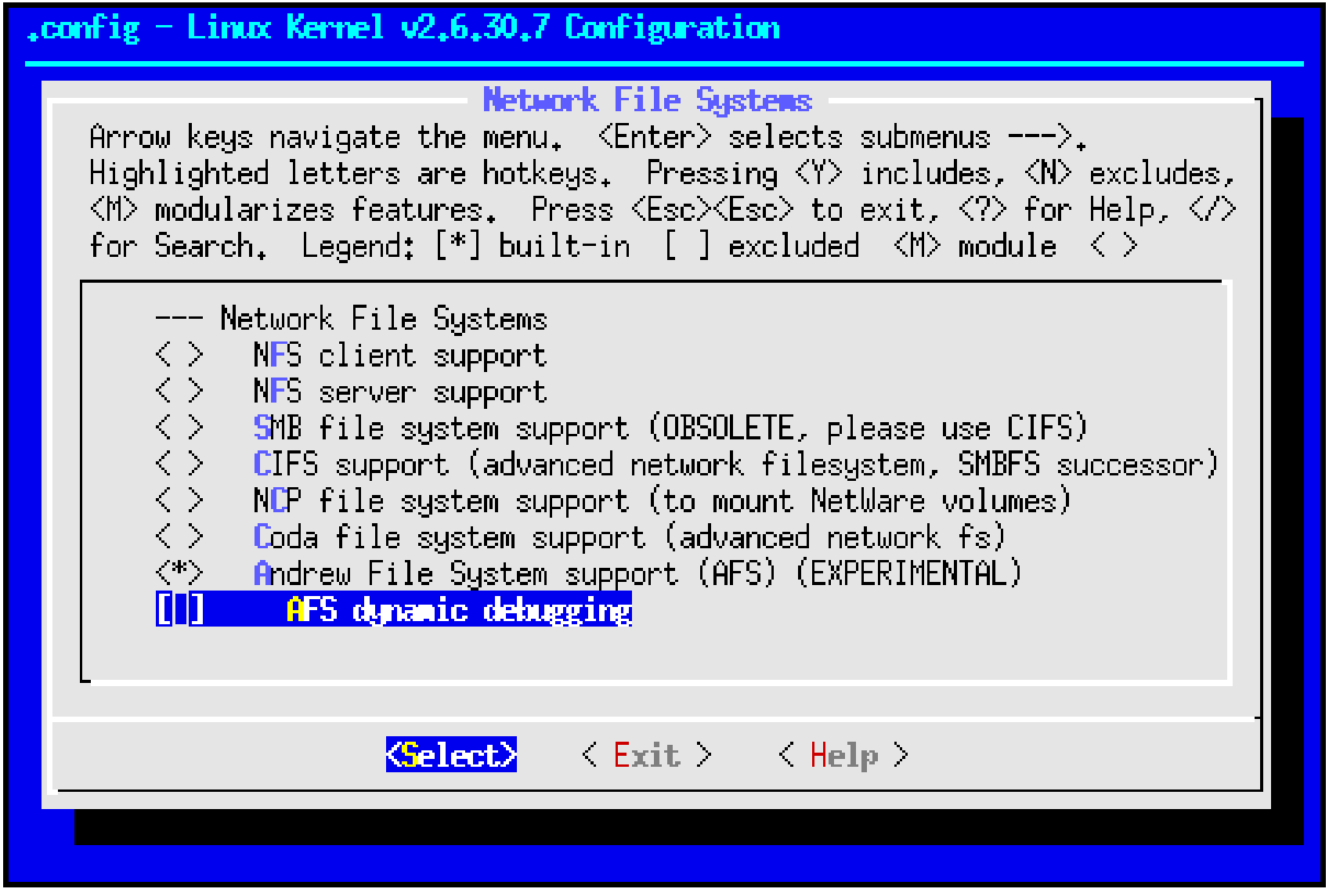

.config - Linux Kernel v2.6.30.7 Configuration
File systems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
<*> Second extended fs support
< > Ext3 journalling file system support
< > The Extended 4 (ext4) filesystem
< > Reiserfs support
< > JFS filesystem support
< > XFS filesystem support
< > GFS2 file system support
< > OCFS2 file system support
< > Btrfs filesystem (EXPERIMENTAL) Unstable disk format
[*] Dnotify support
v(+)
<Select> < Exit > < Help >

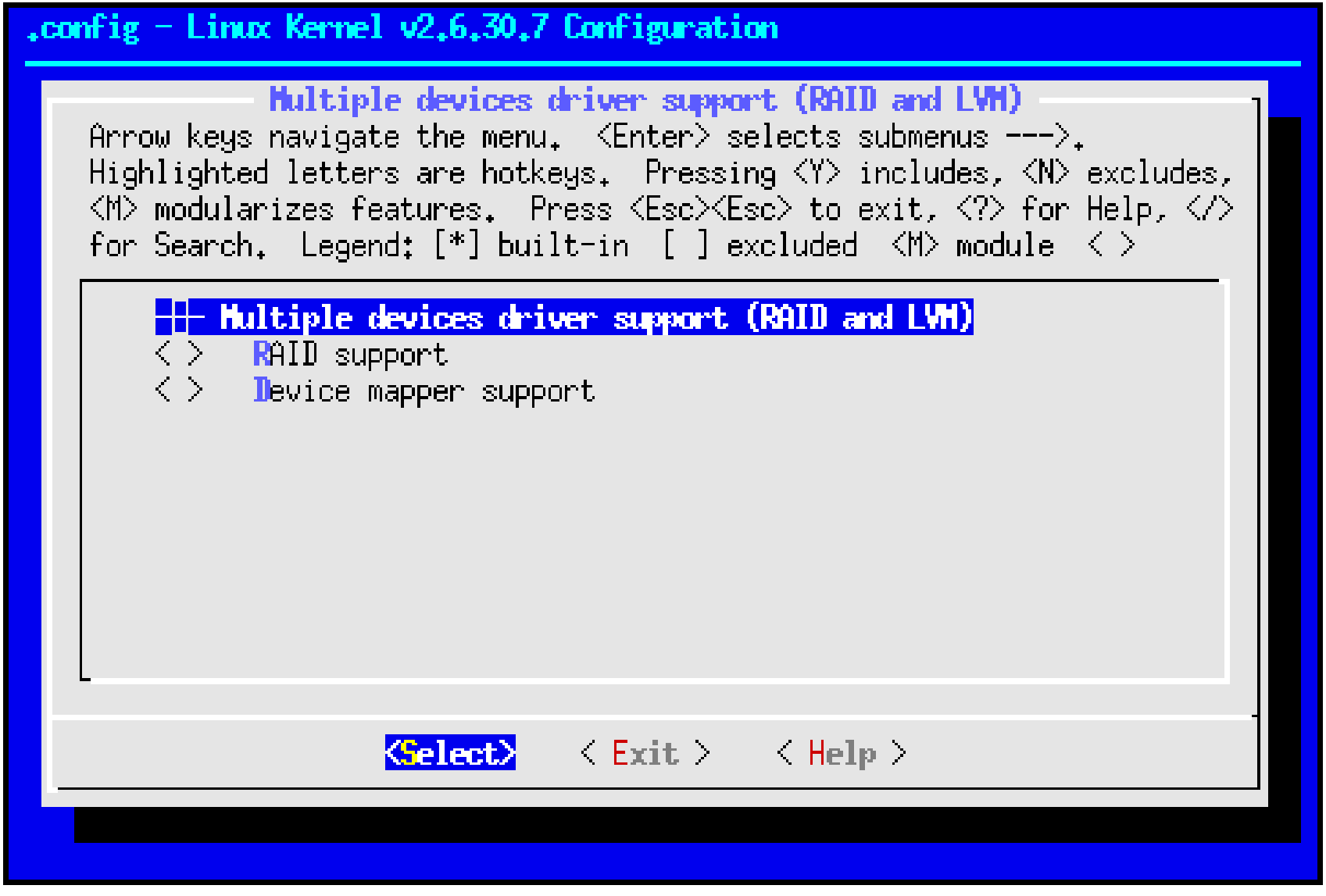

.config - Linux Kernel v2.6.30.7 Configuration
File systems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
<M> FUSE (Filesystem in Userspace) support
Caches --->
CD-ROM/DVD Filesystems --->
DOS/FAT/NT Filesystems --->
Pseudo filesystems --->
[*] Miscellaneous filesystems --->
[*] Network File Systems --->
Partition Types --->
{M} Native language support --->
<M> Distributed Lock Manager (DLM) --->
<Select> < Exit > < Help >


.config - Linux Kernel v2.6.30.7 Configuration
Miscellaneous filesystems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
Miscellaneous filesystems
< > ADFS file system support (EXPERIMENTAL)
< > Amiga FFS file system support (EXPERIMENTAL)
< > eCrypt filesystem layer support (EXPERIMENTAL)
< > Apple Macintosh file system support (EXPERIMENTAL)
< > Apple Extended HFS file system support
< > BeOS file system (BeFS) support (read only) (EXPERIMENTAL)
< > BFS file system support (EXPERIMENTAL)
< > EFS file system support (read only) (EXPERIMENTAL)
< > Journalling Flash File System v2 (JFFS2) support
v(+)
<Select> < Exit > < Help >


.config - Linux Kernel v2.6.30.7 Configuration
Miscellaneous filesystems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
^(-)
< > UBIFS file system support
<M> Compressed ROM file system support (cramfs)
< > SquashFS 4.0 - Squashed file system support
< > FreeVxFS file system support (VERITAS VxFS(TM) compatible)
< > Minix file system support
< > SonicBlue Optimized MPEG File System support
< > OS/2 HPFS file system support
< > QNX4 file system support (read only)
<M> ROM file system support
RomFS backing stores (Block device-backed ROM file system
v(+)
<Select> < Exit > < Help >


.config - Linux Kernel v2.6.30.7 Configuration
Miscellaneous filesystems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
^(-)
< > FreeVxFS file system support (VERITAS VxFS(TM) compatible)
< > Minix file system support
< > SonicBlue Optimized MPEG File System support
< > OS/2 HPFS file system support
< > QNX4 file system support (read only)
<M> ROM file system support
RomFS backing stores (Block device-backed ROM file system
< > System V/Xenix/V7/Coherent file system support
< > UFS file system support (read only)
< > NILFS2 file system support (EXPERIMENTAL)
<Select> < Exit > < Help >


.config - Linux Kernel v2.6.30.7 Configuration
Network File Systems
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
--- Network File Systems
< > NFS client support
< > NFS server support
< > SMB file system support (OBSOLETE, please use CIFS)
< > CIFS support (advanced network filesystem, SMBFS successor)
< > NCP file system support (to mount NetWare volumes)
< > Coda file system support (advanced network fs)
<*> Andrew File System support (AFS) (EXPERIMENTAL)
[ ] AFS dynamic debugging
<Select> < Exit > < Help >


.config - Linux Kernel v2.6.30.7 Configuration
Multiple devices driver support (RAID and LVM)
Arrow keys navigate the menu. <Enter> selects submenus --->.
Highlighted letters are hotkeys. Pressing <Y> includes, <N> excludes,
<M> modularizes features. Press <Esc><Esc> to exit, <?> for Help, </>
for Search. Legend: [*] built-in [ ] excluded <M> module < >
Multiple devices driver support (RAID and LVM)
< > RAID support
< > Device mapper support
<Select> < Exit > < Help >

Partition de 50GB :

- mkfs.ext3 : 23.998s
- mkfs.reiserfs : 2.759s

Partition de 50GB :

- mkfs.ext3 : 23.998s
- mkfs.reiserfs : 2.759s
- mkfs.xfs : 0.522s

Filesystem par default dans Linux

- ZFS
- ext4
- reiser4
- btrfs

Hexdump obligatoire

```
01b0  00 00 00 00 00 00 00 00  18 7a 07 00 00 00 00 01  |.........z......|
01c0  01 00 83 fe 3f 81 3f 00  00 00 c3 dd 1f 00 00 00  |....?.?.........|
01d0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
*
01f0  00 00 00 00 00 00 00 00  00 00 00 00 00 00 55 aa  |..............U.|
```

- La table des partitions commence à 0x1be (446)

Just kidding !

