STOCKAGE SAN BLOCK
SNAPSHOTS ET THIN PROVISIONING

C:YLLENE

OPÉRATEUR DE
TRANSFORMATION DIGITALE

Alexandre Derumier

alexandre.derumier@groupe-cyllene.com
https://twitter.com/aderumier

Stockages et Features

Description	Plugin type	Level	Shared	Snapshots	Thinprovisioning	Trim	Backups	Replication	Chiffrement	Stable
Directory	dir	file	no	yes (.qcow2)	yes (.qcow2)	yes	yes	no	not yet (qcow2)	yes
NFS	nfs	file	yes	yes (.qcow2)	yes (.qcow2)	no (nfs 4.2?)	yes	no	not yet (qcow2)	yes
CIFS	cifs	file	yes	yes (.qcow2)	yes (.qcow2)	no	yes	no	not yet (qcow2)	yes
GlusterFS	glusterfs	file	yes	yes (.qcow2)	yes (.qcow2)	yes	yes	no	not yet (qcow2)	yes
LVM	lvm	block	yes	no	no	no	yes	no	no	yes
LVM-thin	lvmthin	block	no	yes	yes	yes	yes	no	no	yes
ZFS (local)	zfspool	block	no	yes	yes	yes	yes	yes	yes	yes
iSCSI/kernel	iscsi	block	yes	no	no	no	yes	no	no	yes
iSCSI/libiscsi	iscsidirect	block	yes	no	no	no	yes	no	no	yes
Ceph/RBD	rbd	block	yes	yes	yes	yes	yes	yes	yes	yes
ZFS over iSCSI	zfs	block	yes	yes	yes	yes	yes	yes	yes	yes
BTRFS	btrfs	file	no	yes	yes	yes	yes	not yet	no	technology preview

ZFS

- Performance : ok
- Avantages :

– replication via export|import snapshots

–Thin-provisioning

–Compression, Chiffrement

- Inconvénients:

