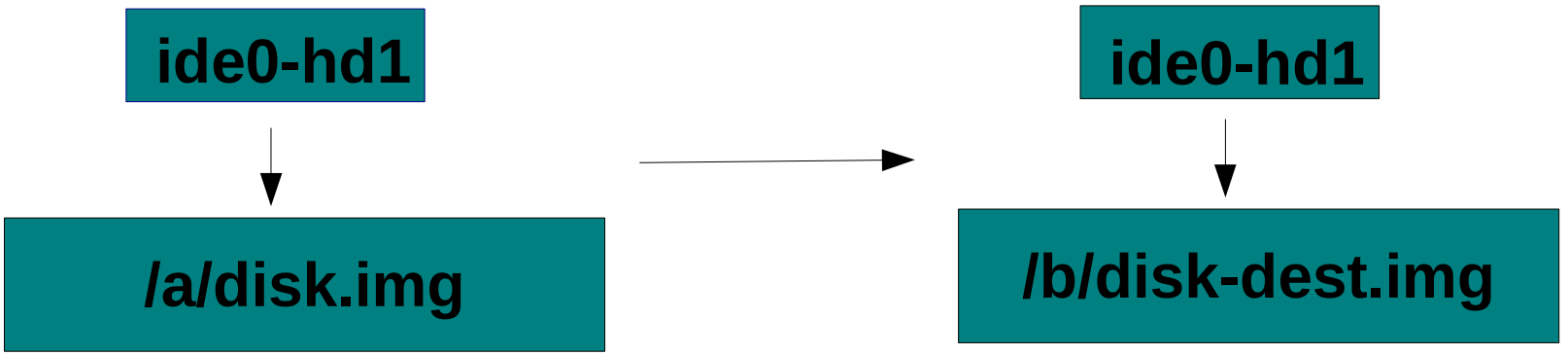
QEMU live block copy

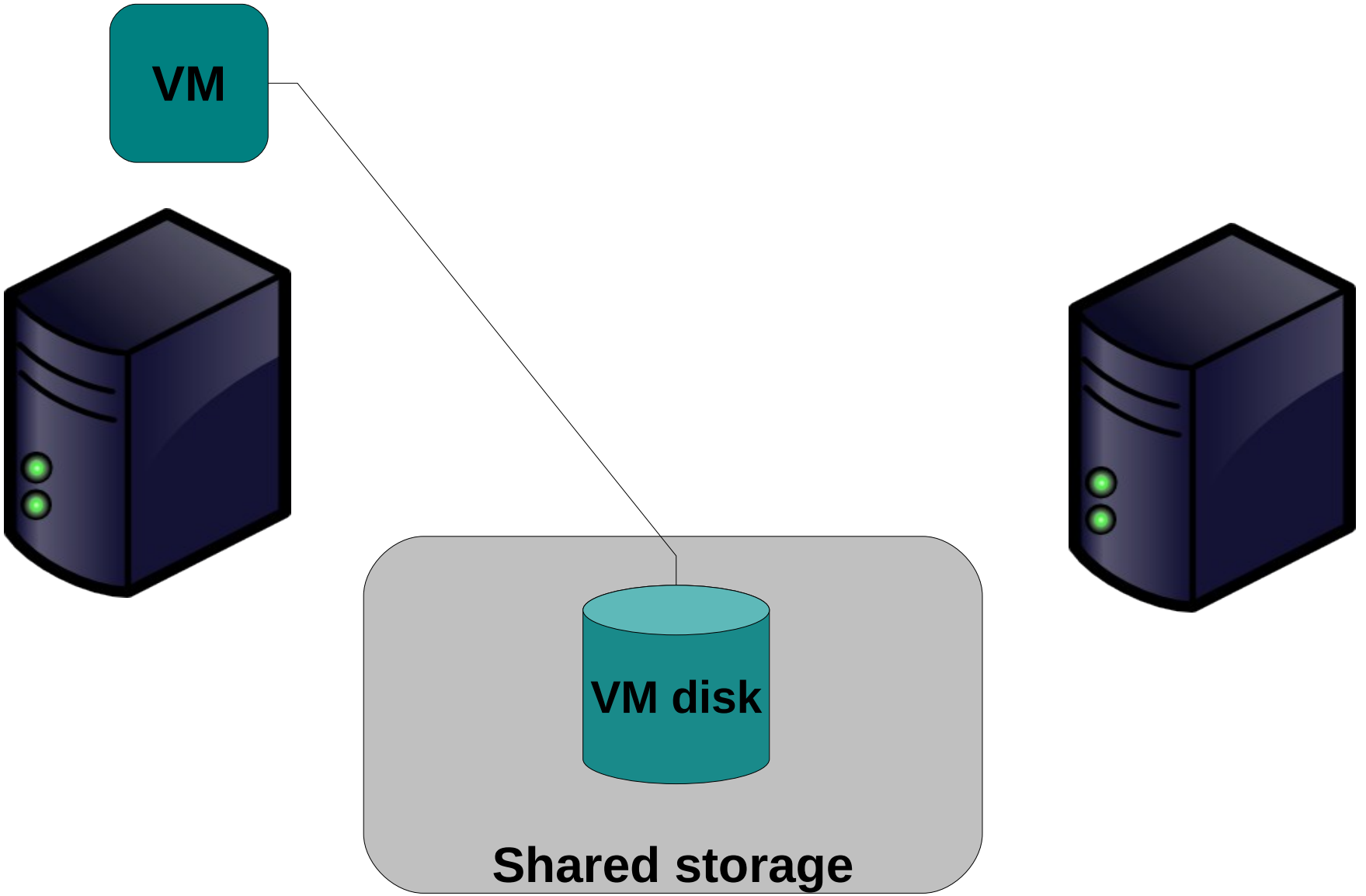
Marcelo Tosatti
KVM Forum 2011 – Vancouver, CA