Je veux mes fichiers maintenant

J'accepte d'acceder à des données peut-être corrompues dans un temps irraisonnable

- Fiabilité
- Performances

Contenus

Définitions

Fiabilité
- fsck
- journal
- checksums
- checkpoints

Performances

Fiabilité et performances

Définitions

Fiabilité
- fsck
- journal
- checksums
- checkpoints

Performances

Fiabilité et performances

- façon d'organiser les données
  - sur les disques
  - sur le réseau
  - dans la RAM

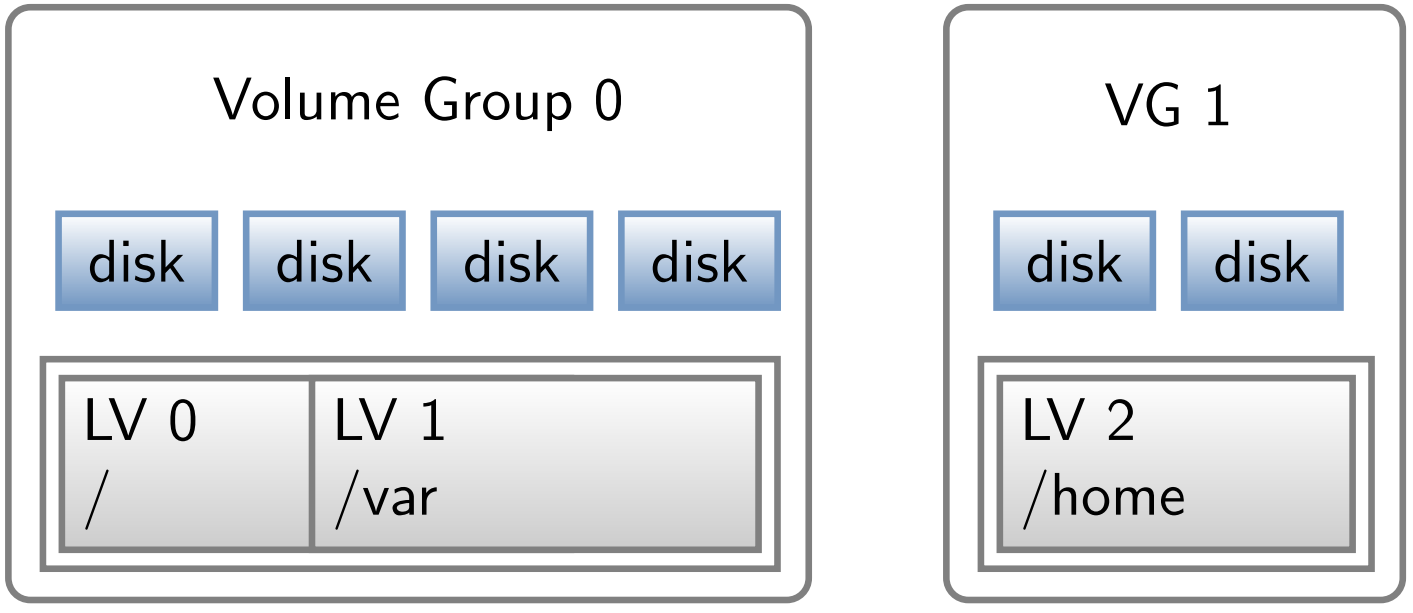
Gestion par volumes logiques

- pas tous les systèmes de fichiers l'incluent
- on peut s'appuyer sur LVM

Le système de fichiers au sense strict

- sur un volume

Illustration



Définitions

Fiabilité
- fsck
- journal
- checksums
- checkpoints

Performances

Fiabilité et performances

- Contrôle d'intégrité
  - fsck
  - comparaisons avec un log (journal)
  - checksums

- Contrôle d'intégrité
  - fsck
  - comparaisons avec un log (journal)
  - checksums
- Consistance des données à tout moment
  - checkpoints

- pour vérifier la cohérence d'un système de fichiers
- et corriger les incohérences
- lent, en général exécuté au démarrage après un arrêt brutal
- en général exécuté offline

- pour vérifier la cohérence d'un système de fichiers
- et corriger les incohérences
- lent, en général exécuté au démarrage après un arrêt brutal
- en général exécuté offline

Temps irraisonnable

- les opérations d'écriture sont enregistrées dans un log avant d'être exécutées
- on peut logguer uniquement les métadonnées
- les données sont écrites deux fois !
- log-structured filesystems
- optimisations : regrouper les écritures

- les opérations d'écriture sont enregistrées dans un log avant d'être exécutées
- on peut logguer uniquement les métadonnées
- les données sont écrites deux fois !
- log-structured filesystems
- optimisations : regrouper les écritures

Est-ce adapté aux systèmes de fichiers de grande taille ?

- comment savoir qu'il n'y a pas eu corruption des données
- checksum du fichier, des blocks,...
- stocké dans le "bloc parent"

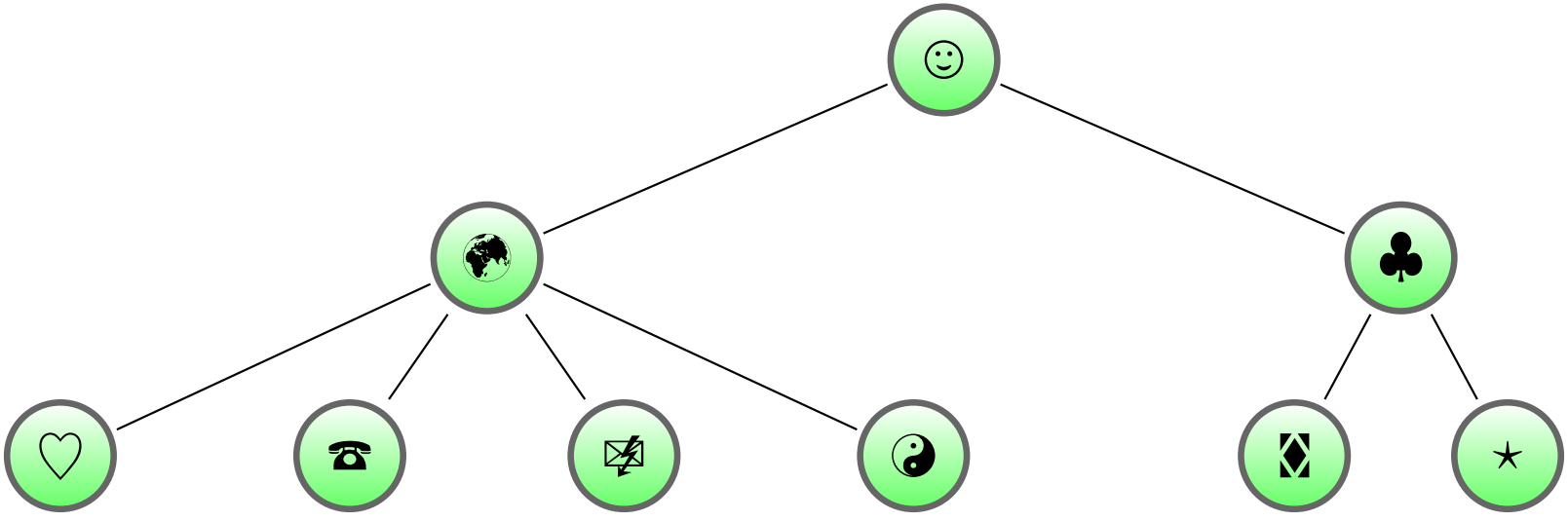
- éliminent la nécessité de journal et de fsck
- car le système de fichiers est toujours consistant

- éliminent la nécessité de journal et de fsck
- car le système de fichiers est toujours consistant