– local only

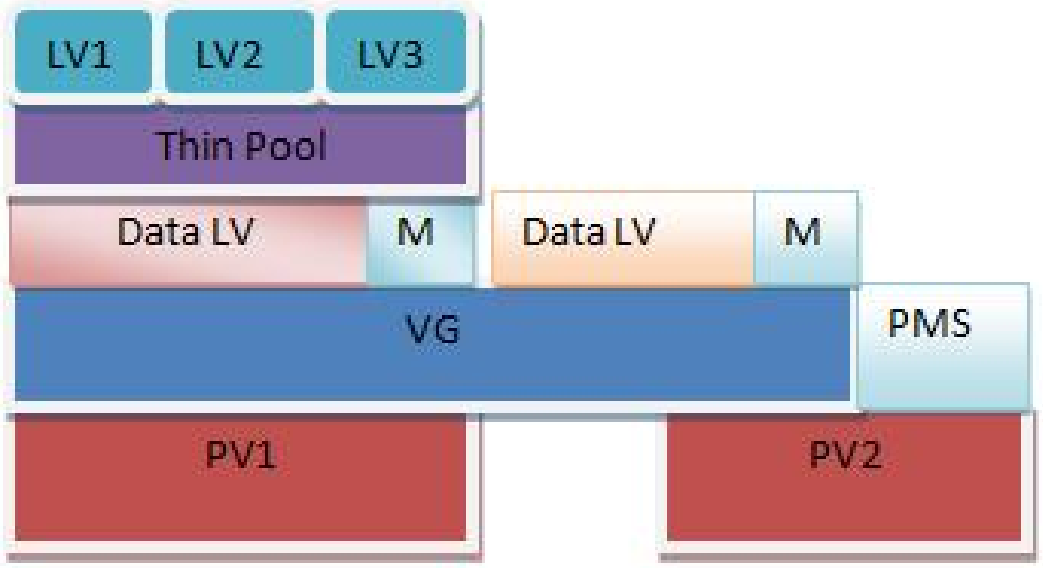
– rollback sur le dernier snapshot

– impossible de cloner un snapshot

–- memory hungry ^_^

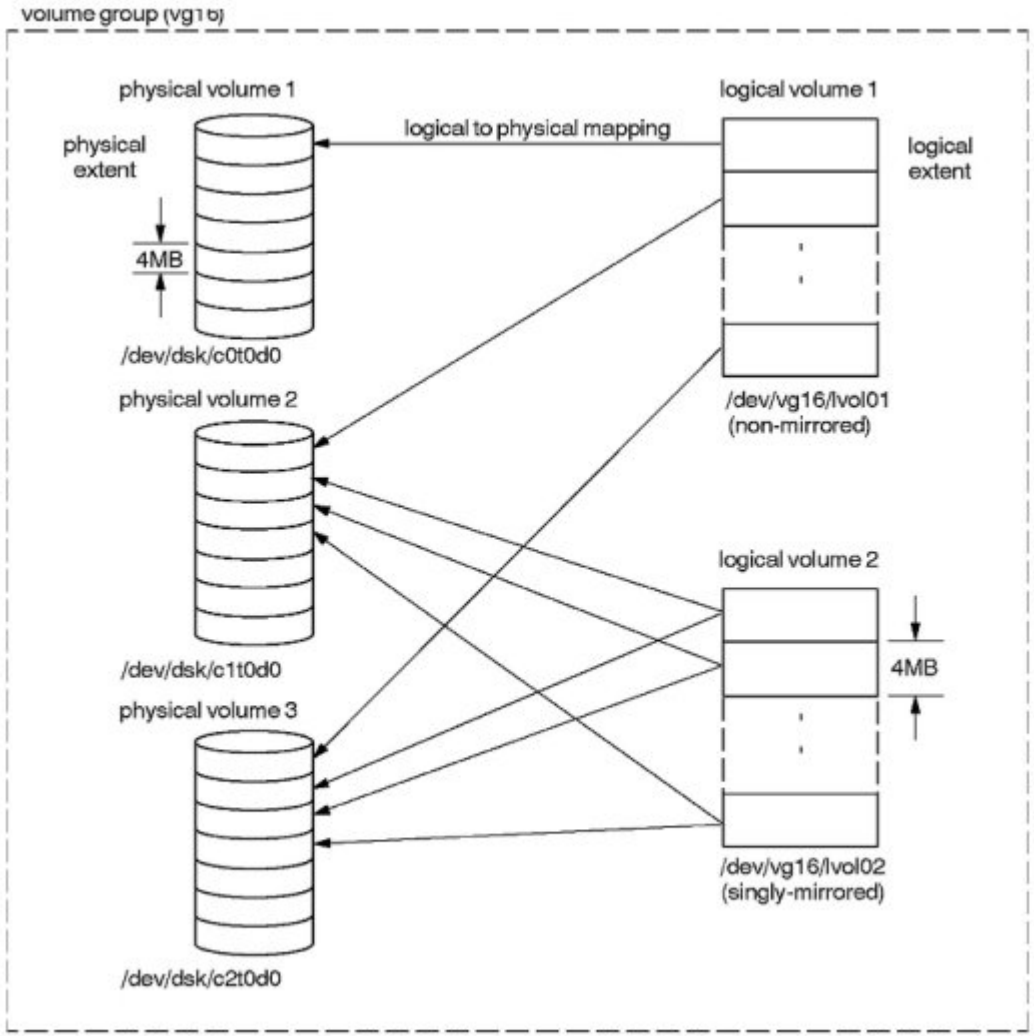
LVM-THIN

- Performance : ok
- Avantages :
  - Thin-provisioning
  - Snapshots (rollback && export ancien snasphot possible)
- Inconvénients:
  - local only ! Volume metadata non partageable.
  - pas de réplication
  - pas de compression, chiffrement



LVM-THICK

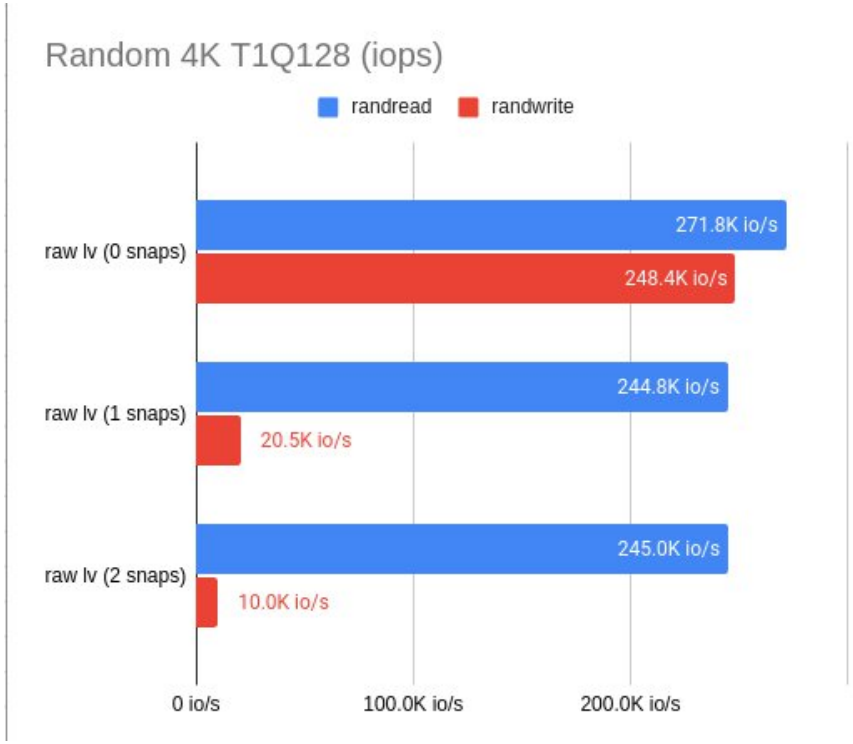
- Performance : ok
- Inconvénients:
  - local only
  - Pas de snapshot (Possible, mais lent)
  - provisioning
  - Pas de réplication
  - .....



LVM-THICK

WRITE AMPLIFICATION x 1000!