Introduction: live copy operation

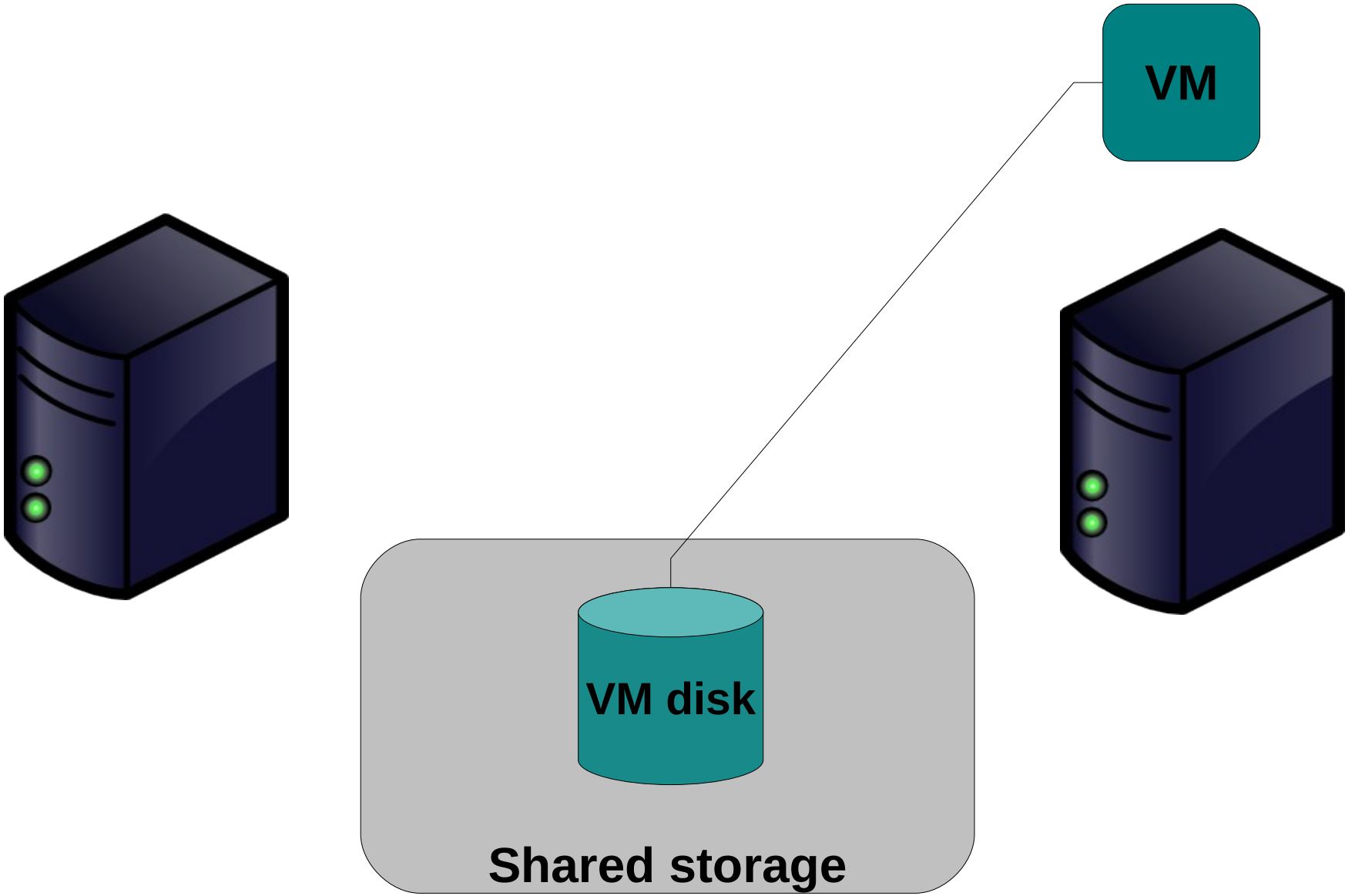
- Copies in use guest disk image to destination image.
- Switches guest disk to destination image.