La solution à tous nos problèmes

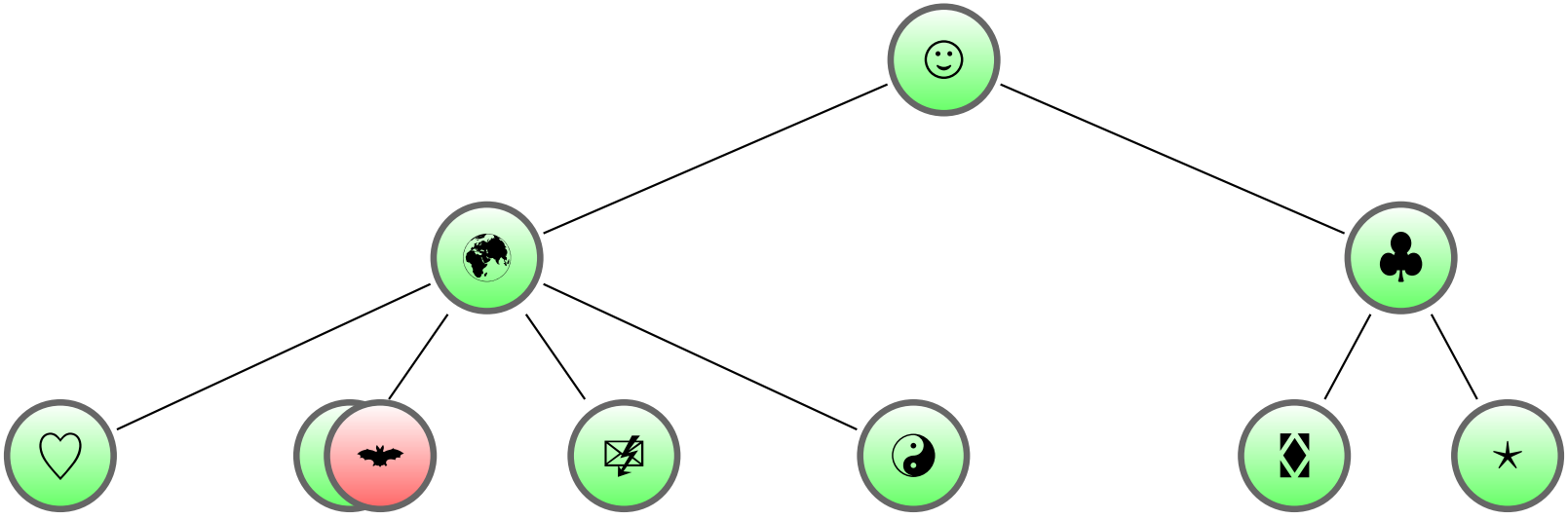
checkpoints - comment ça marche

SB-IT

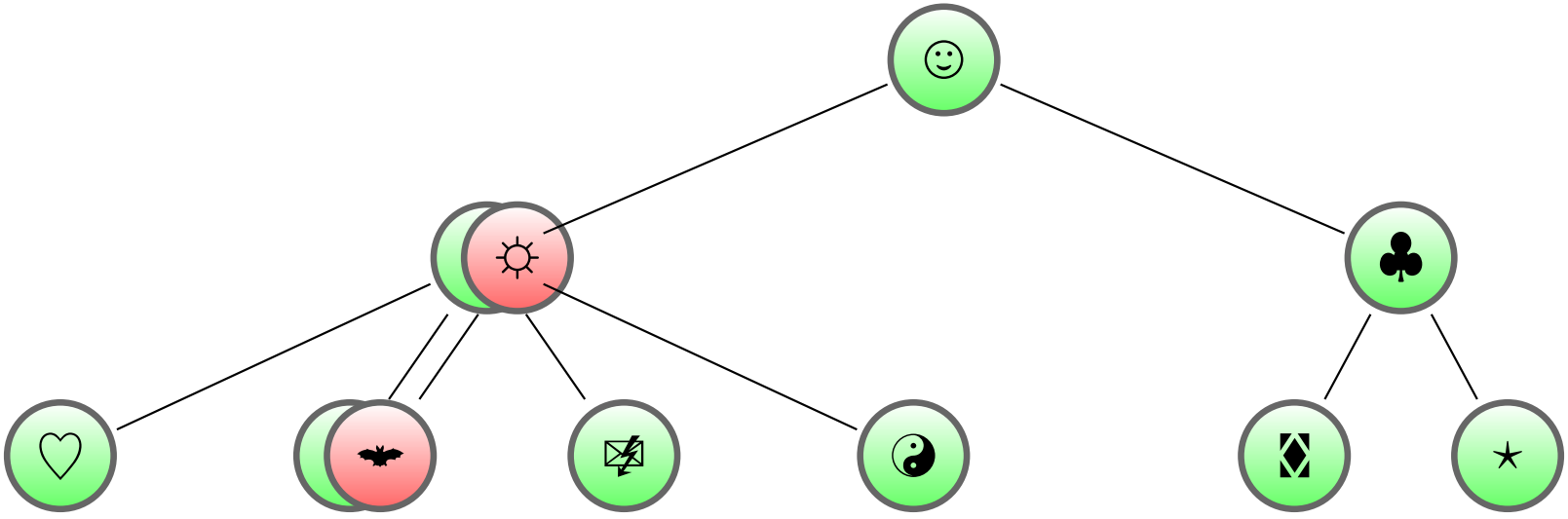


checkpoints - comment ça marche

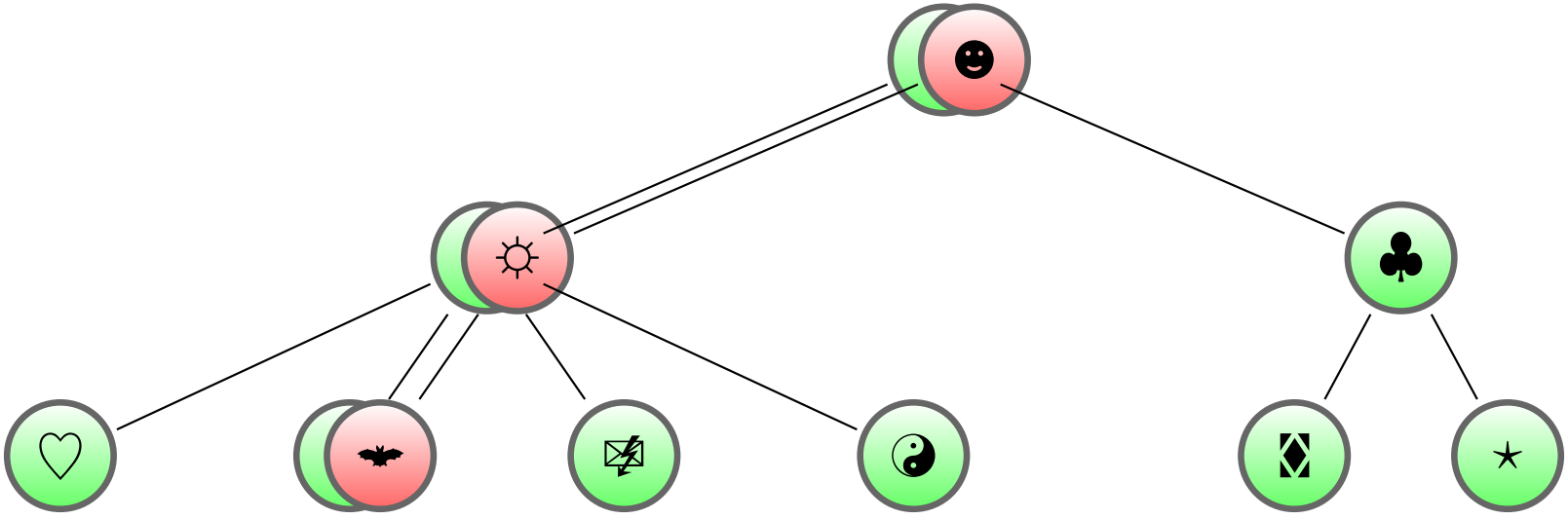
SB-IT



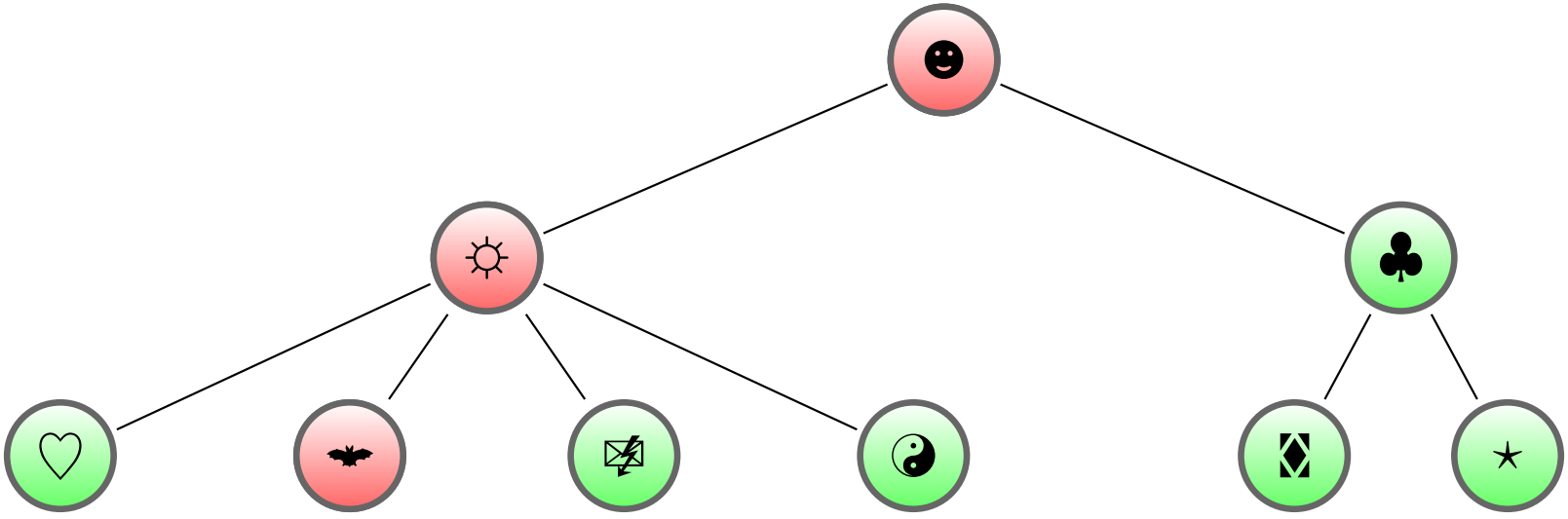
checkpoints - comment ça marche



checkpoints - comment ça marche

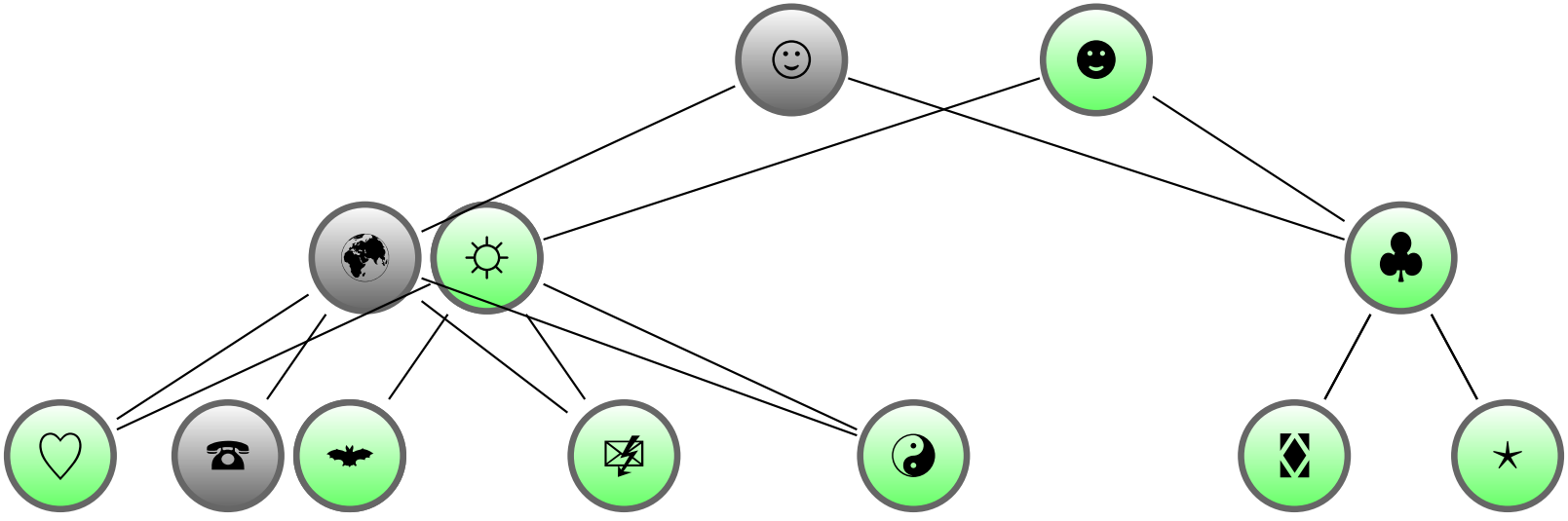


checkpoints - comment ça marche



checkpoints - comment ça marche

SB-IT



- snapshots (gratuits)
- clones (gratuits)

Quelques concepts

- Le block : séquence de bytes ou bits
- Block-size : taille de cette séquence
- niveau d'abstaction entre le hardware et les datas
- cluster : groupe de blocks
- Superblock (ou überblock) : block à la racine de tout, contient plus d'information que les blocks standard.
- inode : containeur de métadonnées, généralement lié à des blocks de data.

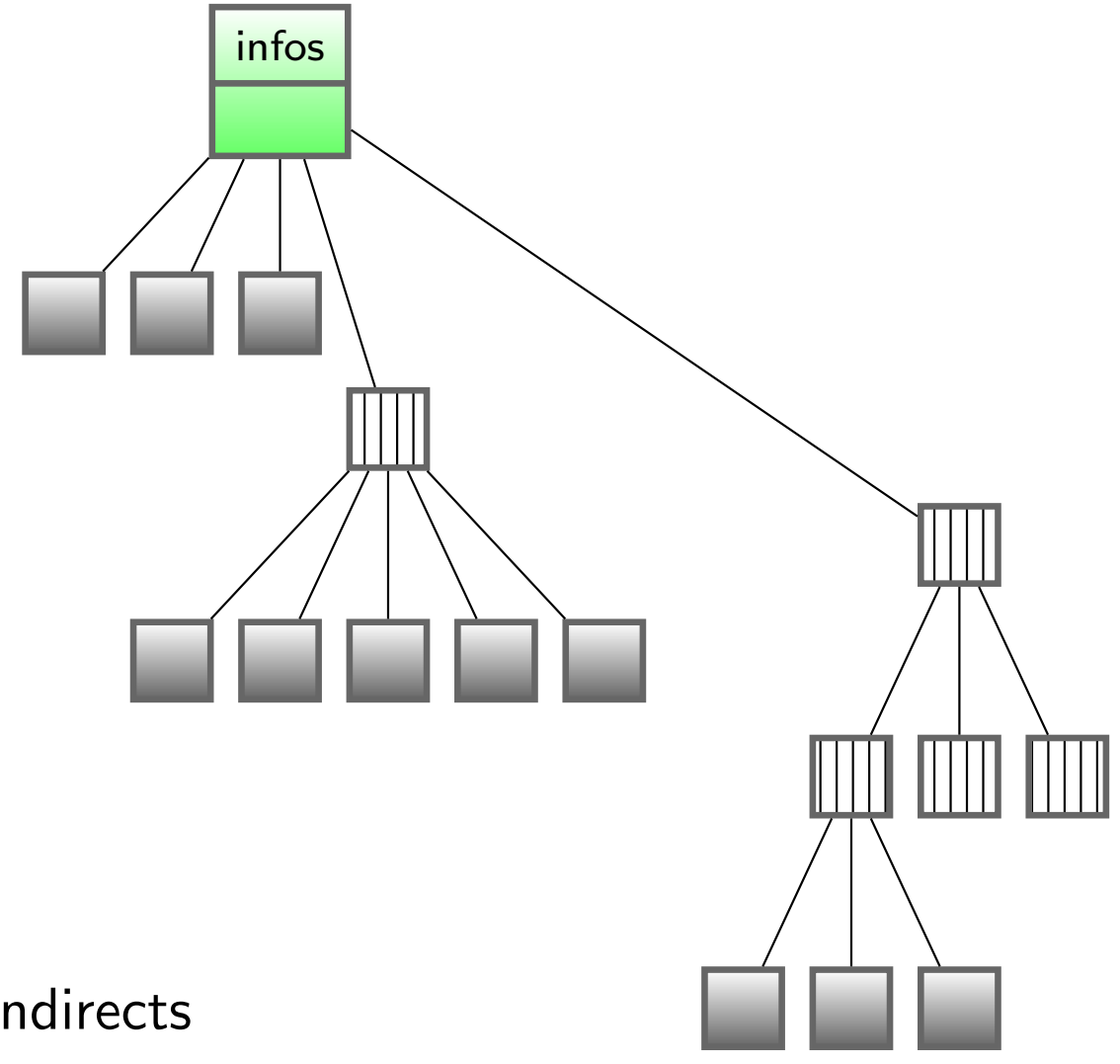
Quelques concepts

inode

blocks directs

blocks indirects

blocks doublement indirects



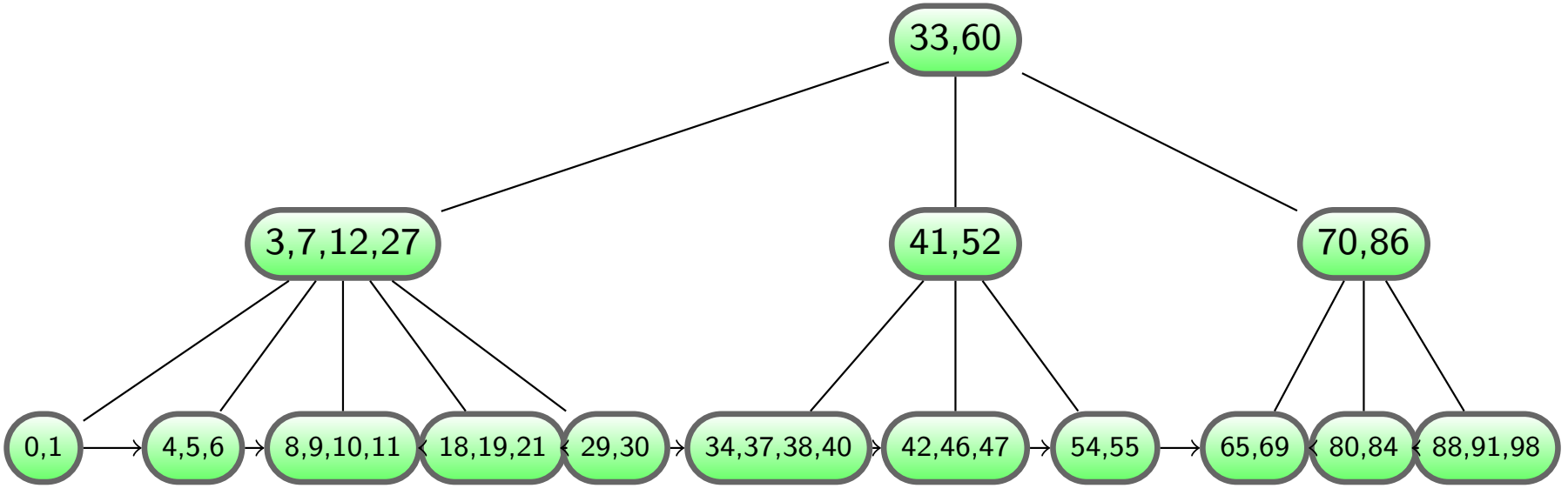
Types de blocks

- superblock
- inodes
- blocks avec uniquement datas
- blocks avec datas et pointeurs

Not to tree

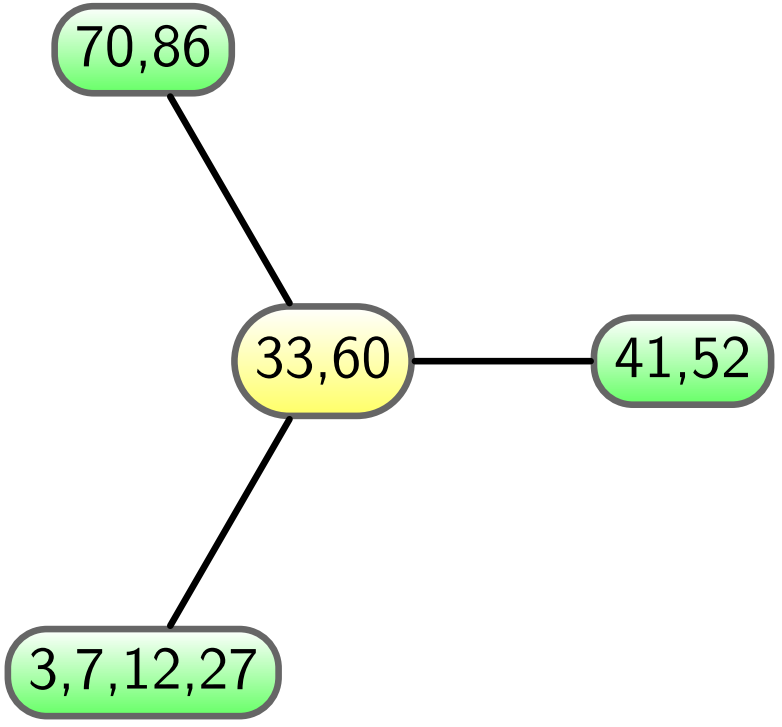
- le superblock se trouve physiquement après le boot record et ses données
- les inodes sont stockés dans un tableau
- les datas aussi, dans une autre zone du disque
- bitmaps pour les blocks des inodes et des datas
- tableau pour le nombre de références (clones, snapshots)