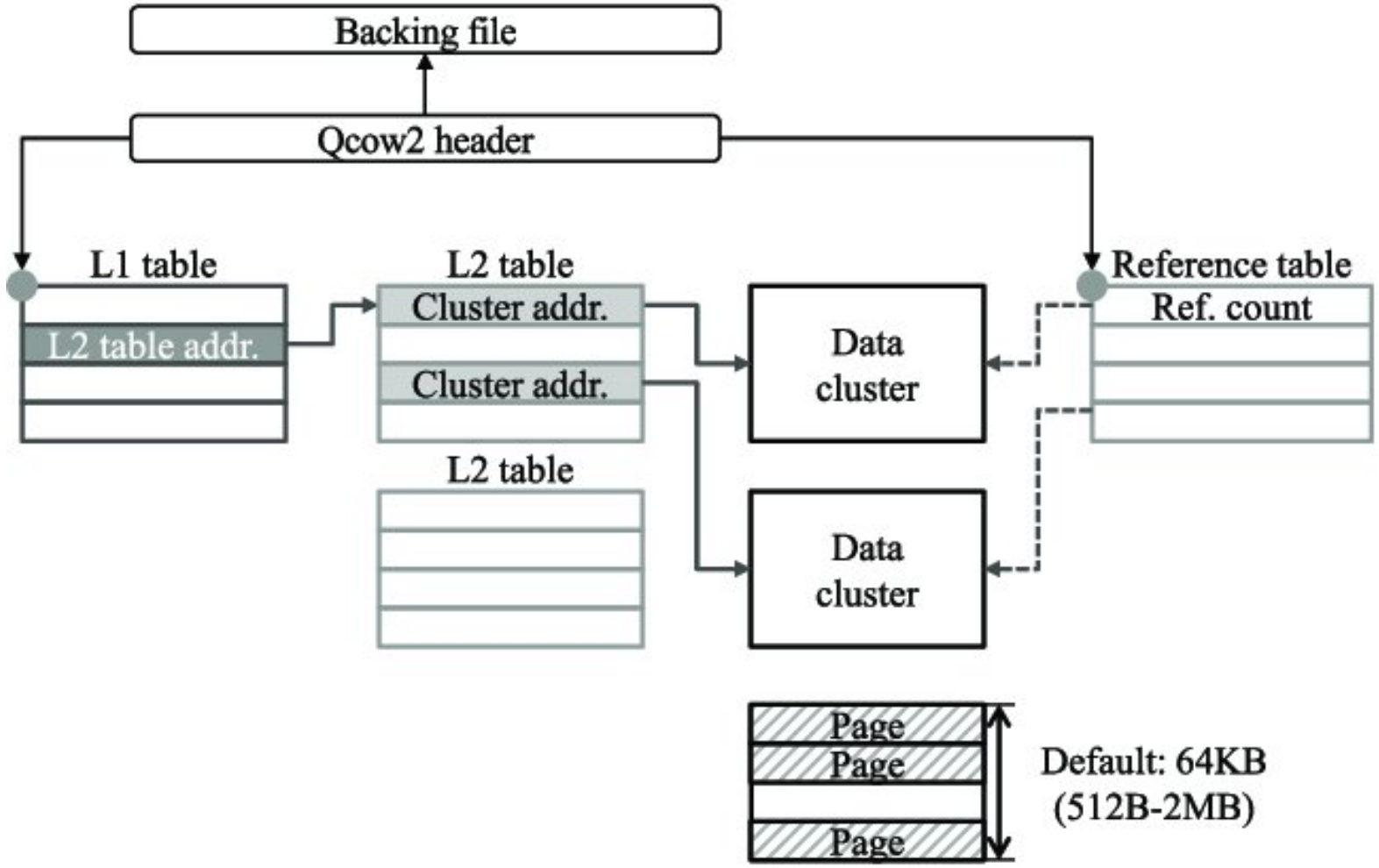
LVM-THICK



Fichiers .qcow2 (Internal snapshots)

- Performance : variable en fonction du workload
- Avantages :
  - Thin-provisioning
  - Snapshots (rollback && export ancien snasphot possible)
  - -1 fichier qcow2 avec l'ensemble des snasphots
- Inconvénients:
  - Filesystem only. (NFS/SMB pour stockage partagés)
  - pas de réplication
  - suppression snapshot : lock complet en écriture ! (internal snapshot)
  - performance allocation nouveau block après snapshot (internal snapshot)

Fichiers .qcow2 (Internal snapshots)



Fixing qcow2 (Internal snapshots)

- Inconvénients:

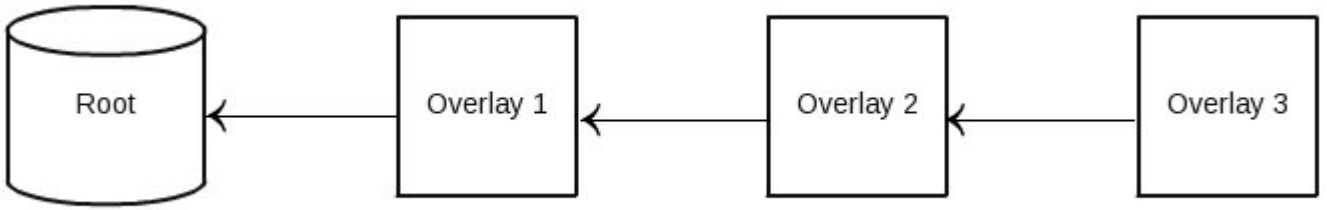
– pas de réplication

– suppression snapshot : lock complet en écriture ! (internal snapshot)

– Performance : variable

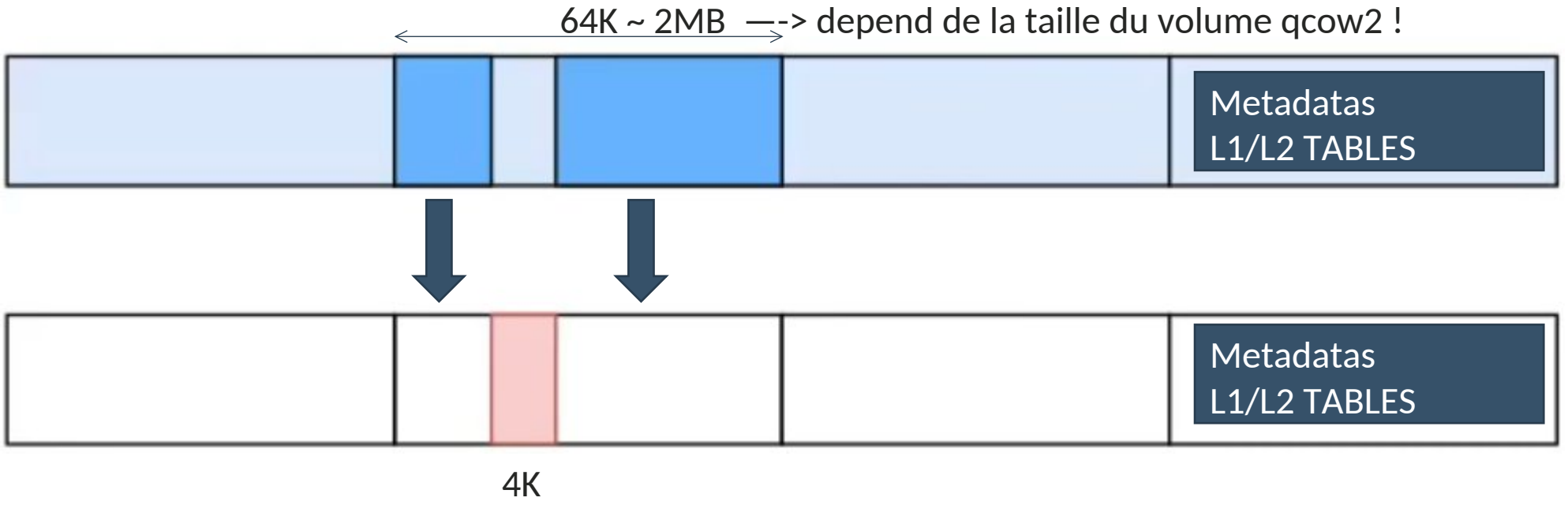
——> EXTERNAL SNAPSHOTS !