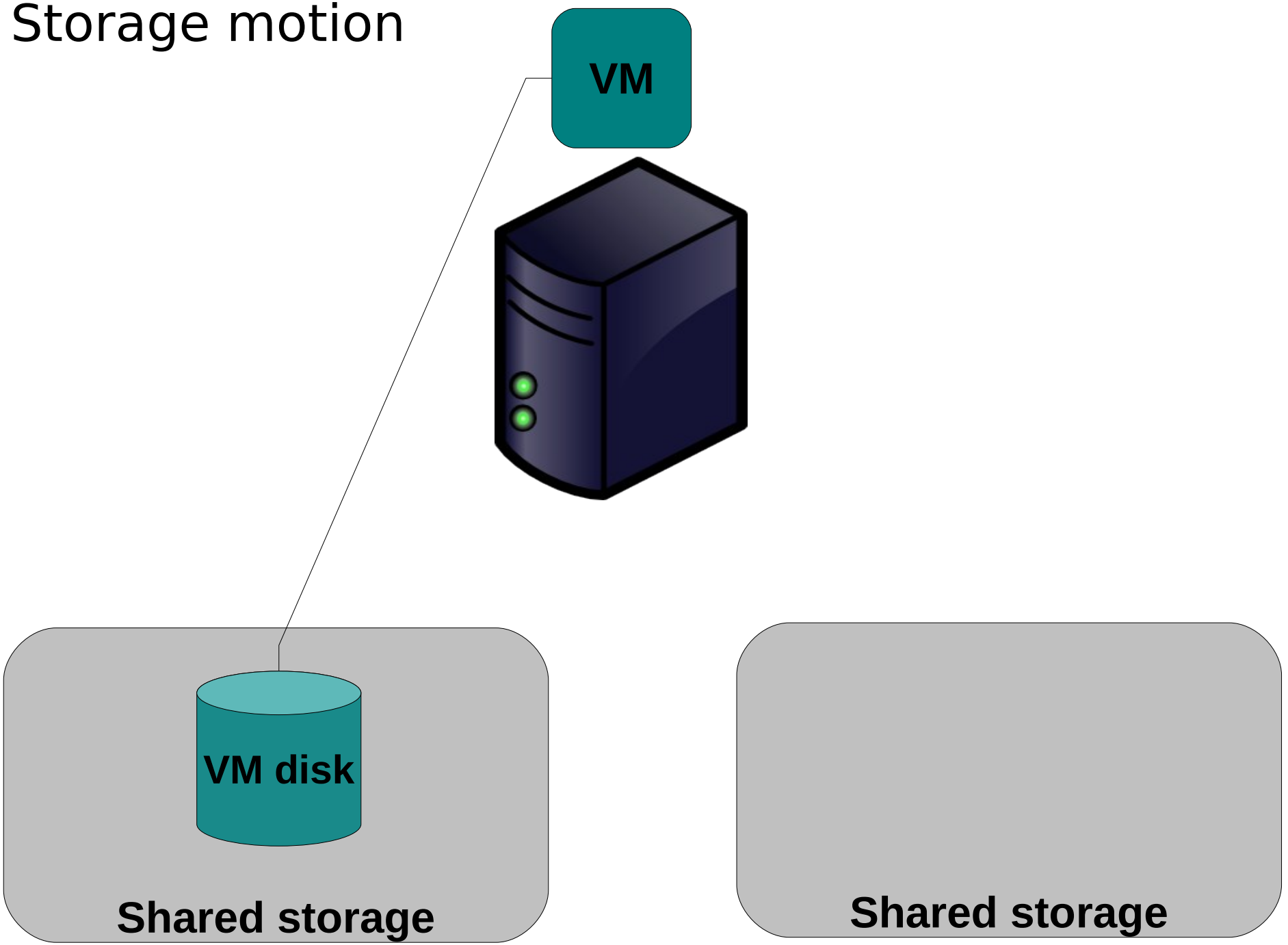
Live migration



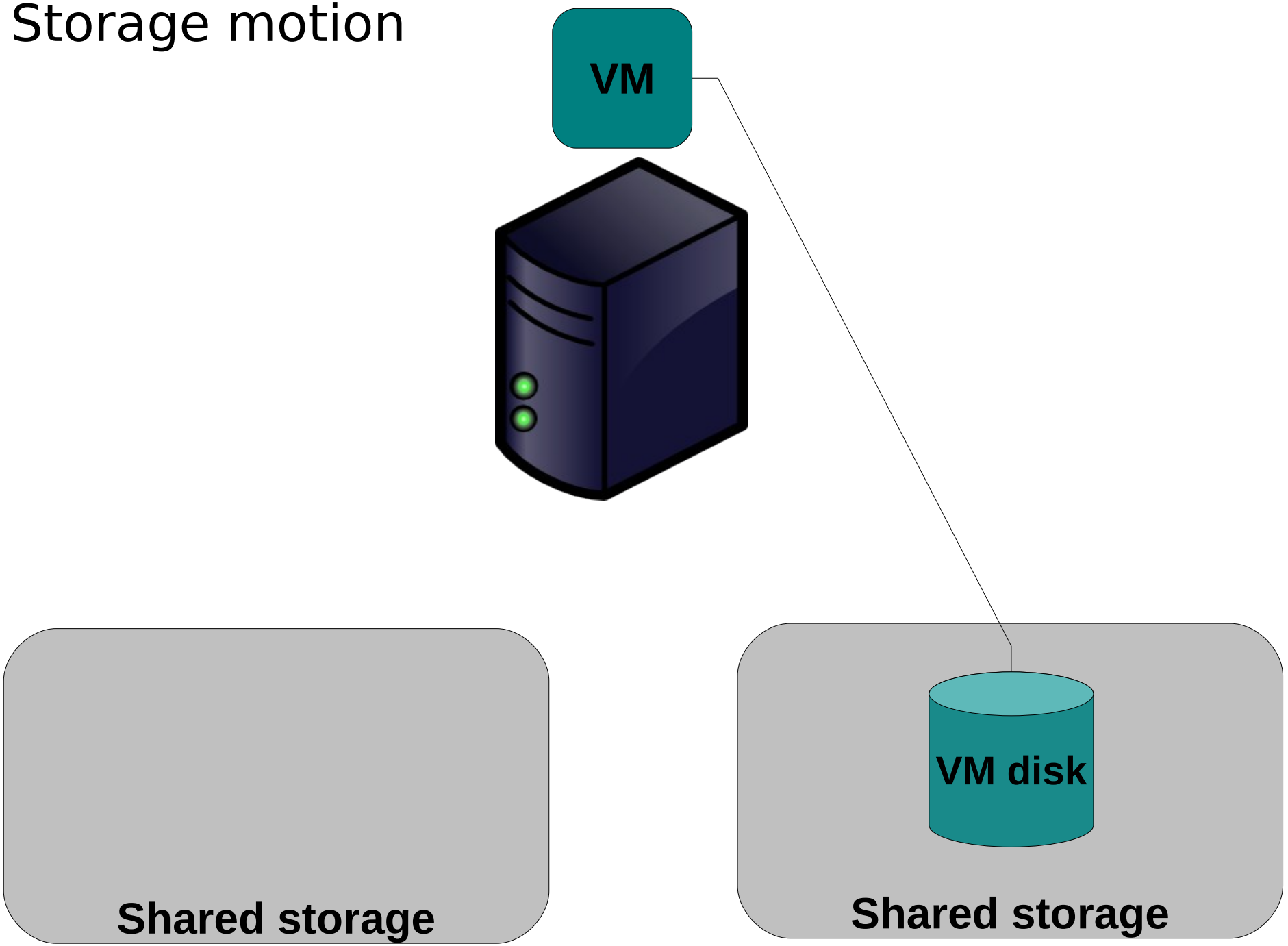
Live migration



Storage motion



Storage motion



Use cases – storage motion

- Move guest image(s) from local storage to SAN storage unit and vice-versa.