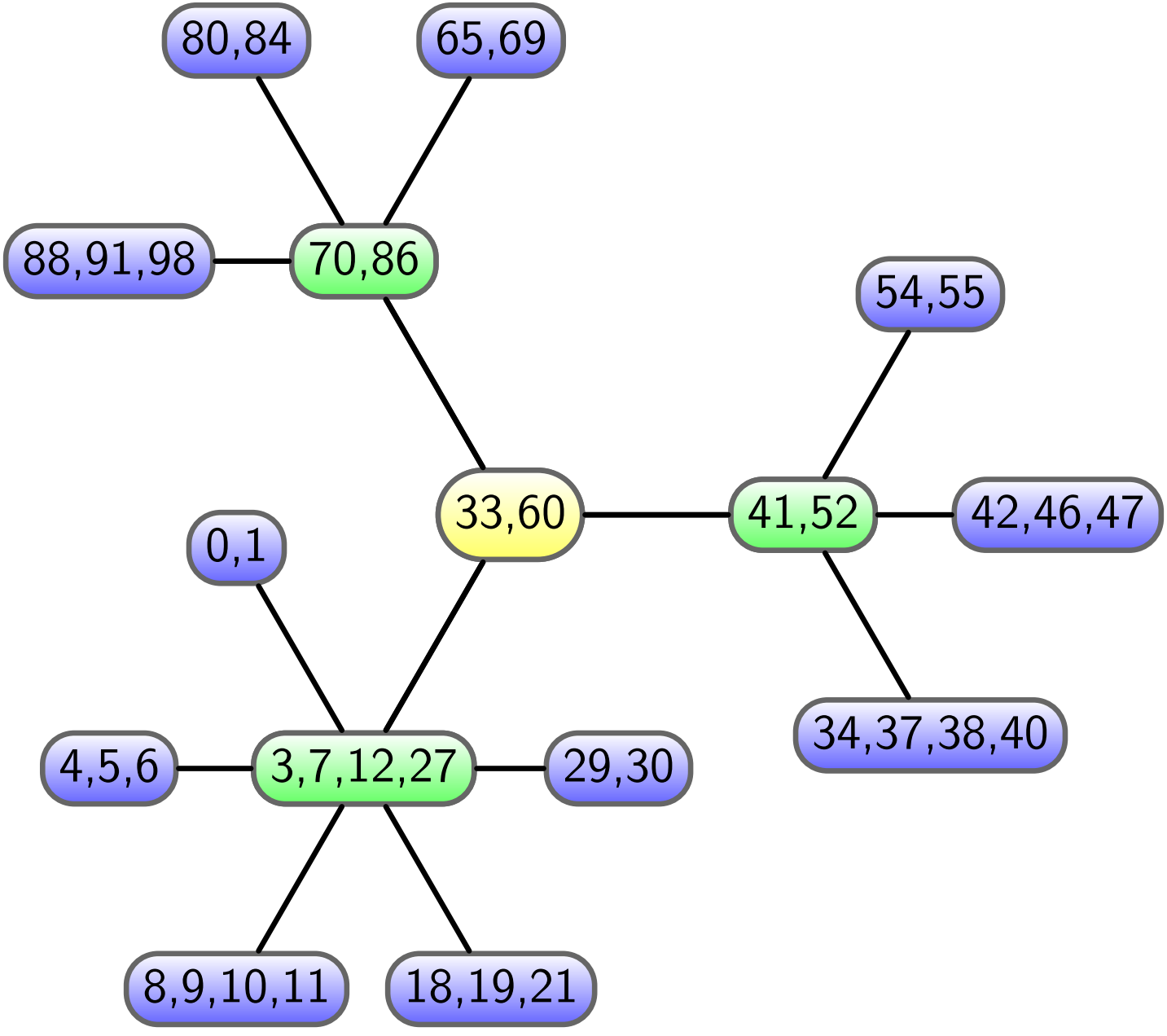
- *Balanced* Tree
- B-Tree d'ordre d : chaque noeud (sauf la racine) a entre d et $2d$ feuilles
- L-U B-Tree : chaque noeud a entre L et U feuilles.
- Clef de tri : index, p. ex ici **52**

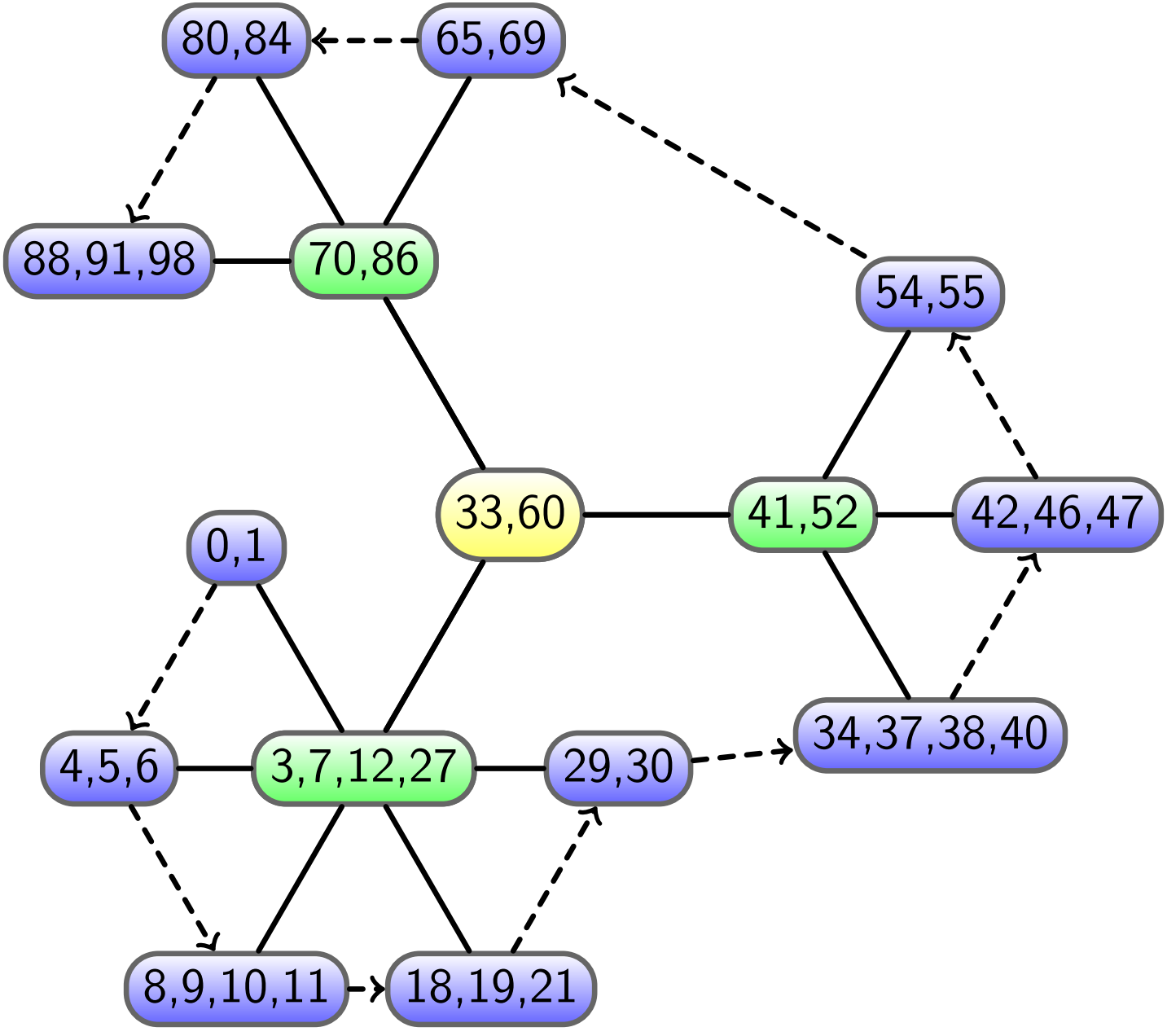


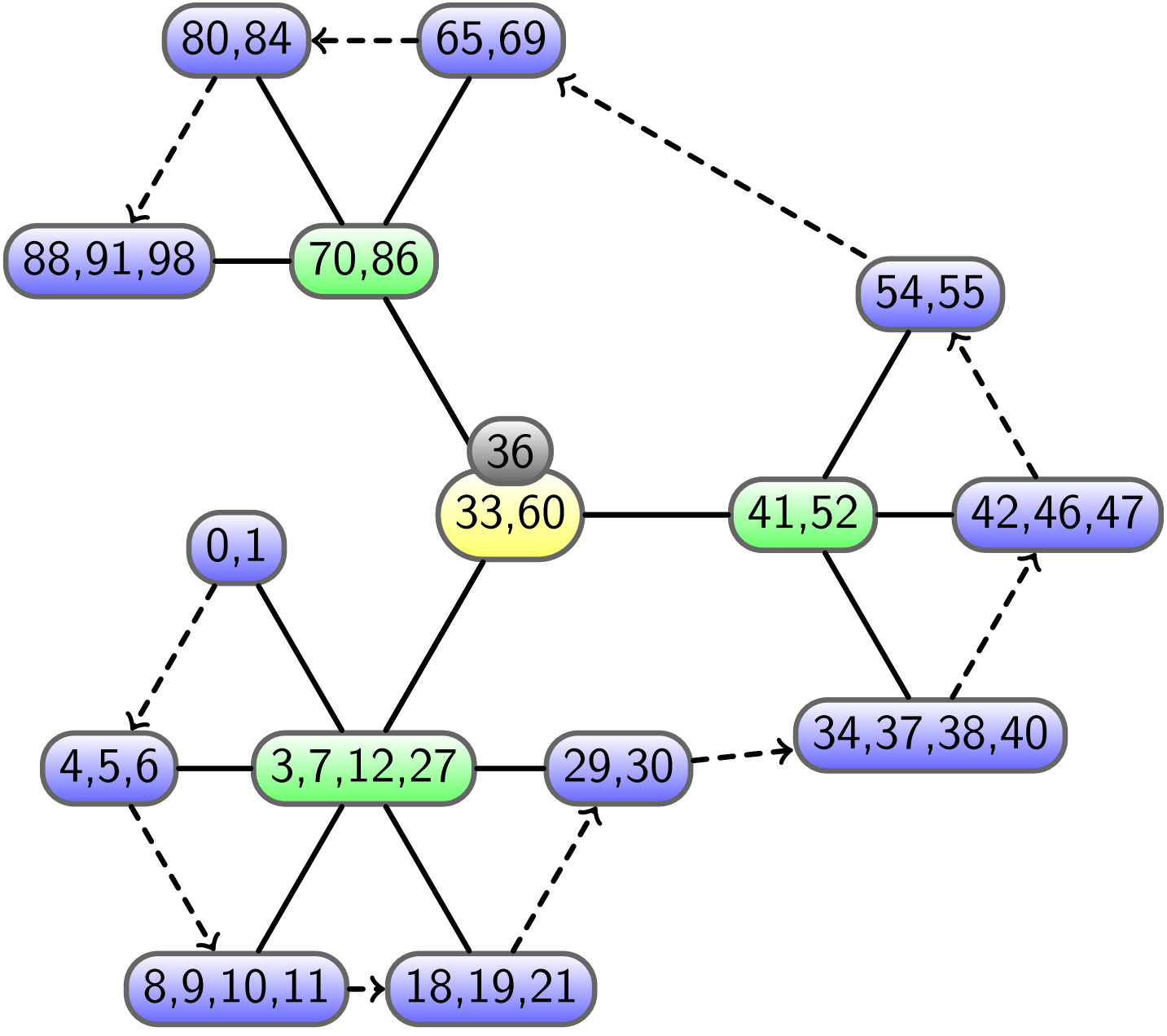
Exemples avec $n = 1M$ et $p = 2$

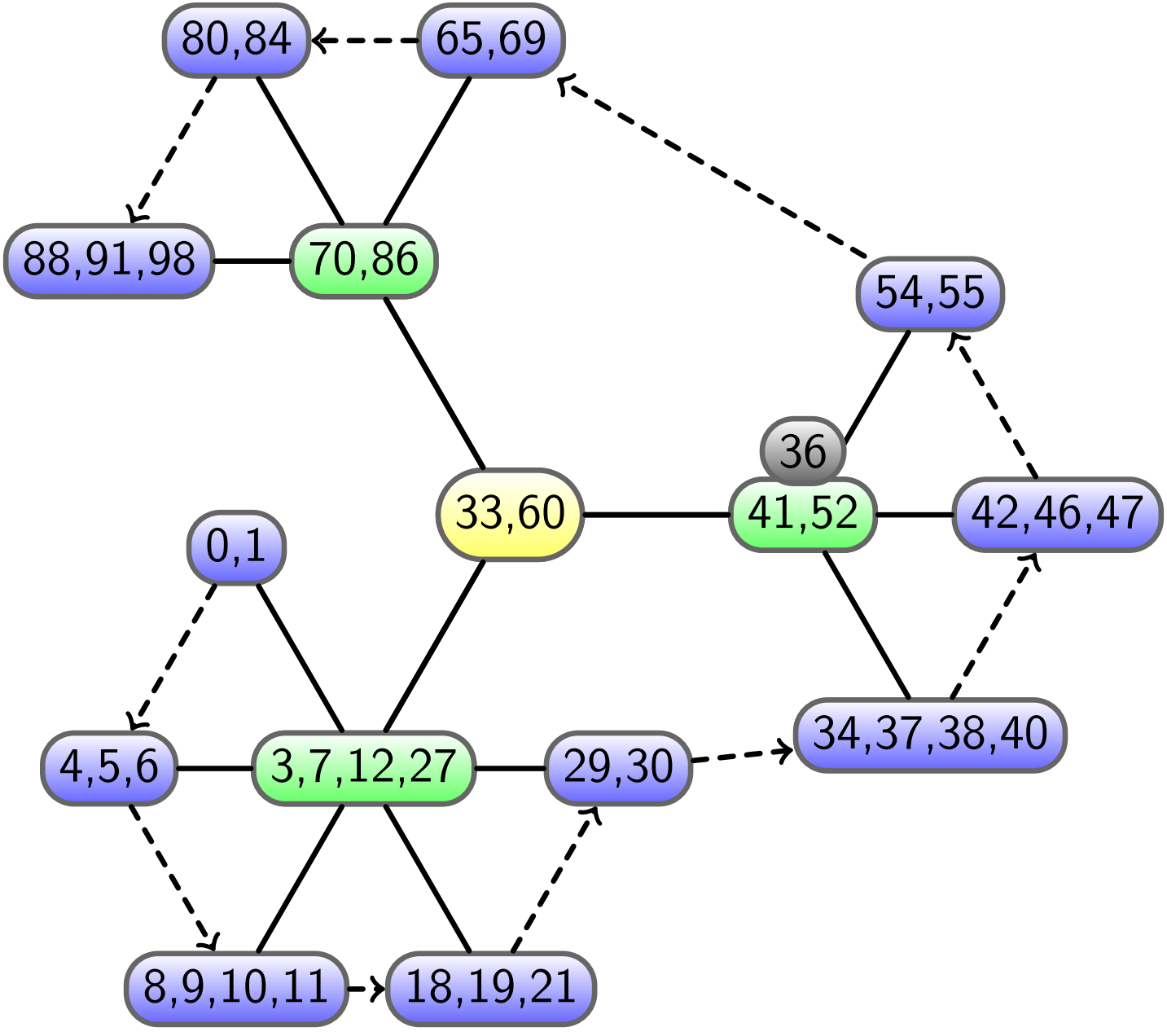
- Recherche $O(log_p(n))$ 19.93
- Insertion $O(plog_p(n))$ 20.93
- Remove $O(plog_p(n))$ 20.93
- n clefs peuvent être casées dans au plus $log_t \frac{n+1}{2}$ niveaux 18
- h niveaux peuvent contenir jusqu'à $2p \sum_{\ell=0}^{h} (2p+1)^{\ell}$ clefs 10^{13}