qcow2 (External snapshots)



- 1 fichier par snapshot

Performance



Disk size	Cluster size	L2 cache	QEMU master	4K slices
16 GB	64 KB	1 MB [8 GB]	5000 IOPS	12700 IOPS
2 TB	2 MB	4 MB [1 TB]	576 IOPS	11000 IOPS

Performance

- QEMU keeps a cache of L2 tables to speed up disk access.
- The maximum amount of L2 metadata depends on the disk size and the cluster size.
- Problem: large images need large amounts of metadata, so we cannot keep everything in memory.

Cluster size (=L2 table size)	Max. L2 size per TB
64 KB	128 MB (2048 tables)
128 KB	64 MB (512 tables)
256 KB	32 MB (128 tables)
512 KB	16 MB (32 tables)
1 MB	8 MB (8 tables)
2 MB	4 MB (2 tables)

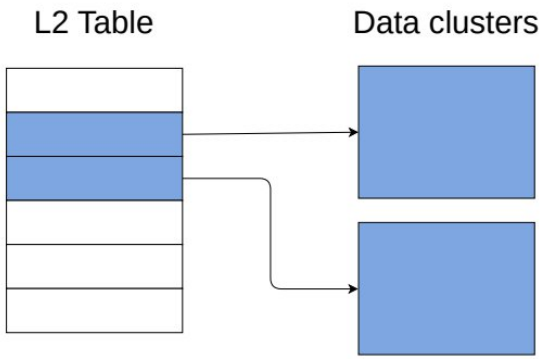
Qcow2 Extended L2 (sub allocation cluster)

$ qemu-img create -f qcow2 -o **extended_l2=on**,cluster_size=128k img.qcow2 1T

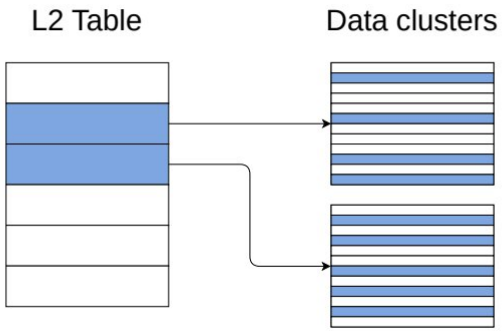
Existe depuis 2020 (qemu 5.2)

Pas implémenté actuellement sur proxmox !

A standard L2 table with entries and their data clusters



An extended L2 table with subcluster allocation



Qemu Extended L2 (sub allocation cluster)

- Having less copy-on-write improves the allocation performance.
- If subcluster size = request size no copy-on-write is needed!
- Average IOPS of random 4KB writes:

With a backing file		
Cluster size	**Without subclusters**	**With subclusters**
16 KB	3600 IOPS	8124 IOPS
32 KB	2557 IOPS	11575 IOPS
64 KB	1634 IOPS	13219 IOPS
128 KB	869 IOPS	12076 IOPS
256 KB	577 IOPS	9739 IOPS
512 KB	364 IOPS	4708 IOPS
1 MB	216 IOPS	2542 IOPS
2 MB	125 IOPS	1591 IOPS

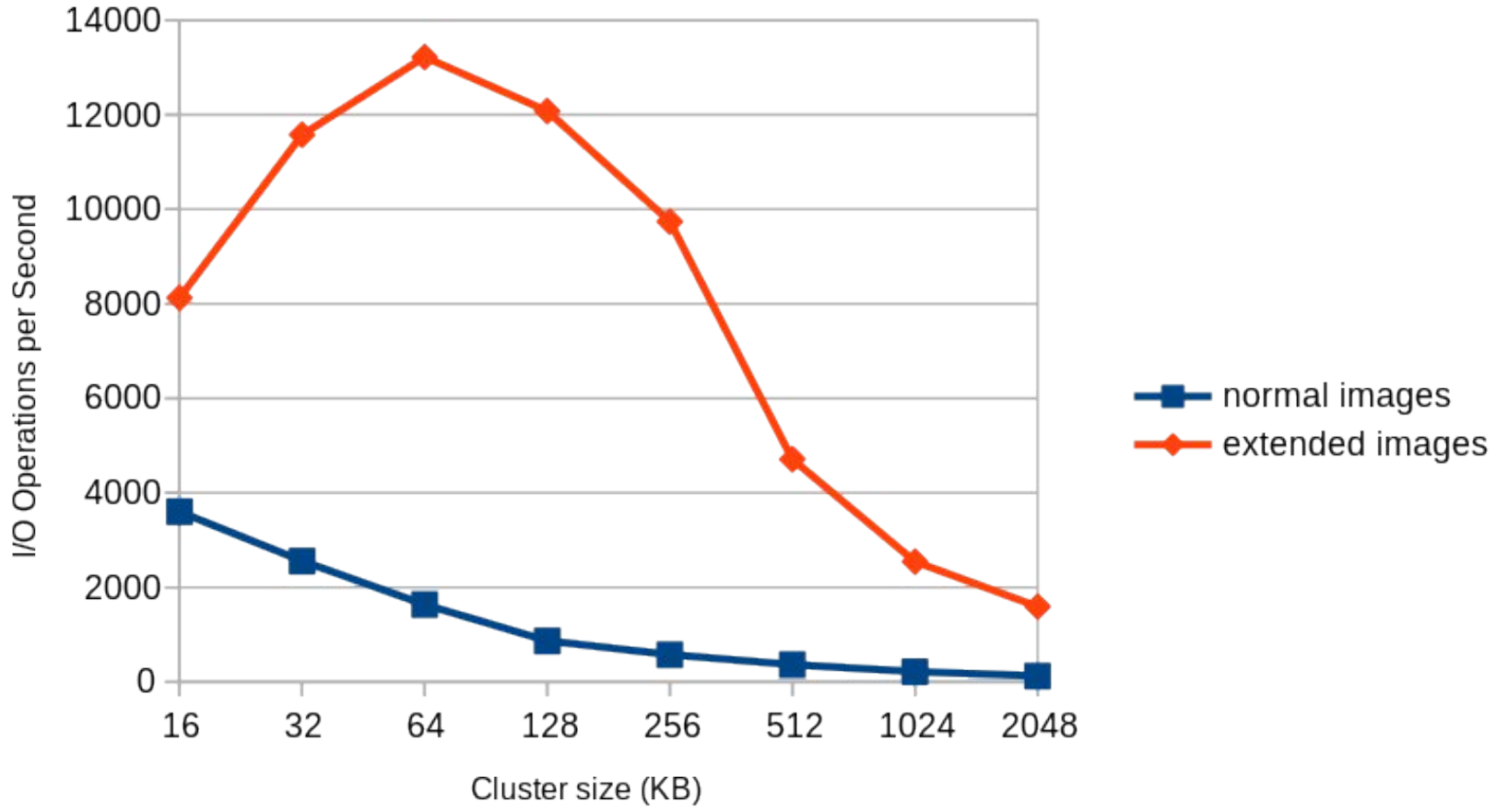
Qemu Extended L2 (sub allocation cluster)