- Useful for repairs, maintenance tasks (eg: move to new storage unit).
- Useful to manage guest images across storage units for speed and capacity arrangements.

Use cases – image format conversion.

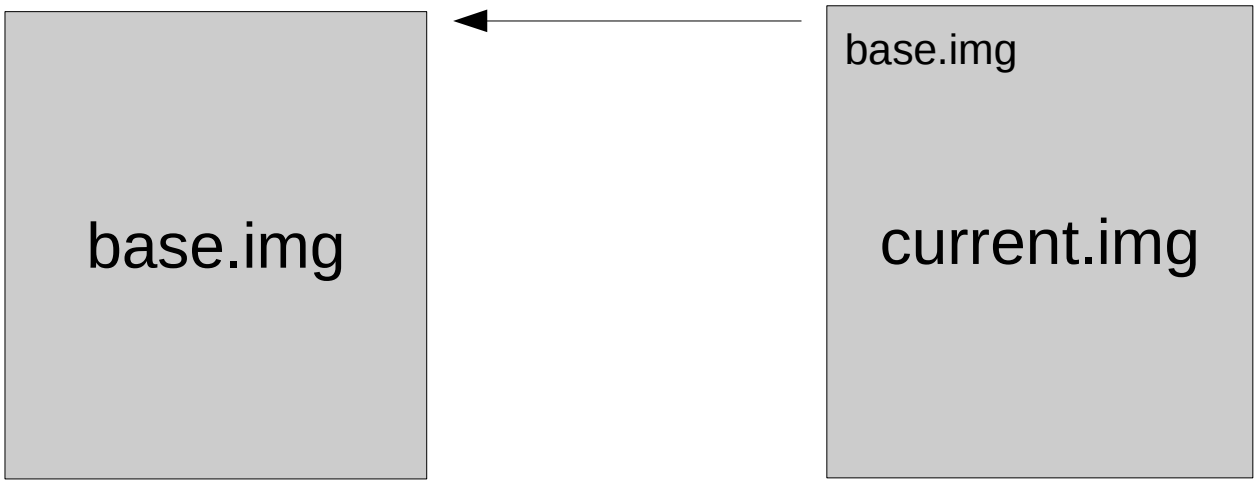
- Convert guest disk image format.