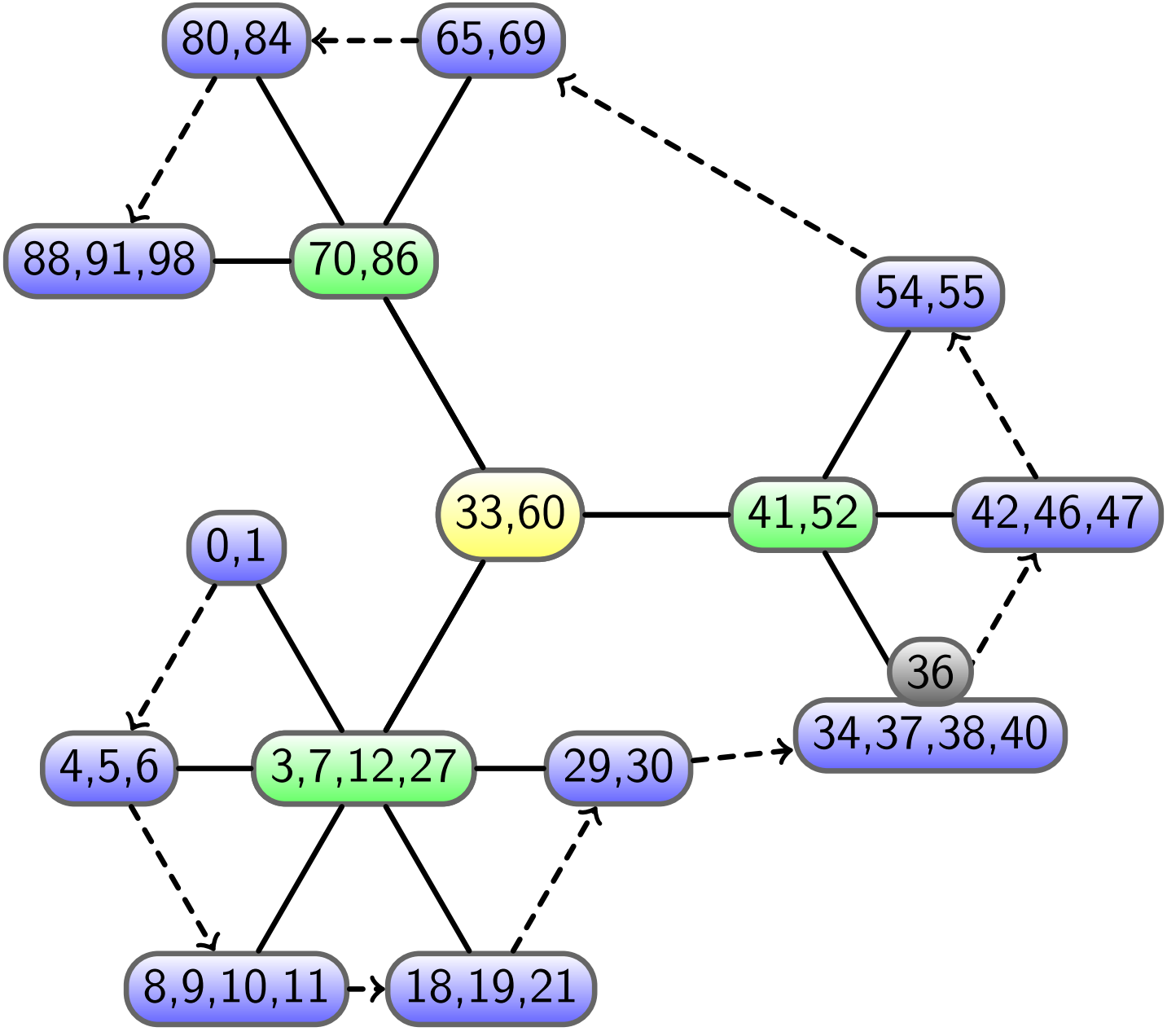

70,86
33,60
41,52
3,7,12,27

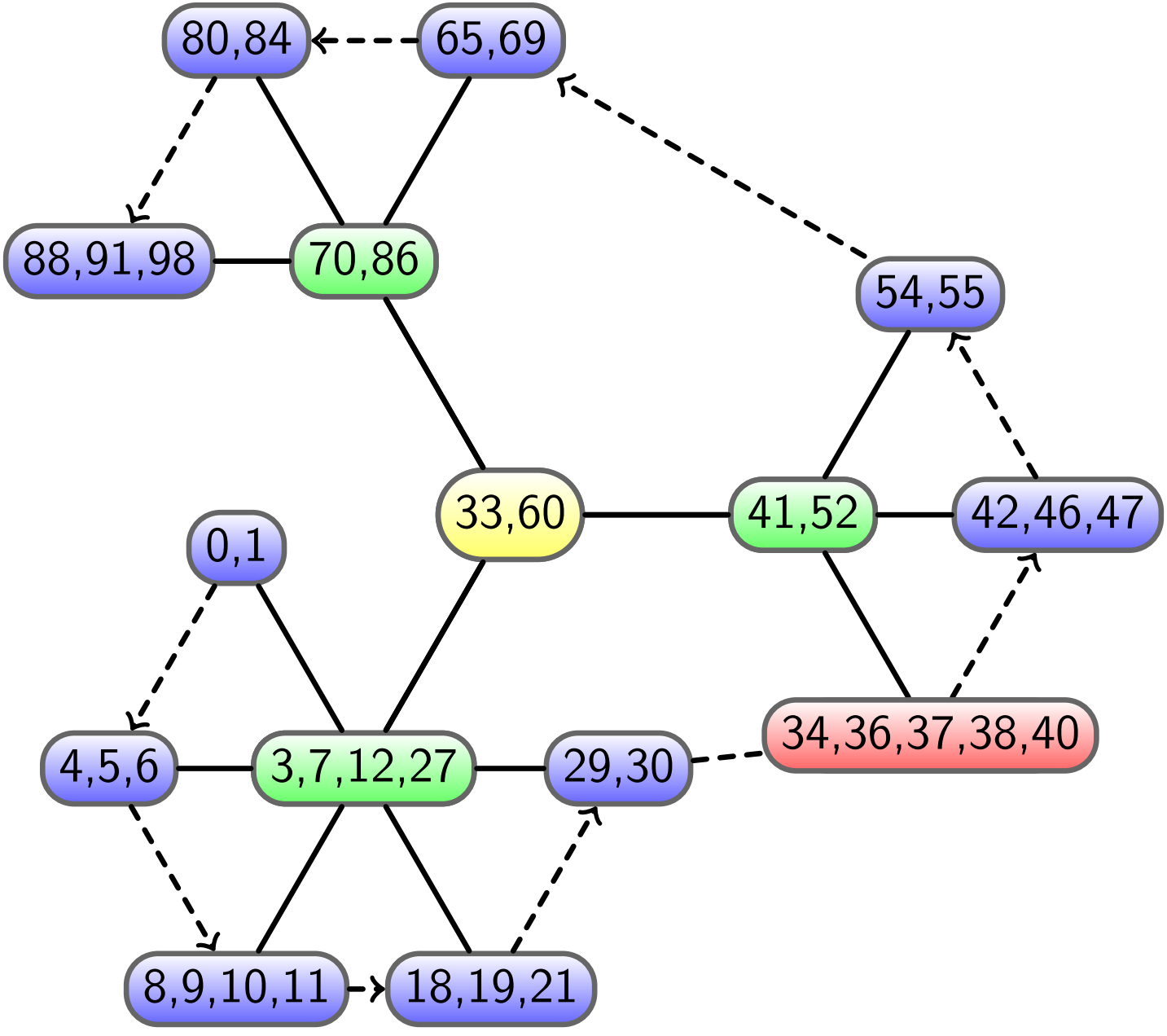

80,84
65,69
88,91,98
70,86
54,55
33,60
41,52
42,46,47
0,1
34,37,38,40
4,5,6
3,7,12,27
29,30
8,9,10,11
18,19,21

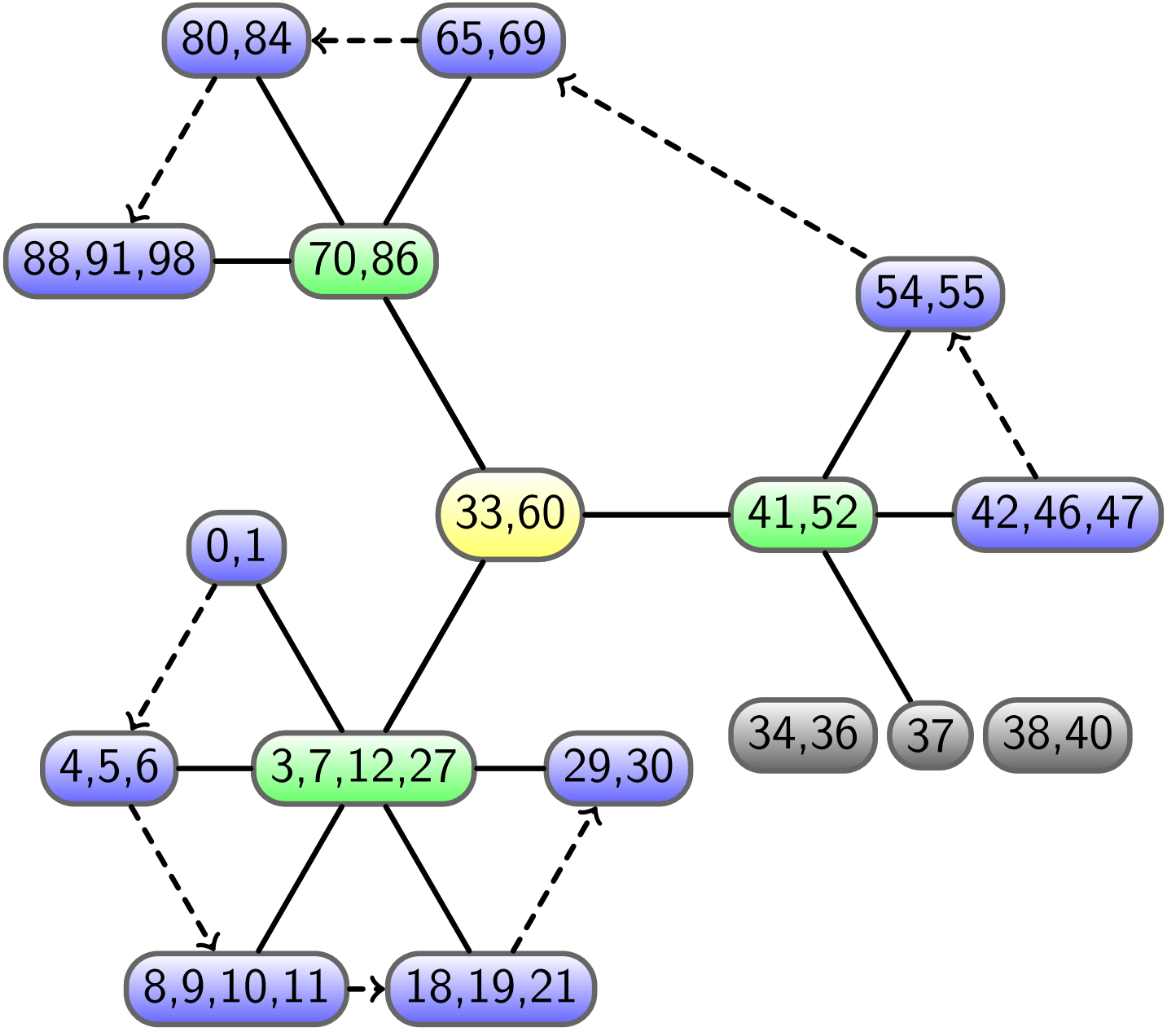

80,84
65,69
88,91,98
70,86
54,55
33,60
41,52
42,46,47
0,1
4,5,6
3,7,12,27
29,30
34,37,38,40
8,9,10,11
18,19,21