- Extended L2 entries are twice as large but each one of them references 32 subclusters.
- As a result we have **16 times less** metadata for the same unit of allocation.
- This table compares the amount of L2 metadata for a 1TB image.

Standard L2 entries	
Cluster size	**Max. L2 size**
4 KB	2 GB
8 KB	1 GB
16 KB	512 MB
32 KB	256 MB
64 KB	128 MB

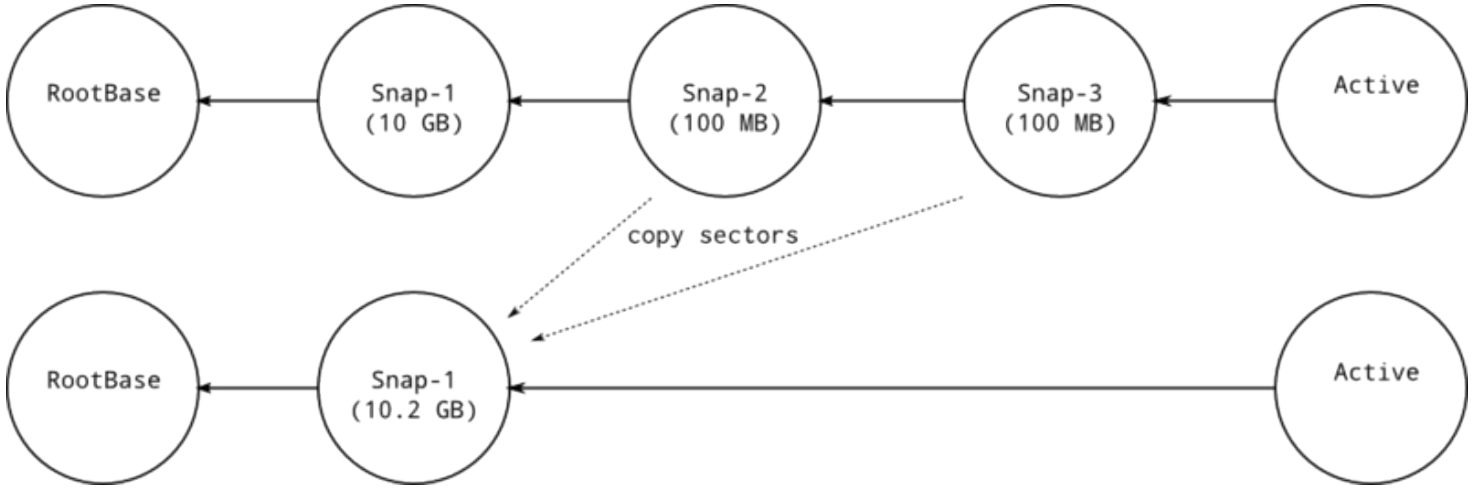
Extended L2 entries	
Subcluster size	**Max. L2 size**
4 KB	128 MB
8 KB	64 MB
16 KB	32 MB
32 KB	16 MB
64 KB	8 MB

Qemu Extended L2 (sub allocation cluster)