Use cases – snapshot merging.

- Collapse (merge) chains formed with QCOW2 external snapshots.

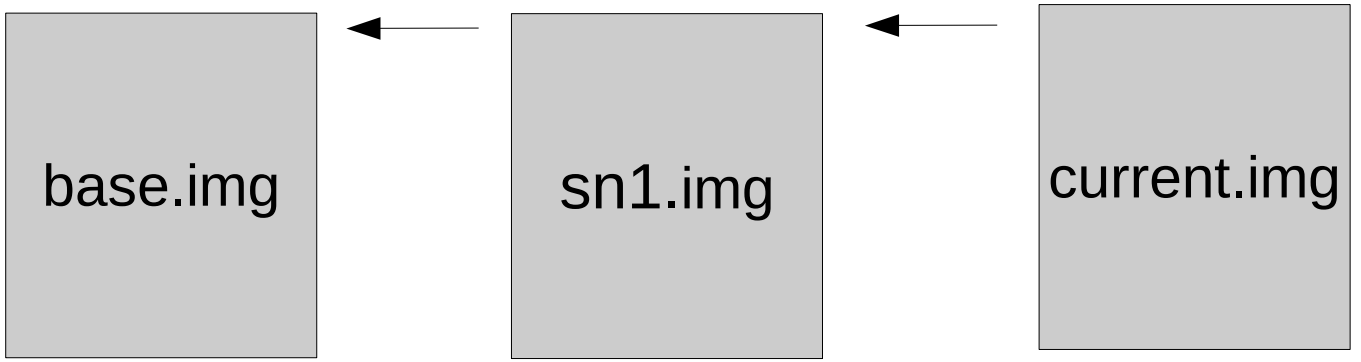
Qcow2 backing files

- Image contains difference to base image.
- Copy-On-Write.



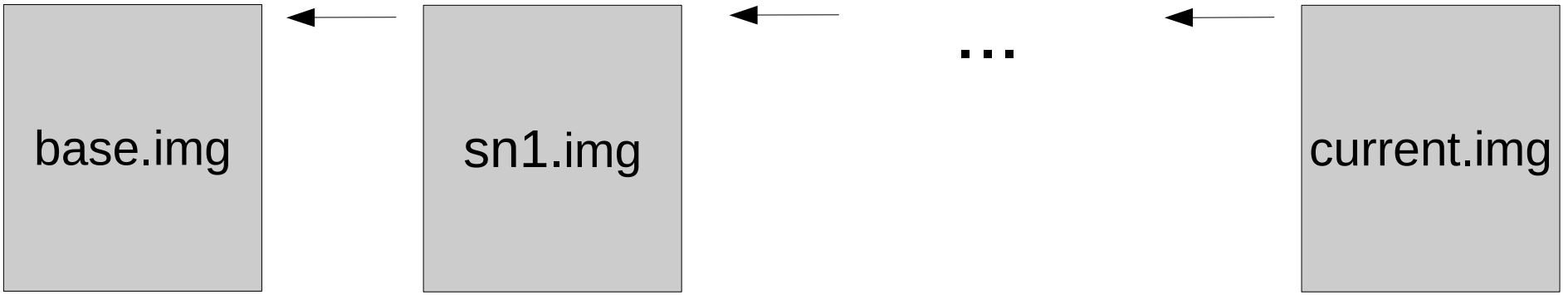
Snapshots with base files

- New image is created to accomodate writes. Previous image becomes a snapshot.
- Live snapshots: snapshot_blkdev command.