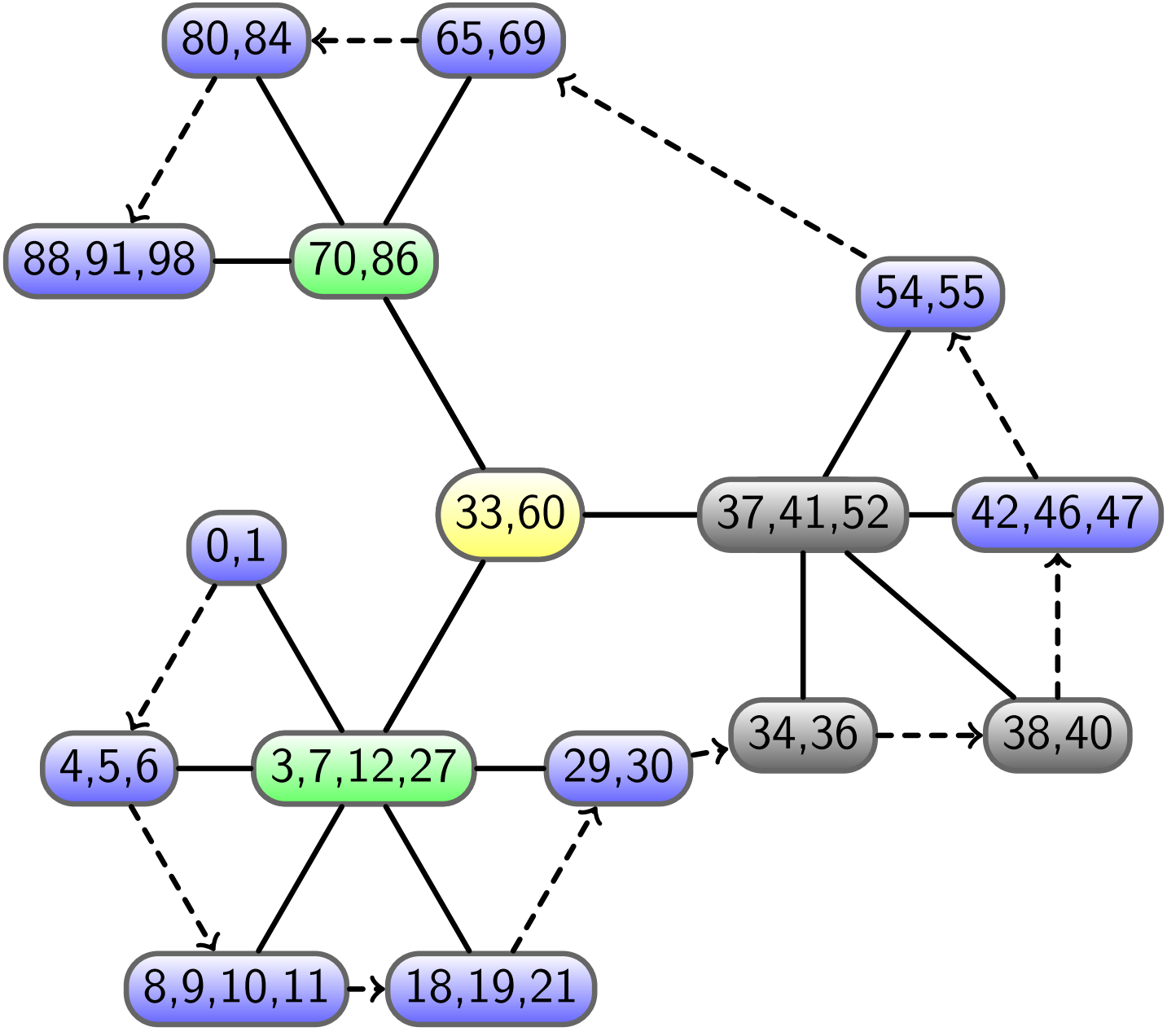

80,84
65,69
88,91,98
70,86
54,55
36
33,60
41,52
42,46,47
0,1
4,5,6
3,7,12,27
29,30
34,37,38,40
8,9,10,11
18,19,21

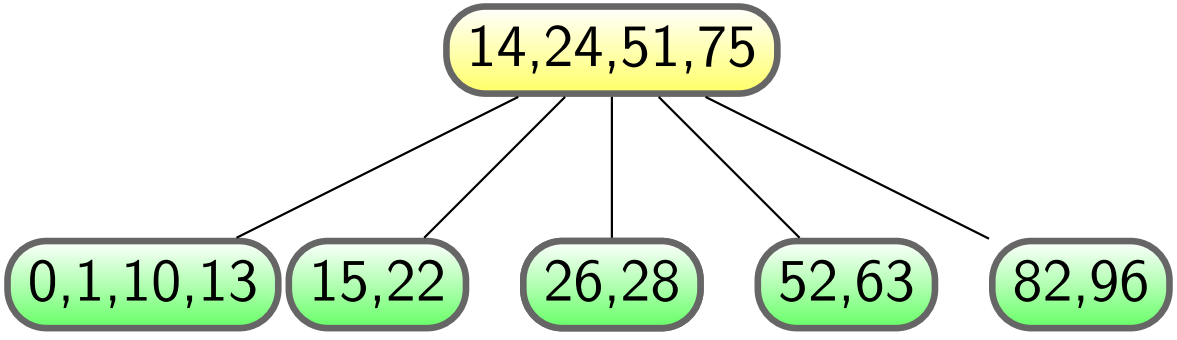

80,84
65,69
88,91,98
70,86
54,55
36
33,60
41,52
42,46,47
0,1
4,5,6
3,7,12,27
29,30
34,37,38,40
8,9,10,11
18,19,21

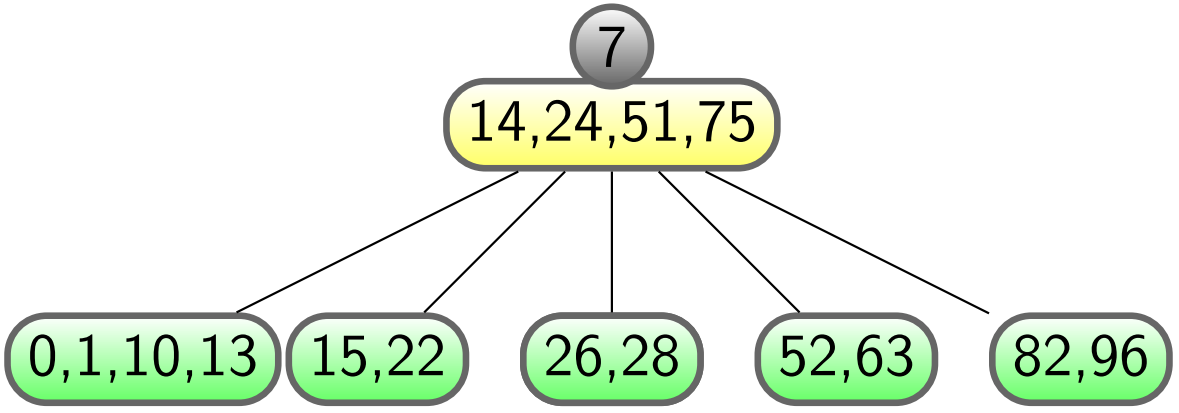

80,84
65,69
88,91,98
70,86
54,55
33,60
41,52
42,46,47
0,1
36
34,37,38,40
4,5,6
3,7,12,27
29,30
8,9,10,11
18,19,21

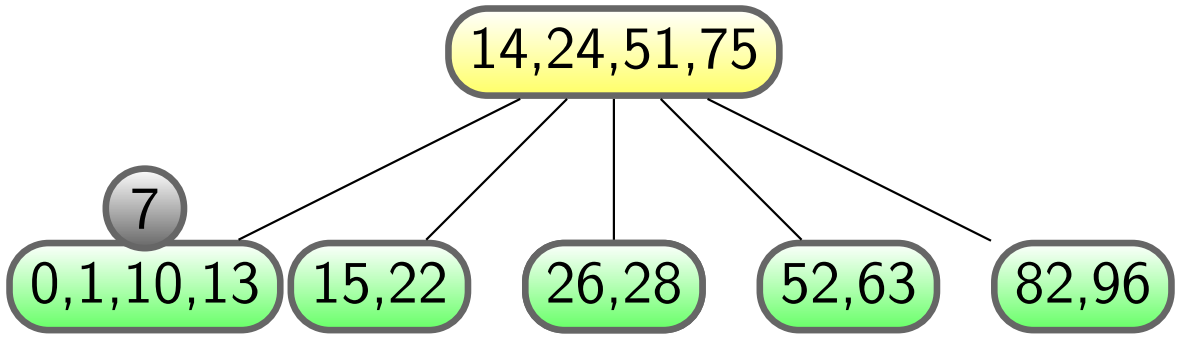

80,84
65,69
88,91,98
70,86
54,55
33,60
41,52
42,46,47
0,1
4,5,6
3,7,12,27
29,30
34,36,37,38,40
8,9,10,11
18,19,21

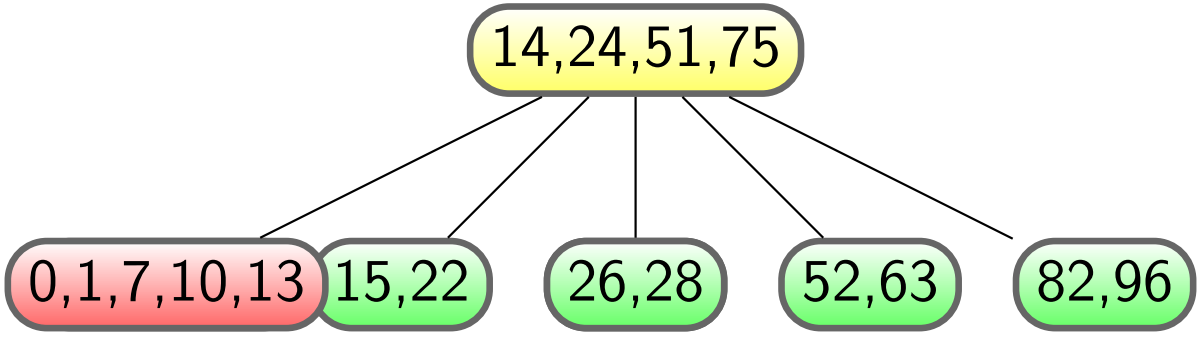

80,84
65,69
88,91,98
70,86
54,55
33,60
41,52
42,46,47
0,1
34,36
37
38,40
4,5,6
3,7,12,27
29,30
8,9,10,11
18,19,21

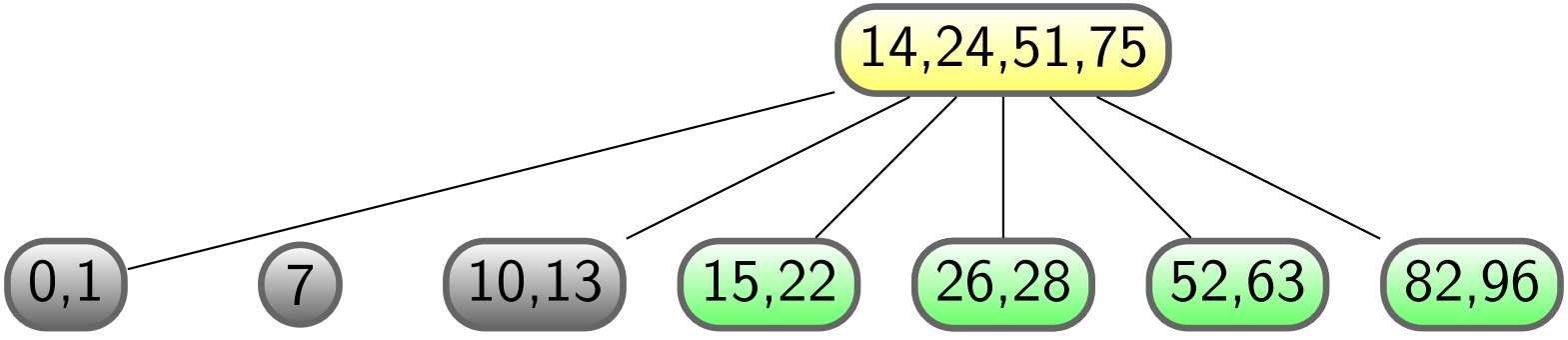

80,84
65,69
88,91,98
70,86
54,55
33,60
37,41,52
42,46,47
0,1
4,5,6
3,7,12,27
29,30
34,36
38,40
8,9,10,11
18,19,21