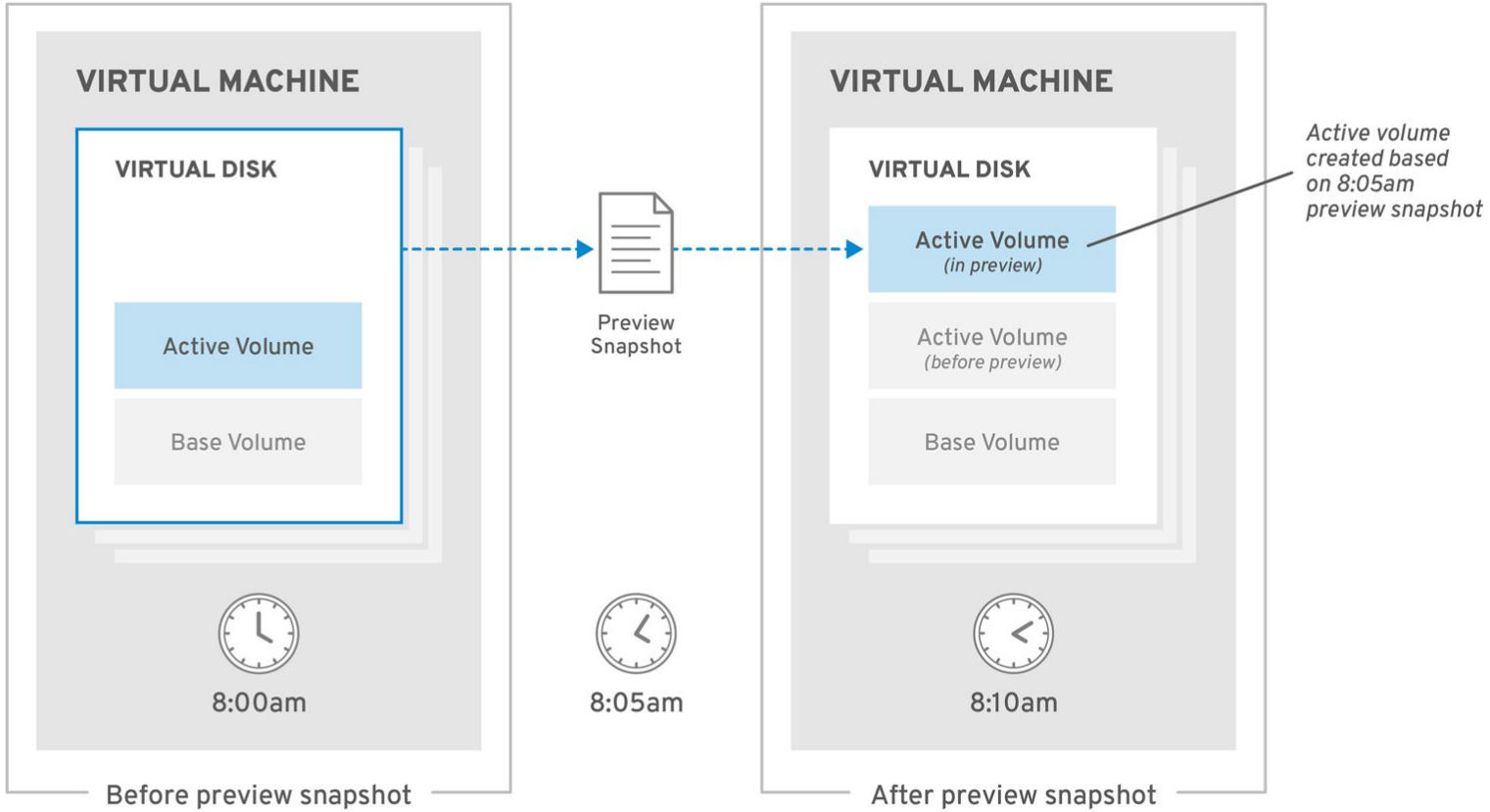


External Snapshot : Suppression

- Merge des données du snasphot dans le fichier parent.
- Job qemu : pas de lock !



External Snapshot : Replication



RHV_453539_0718

QCOW2 + Cluster filesystem ?

- Solutions possibles : OCFS2 ou GFS2

- Problèmes :
- Scalabilité
- Performance Allocation nouveaux block (Distributed lock manager)
- Stabilité (montée de version kernel, bug kernel,...)
- Lag/freeze (30s~1min) du cluster en cas de panne d’un noeud

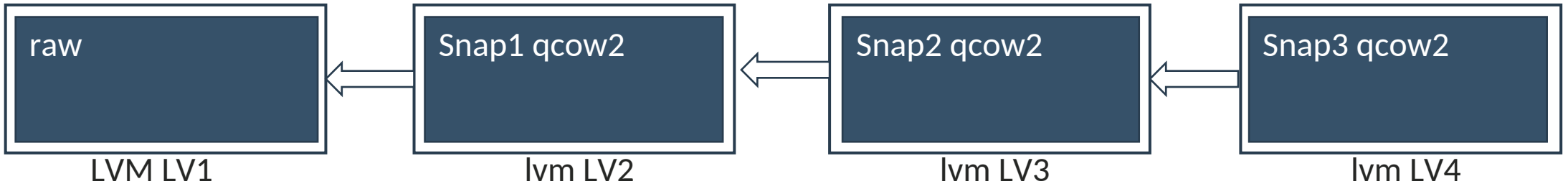
QCOW2 + SHARED BLOCK STORAGE ????

QCOW2 + LVM THICK !

QCOW2 + LVM THICK !

Formatage LVM en QCOW2 sans fichier !

lvcreate -L1000 -n vgtest/vm-100-disk-0
qemu-img create -f qcow2 -o extended_l2=on,cluster_size=128k **/dev/lvm/vgtest/vm-100-disk-0**