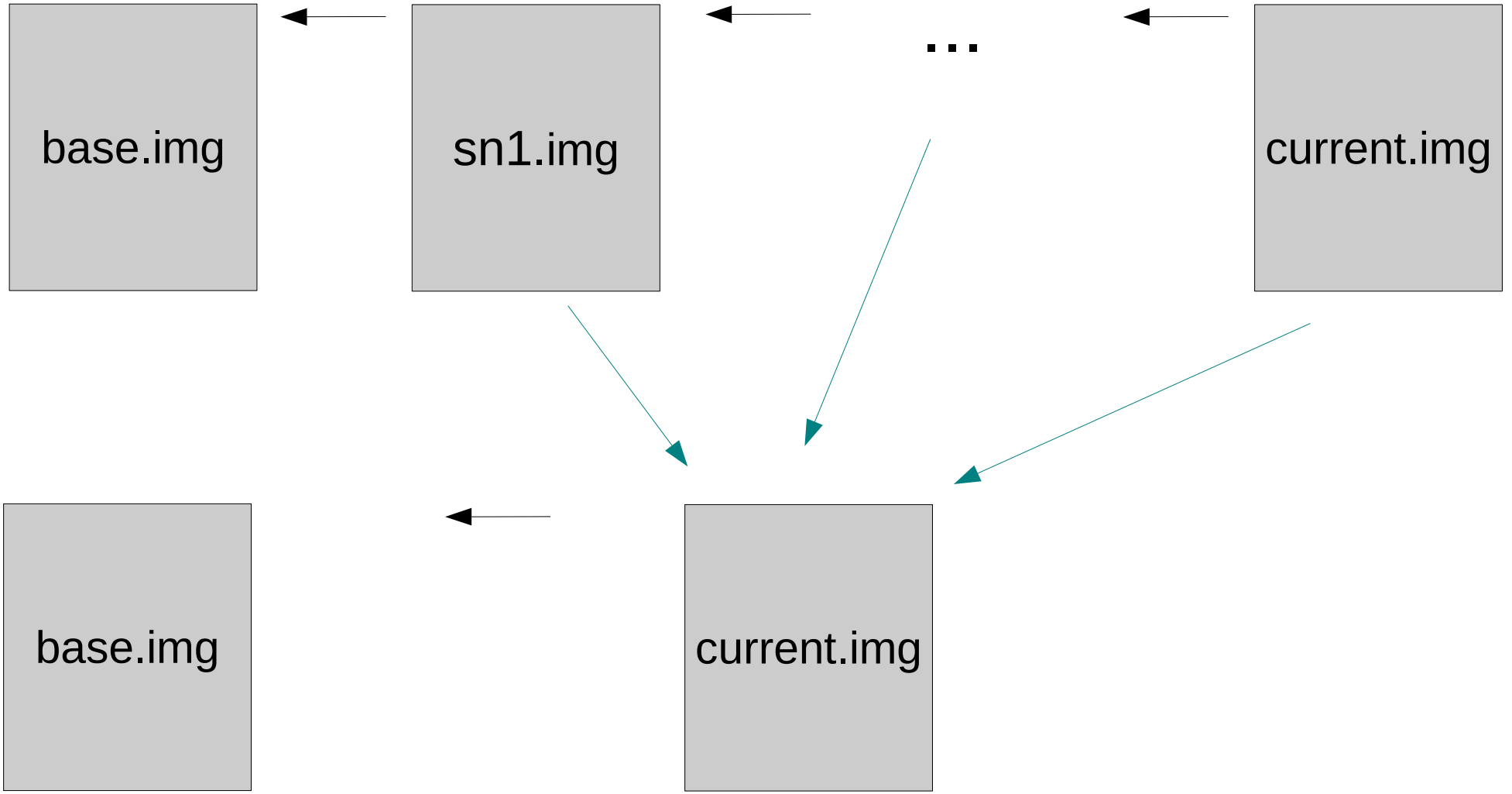


Qcow2 snapshot chains

- After many snapshots...
- Reading data traverses back image chain, reading and caching metadata.



Merging snapshots with live copy



Live block copy interface

- Monitor command:

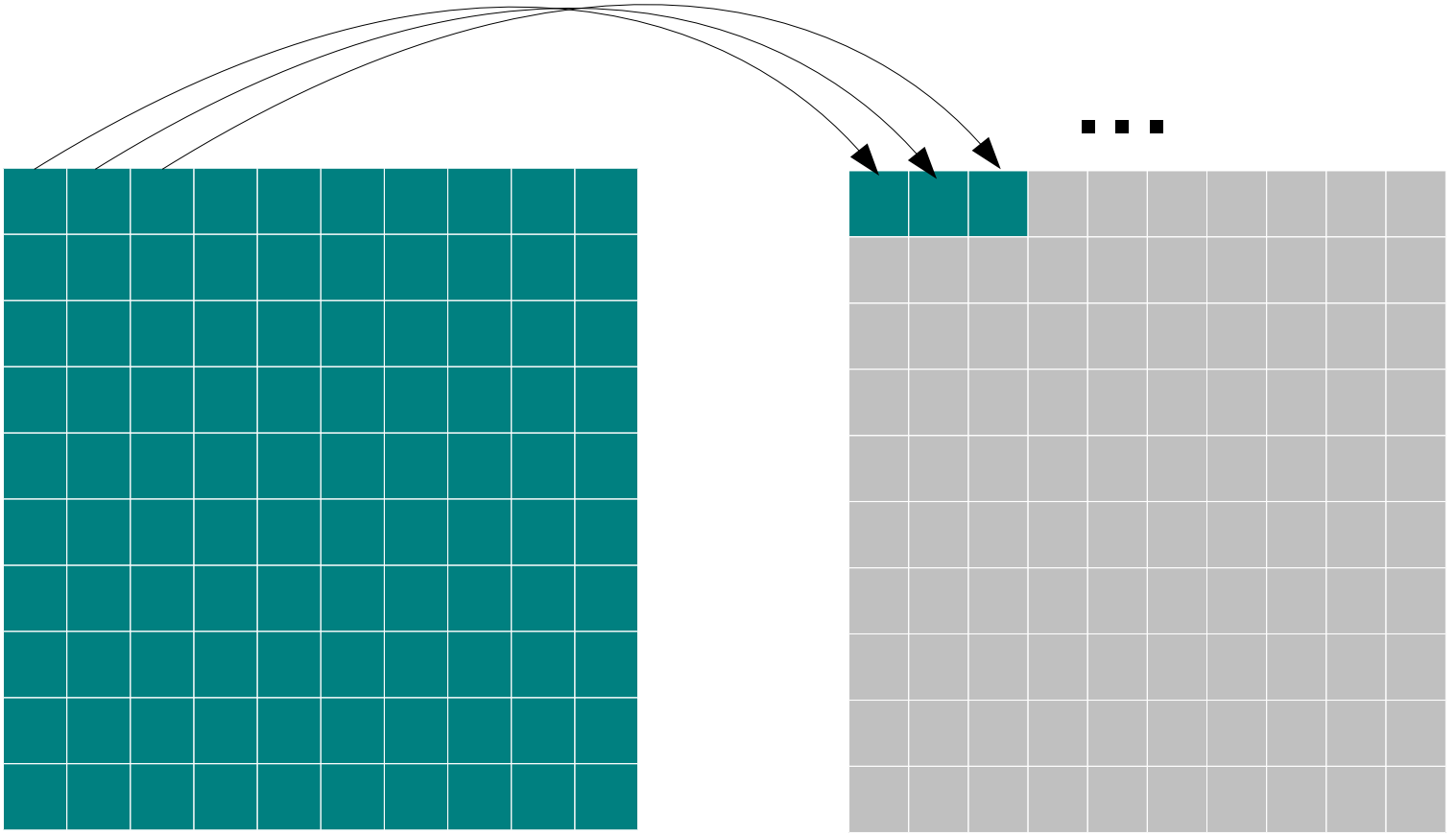
  block_copy guest-disk-ID /path/to/new/image.img
- image.img created externally.