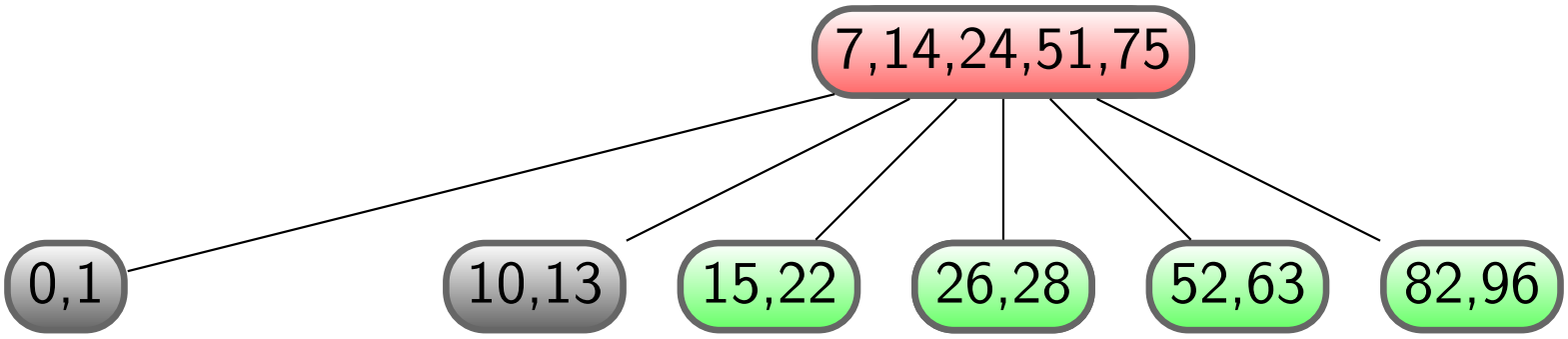

14,24,51,75
0,1,10,13
15,22
26,28
52,63
82,96

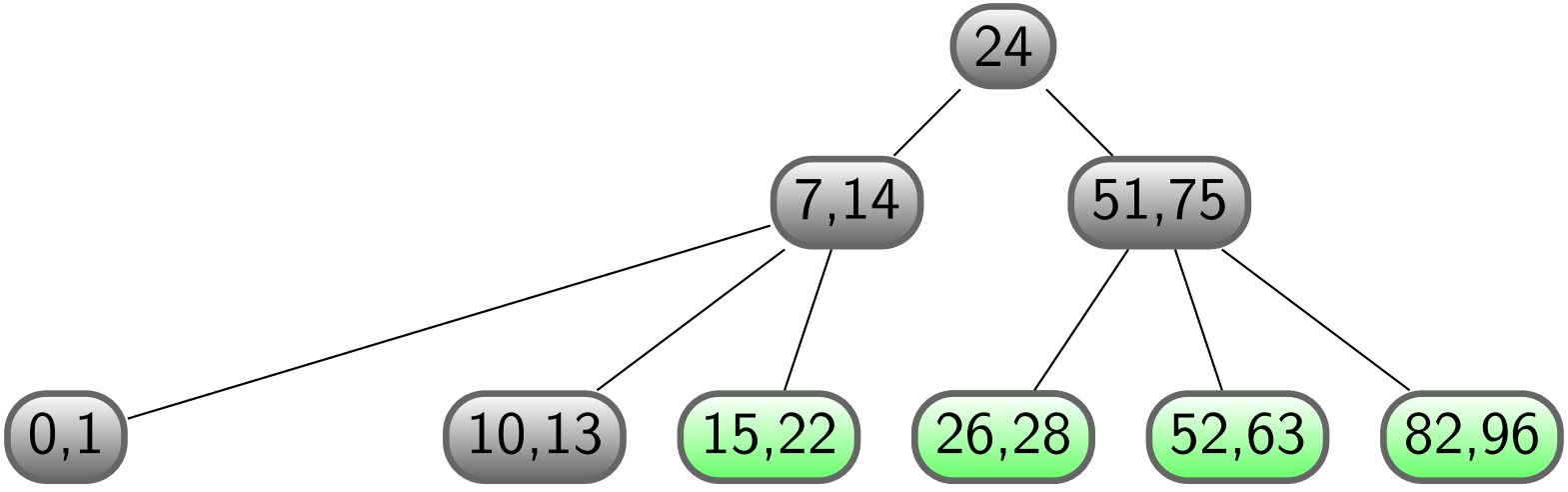

7
14,24,51,75
0,1,10,13
15,22
26,28
52,63
82,96


14,24,51,75
7
0,1,10,13
15,22
26,28
52,63
82,96


14,24,51,75
0,1,7,10,13
15,22
26,28
52,63
82,96


14,24,51,75
0,1
7
10,13
15,22
26,28
52,63
82,96

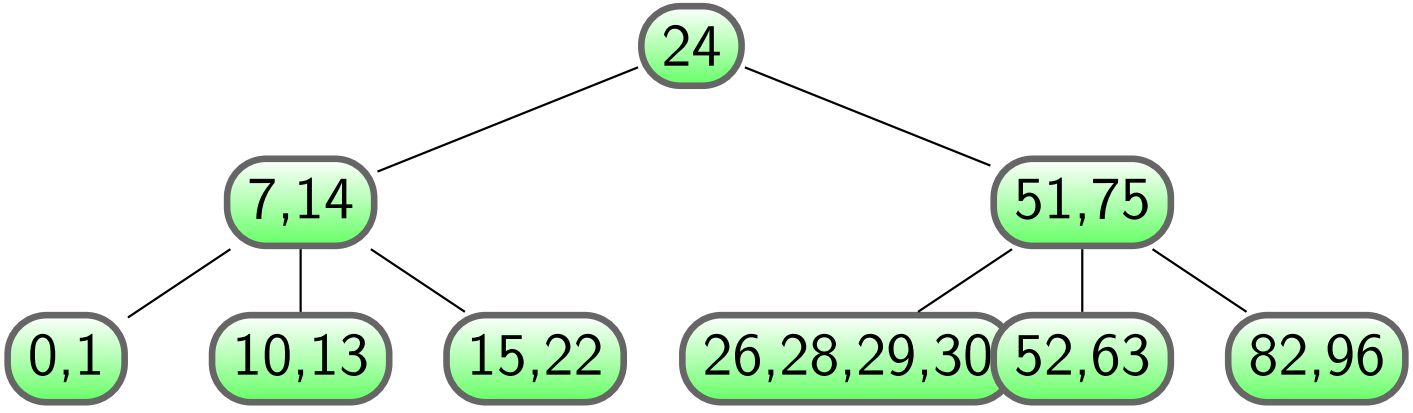

7,14,24,51,75
0,1
10,13
15,22
26,28
52,63
82,96

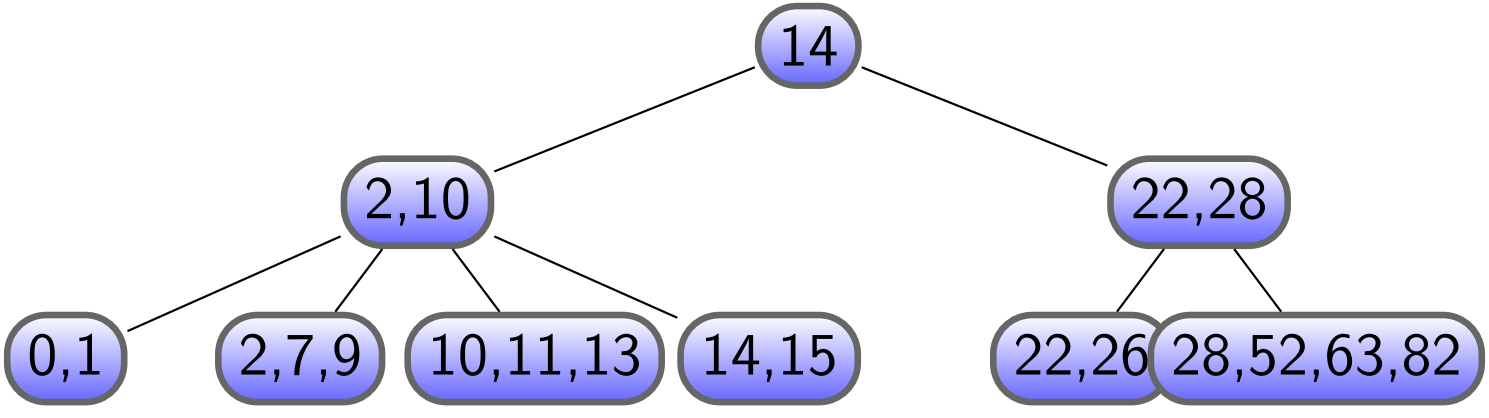

24
7,14
51,75
0,1
10,13
15,22
26,28
52,63
82,96

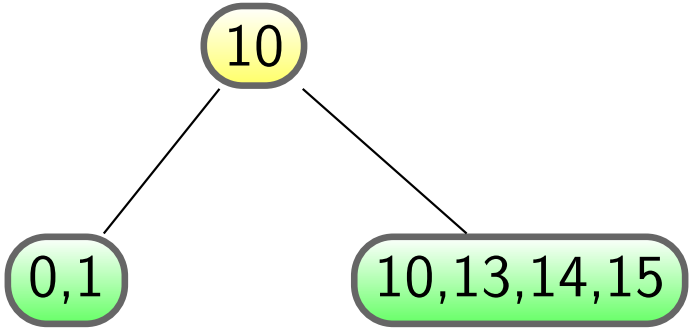
- Deux types de noeuds : les feuilles et les noeuds internes
- Les noeuds internes pour l'indexage
- Les feuilles contiennent les datas

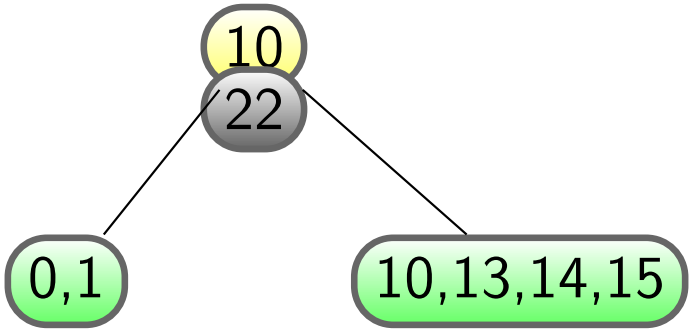
- Odre d'un B+tree = nb max de déscendants d'un noeud
- (B-tree d'ordre 2 $\approx$ B+-tree d'ordre 5)
- In B-+Tree d'ordre p a : au maximum $n = p^h$ entrées

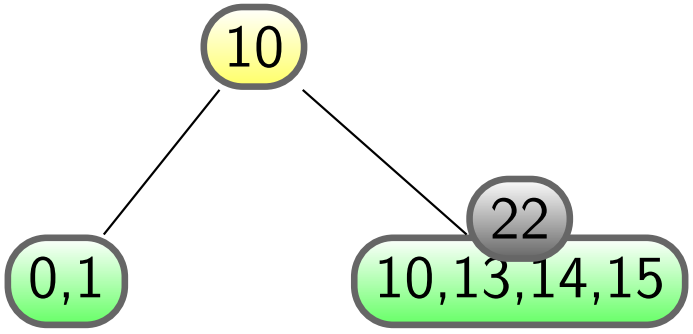
- Les feuilles sont souvent chaînées pour accélerer des requêtes multiples