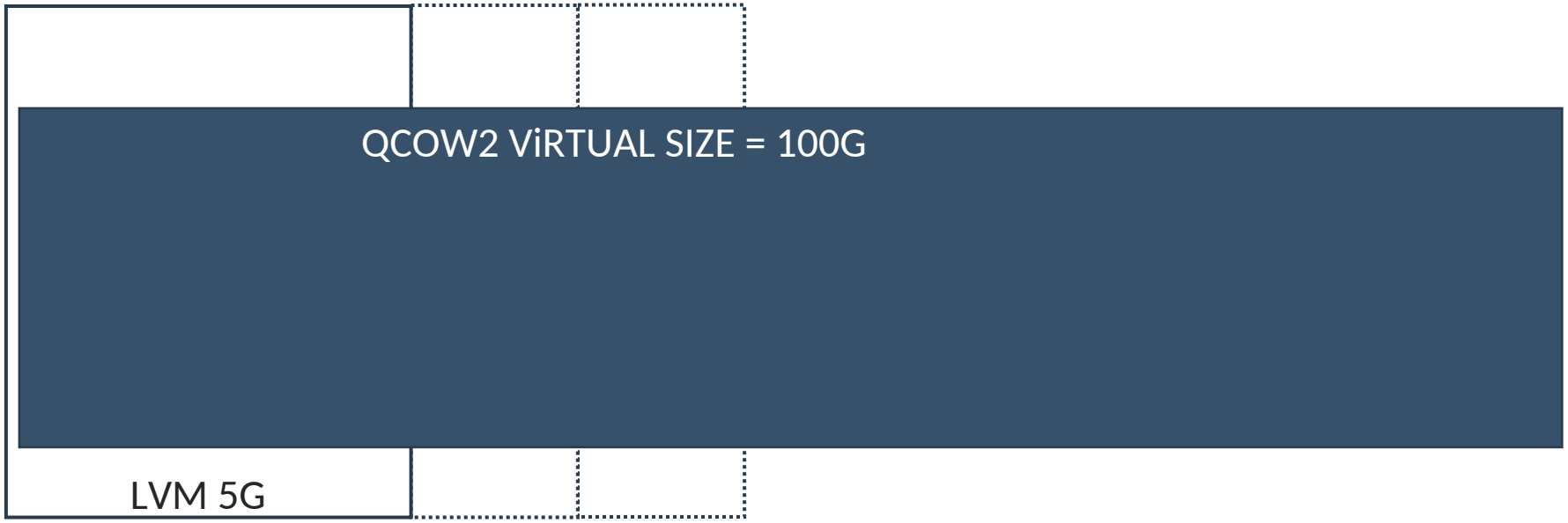


THIN PROVISIONING ?

LVM – THICK

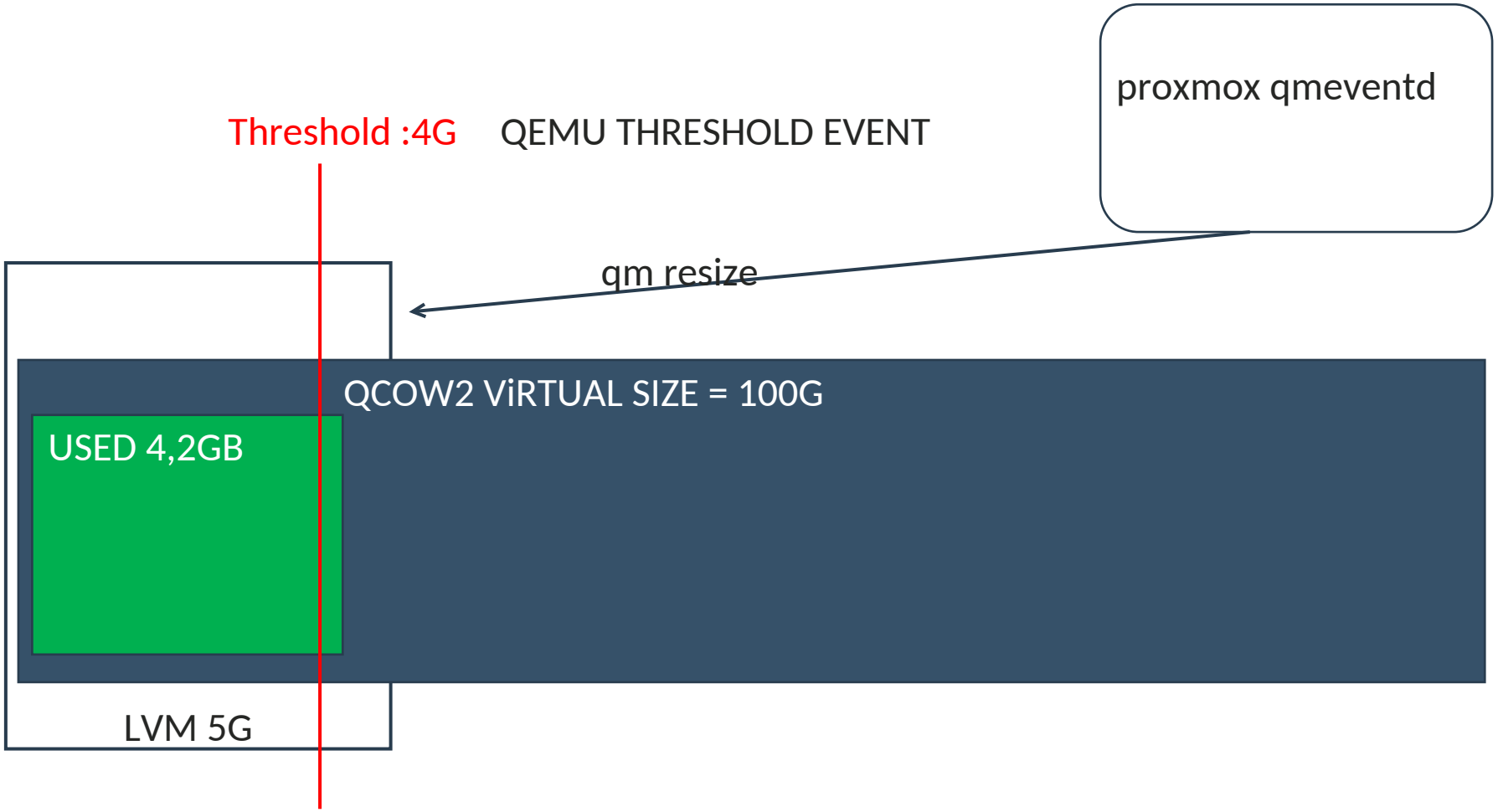
THIN PROVISIONING ?