Live block copy internals

- 3 stages: bulk, dirty and mirrored writes.

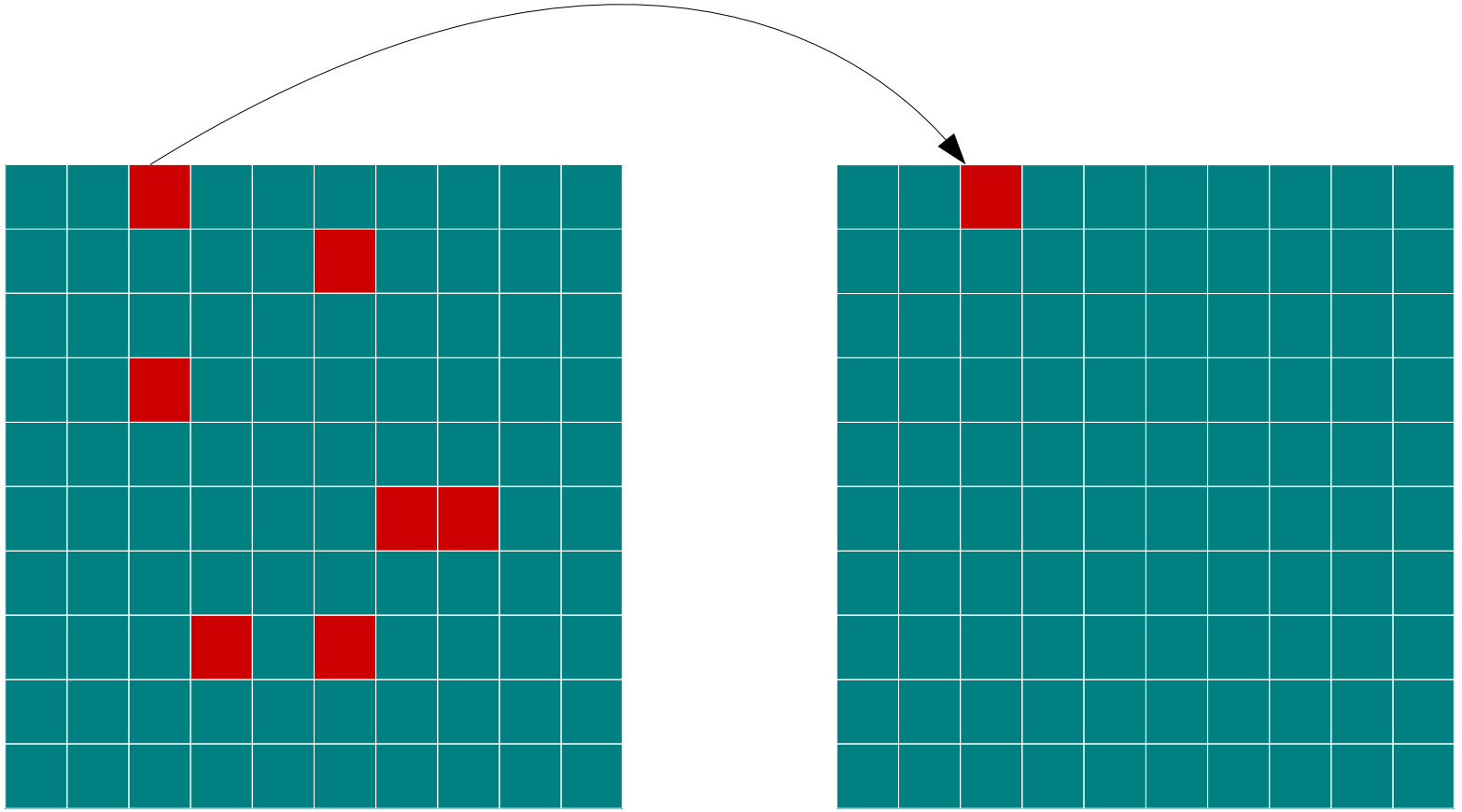
Bulk

- Log guest writes to source block dev (dirty bitmap).
- Copy sectors from 1...LAST_SECTOR to destination block dev.