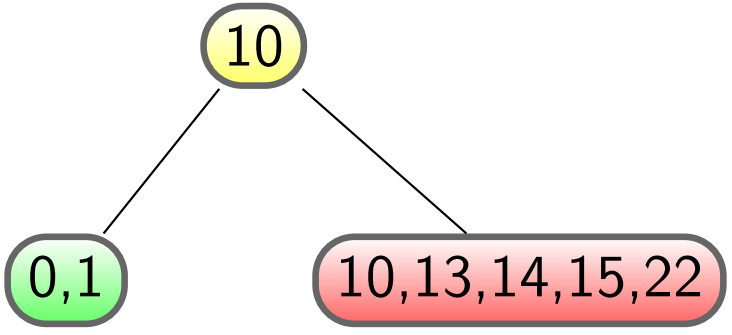

24
7,14
51,75
0,1
10,13
15,22
26,28,29,30
52,63
82,96

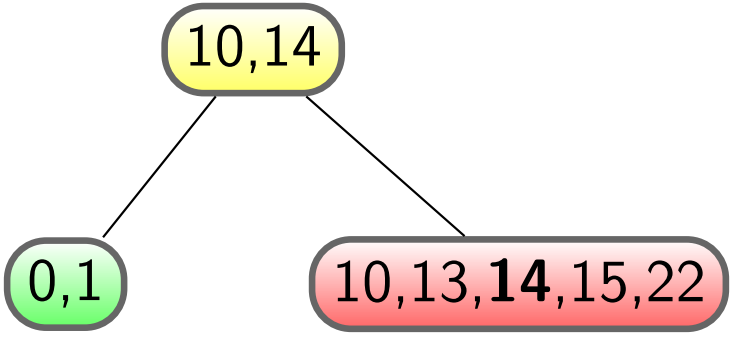

14
2,10
22,28
0,1
2,7,9
10,11,13
14,15
22,26
28,52,63,82

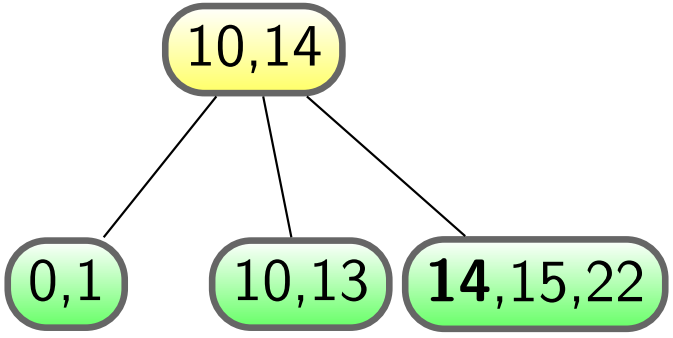

10
0,1
10,13,14,15

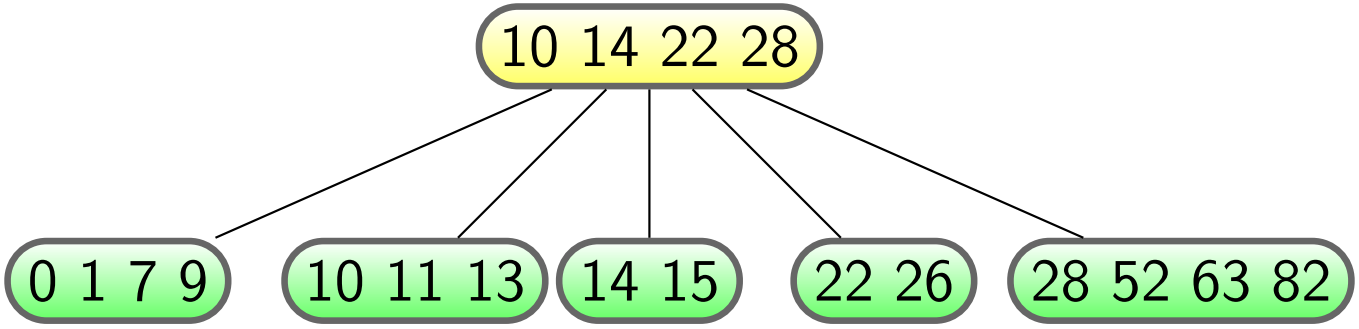

10
22
0,1
10,13,14,15

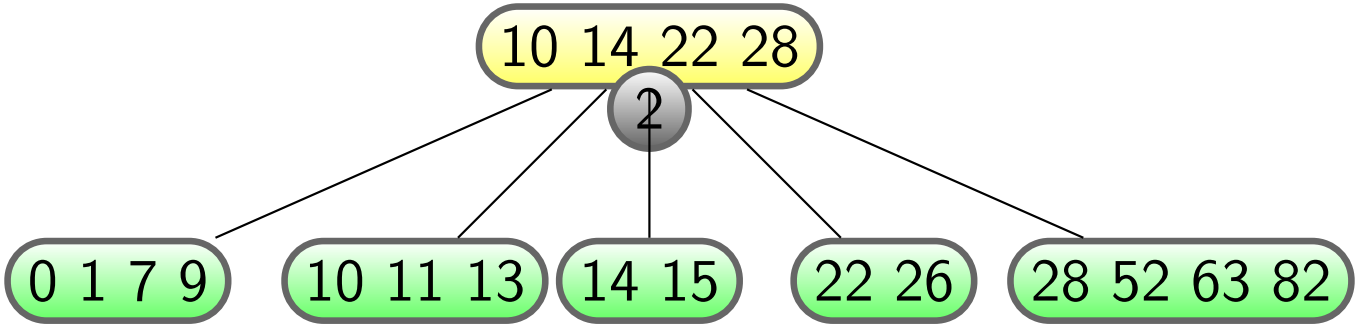

10
22
0,1
10,13,14,15

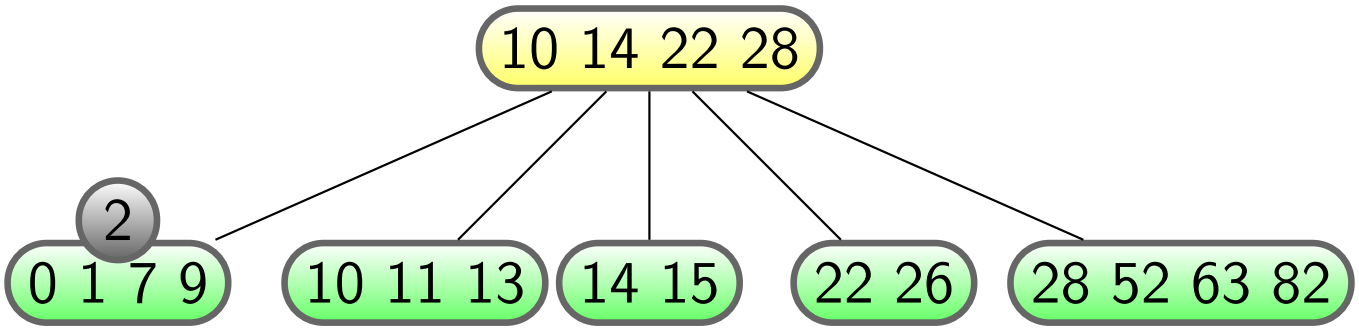

10
0,1
10,13,14,15,22

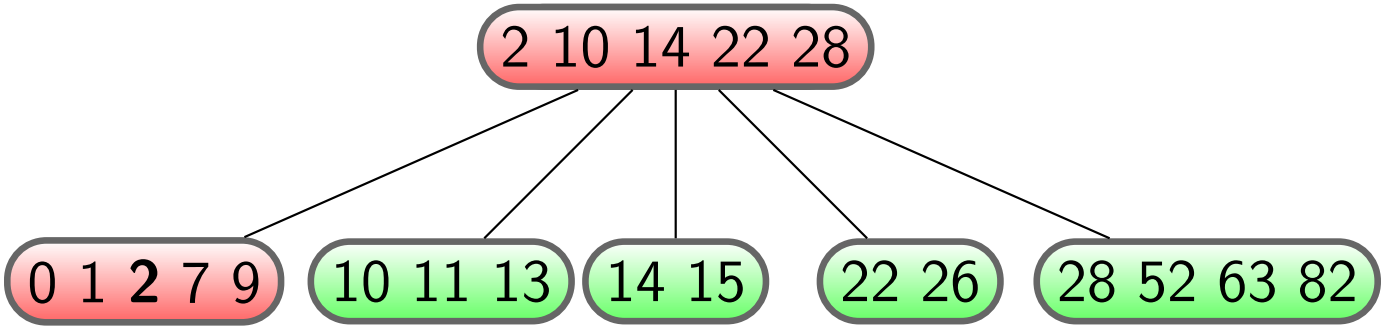

10,14
0,1
10,13,14,15,22

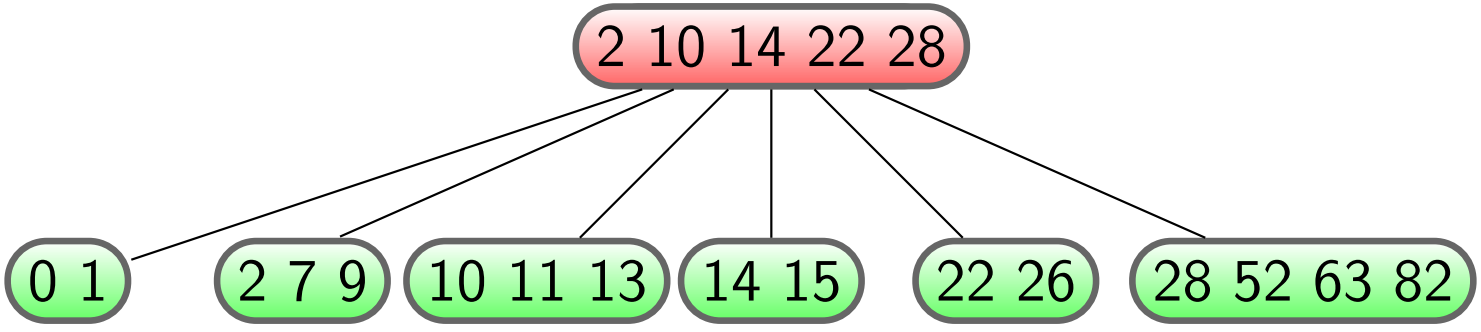

10,14
0,1
10,13
14,15,22

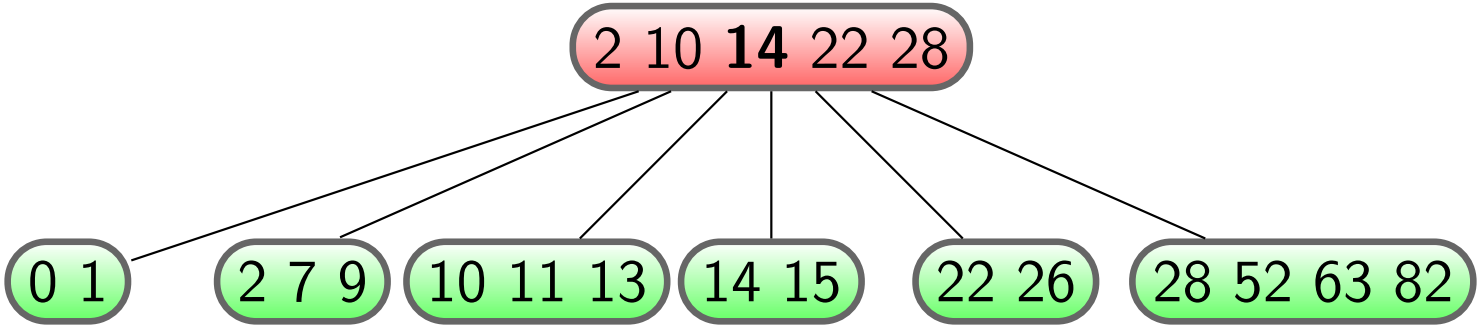

10 14 22 28
0 1 7 9
10 11 13
14 15
22 26
28 52 63 82

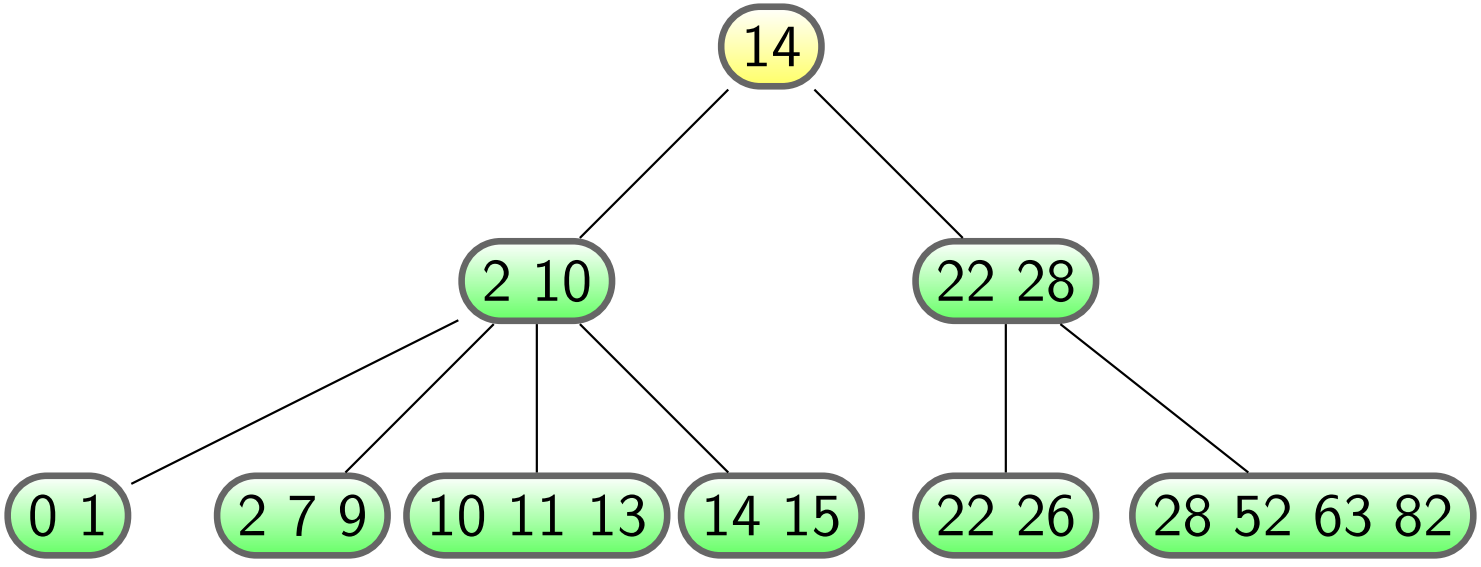

10 14 22 28
2
0 1 7 9
10 11 13
14 15
22 26
28 52 63 82


10 14 22 28
2
0 1 7 9
10 11 13
14 15
22 26
28 52 63 82

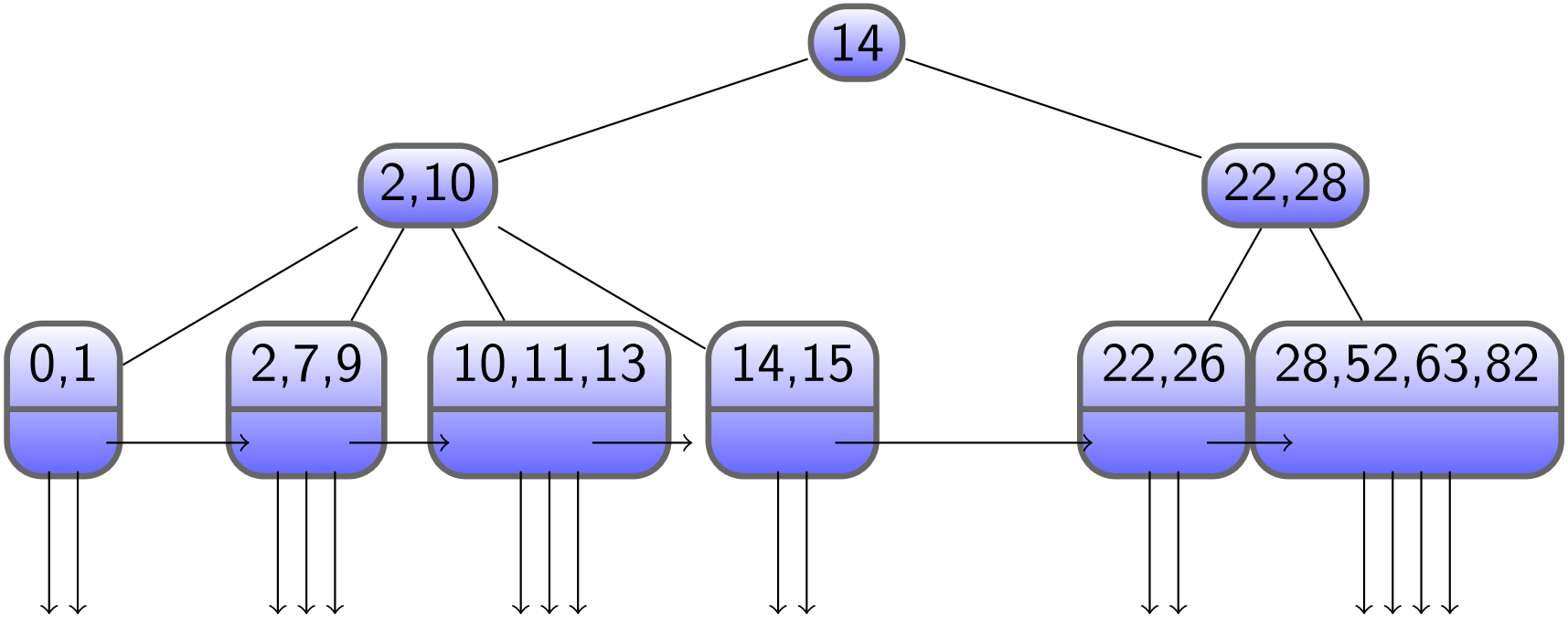

2 10 14 22 28
0 1 2 7 9
10 11 13
14 15
22 26
28 52 63 82


2 10 14 22 28
0 1
2 7 9
10 11 13
14 15
22 26
28 52 63 82


2 10 14 22 28
0 1
2 7 9
10 11 13
14 15
22 26
28 52 63 82


14
2 10
22 28
0 1
2 7 9
10 11 13
14 15
22 26
28 52 63 82

- les noeuds internes pointent vers leur noeuds fils
- les feuilles pointent vers les datas
- et vers la feuille successive



To tree

To tree

Liste exhaustive des fs qui utilisent les arbres balancés

Fiabilité et performances

Not to tree

- les checkpoints sont incompatibles avec les arbres comme on les connaît

Not to tree

- les checkpoints sont incompatibles avec les arbres comme on les connaît
- il faut re-écrire (la moitié de) la théorie sur les B-Trees.

Not to tree

- les checkpoints sont incompatibles avec les arbres comme on les connaît
- il faut re-écrire (la moitié de) la théorie sur les B-Trees.
- ça a été fait **en 2007**

Not to tree

- principale incompatibilité entre arbres et checkpoints : le fait que les feuilles sont enchainées
  $\Rightarrow$ on enlève ces liens, en renonçant entre autre à l'accès rapide à des blocs contigus
- comment compter les références ?

Questions ?

Trolls ?

Pour en savoir plus...

- Dave Hitz et al. : File System Design for an NFS File Server Appliance
- Zachary N. J. Peterson : Ext3cow : A Time-Shifting File System for Regulatory Compliance
- Jeff Bonwick et al. : The Zettabyte File System
- Ohad Rodeh : B-trees, Shadowing, and Clones