- Qcow2 virtual size > LVM real size
- Agrandissement LVM dynamique fait par proxmox par chunk de x Gb

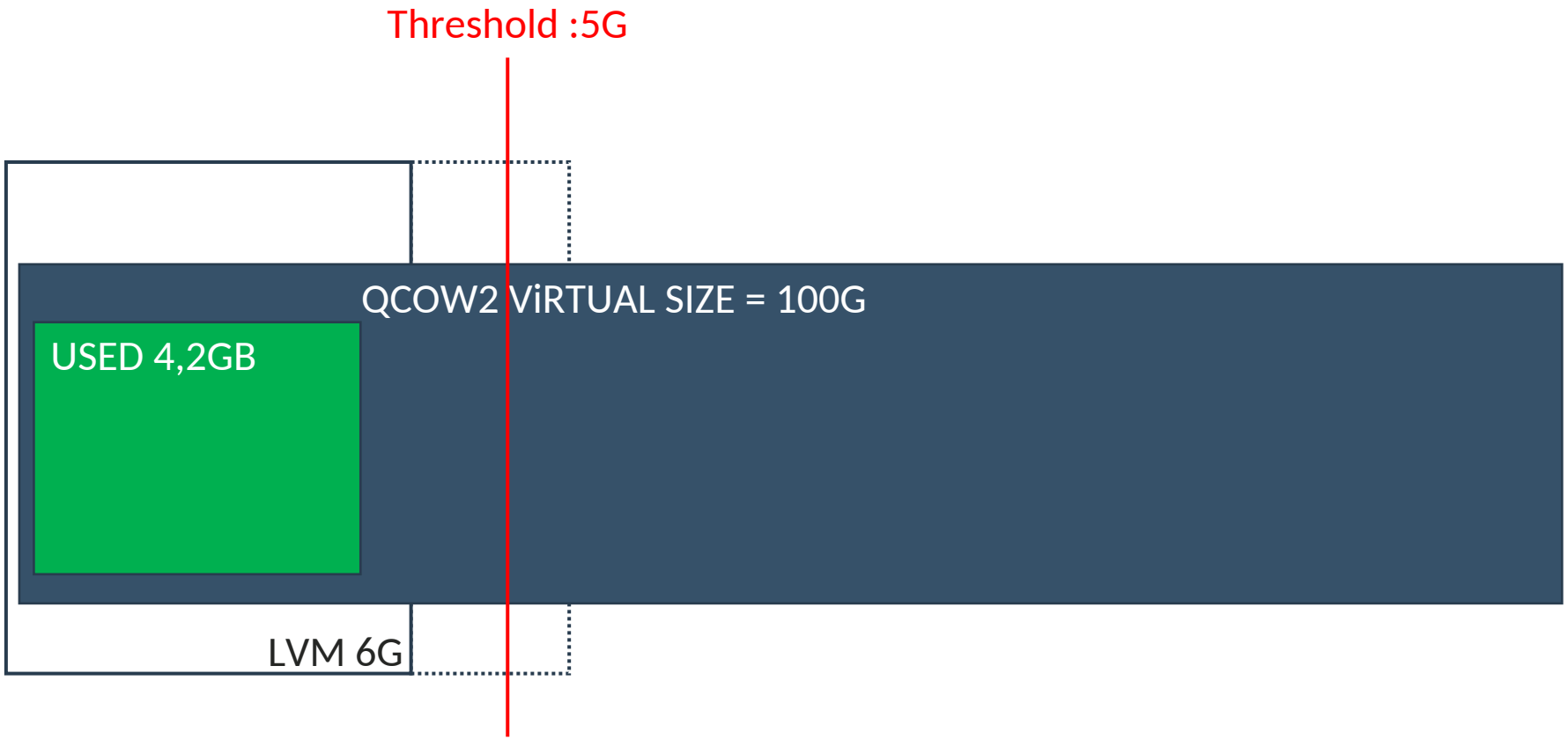


THIN PROVISIONING ?

THIN PROVISIONING ?



THIN PROVISIONING ?



LXC ?

dm-qcow2 -> still beta

https://lore.kernel.org/lkml/164846619932.251310.3668540533992131988.stgit@pro/T/

Stockages et Features

Description	Plugin type	Level	Shared	Snapshots	Thinprovisioning	Trim	Backups	Replication	Chiffrement	Stable
lvmqcow2	lvmqcow2	block	yes	yes	yes	no	yes	yes	not yet (qcow2)	not yet
Directory	dir	file	no	yes (.qcow2)	yes (.qcow2)	yes	yes	no	not yet (qcow2)	yes
NFS	nfs	file	yes	yes (.qcow2)	yes (.qcow2)	no (nfs 4.2?)	yes	no	not yet (qcow2)	yes
CIFS	cifs	file	yes	yes (.qcow2)	yes (.qcow2)	no	yes	no	not yet (qcow2)	yes
GlusterFS	glusterfs	file	yes	yes (.qcow2)	yes (.qcow2)	yes	yes	no	not yet (qcow2)	yes
LVM	lvm	block	yes	no	no	no	yes	no	no	yes
LVM-thin	lvmthin	block	no	yes	yes	yes	yes	no	no	yes
ZFS (local)	zfspool	block	no	yes	yes	yes	yes	yes	yes	yes
iSCSI/kernel	iscsi	block	yes	no	no	no	yes	no	no	yes
iSCSI/libiscsi	iscsidirect	block	yes	no	no	no	yes	no	no	yes
Ceph/RBD	rbd	block	yes	yes	yes	yes	yes	yes	yes	yes
ZFS over iSCSI	zfs	block	yes	yes	yes	yes	yes	yes	yes	yes
BTRFS	btrfs	block	no	yes	yes	yes	yes	no	no	technology preview

LILLE - PARIS - LYON - BORDEAUX - STRASBOURG - MONTBÉLIARD - TROYES - SAINT-BRIEUC – ARRAS - NANTES

contact@groupe-cyllene.com
T. +33 (1) 41 19 40 40
WWW.GROUPE-CYLLENE.COM