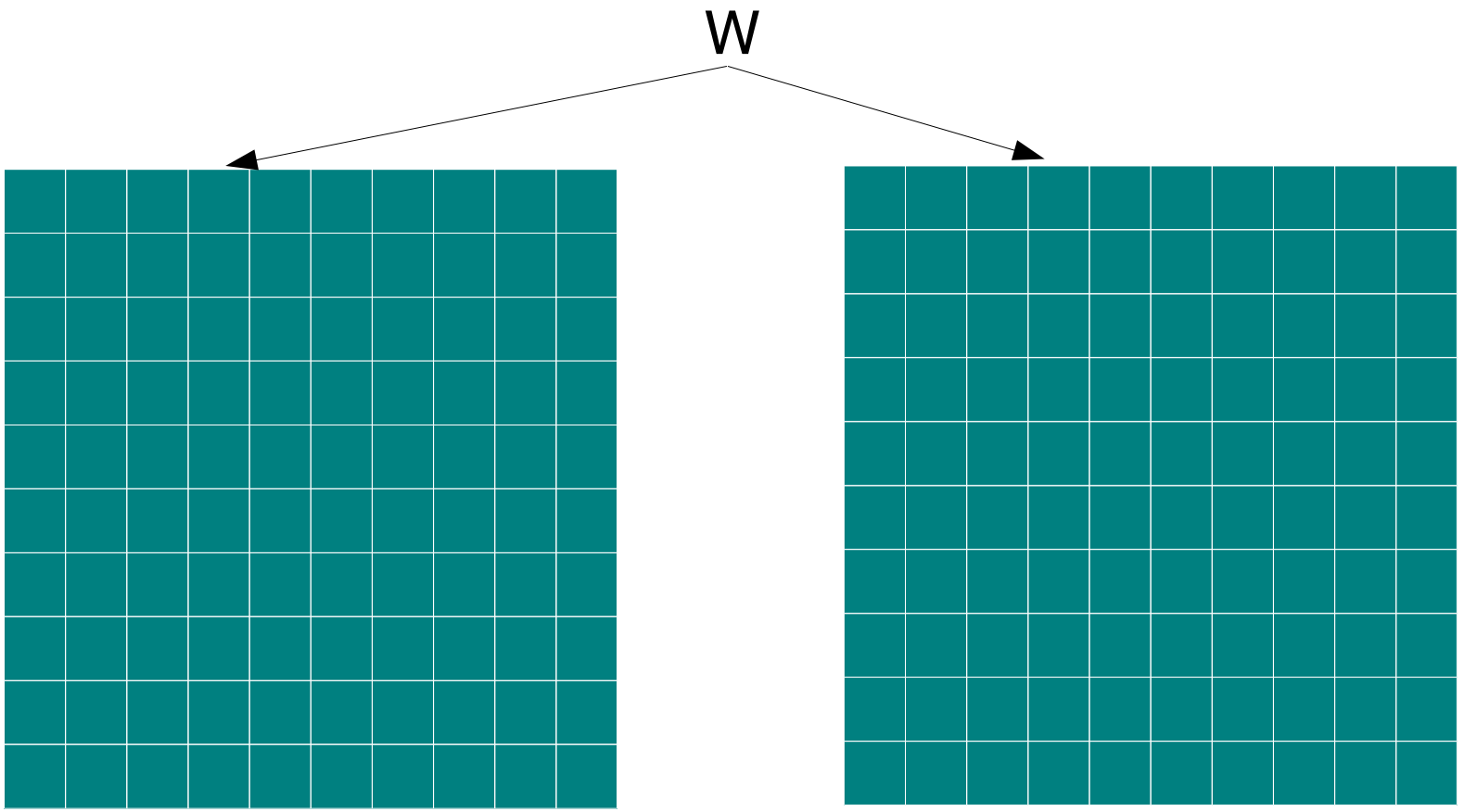
Dirty

- Copy dirty blocks (that have been modified by the guest during bulk stage), from dirty bitmap.



Mirrored writes

- Duplicate writes to source and destination.
- Both images are valid (crash scenario).



Mirrored writes

- Until receives switch command from management.
- Writes to destination only.

In the meantime...

- Requirement arises to quickly deploy guest whose base image is on slow remote storage.
- Copying entire image takes too long.

Image streaming

- Copy On Read.

guest: read sector y

remote.img

local.img

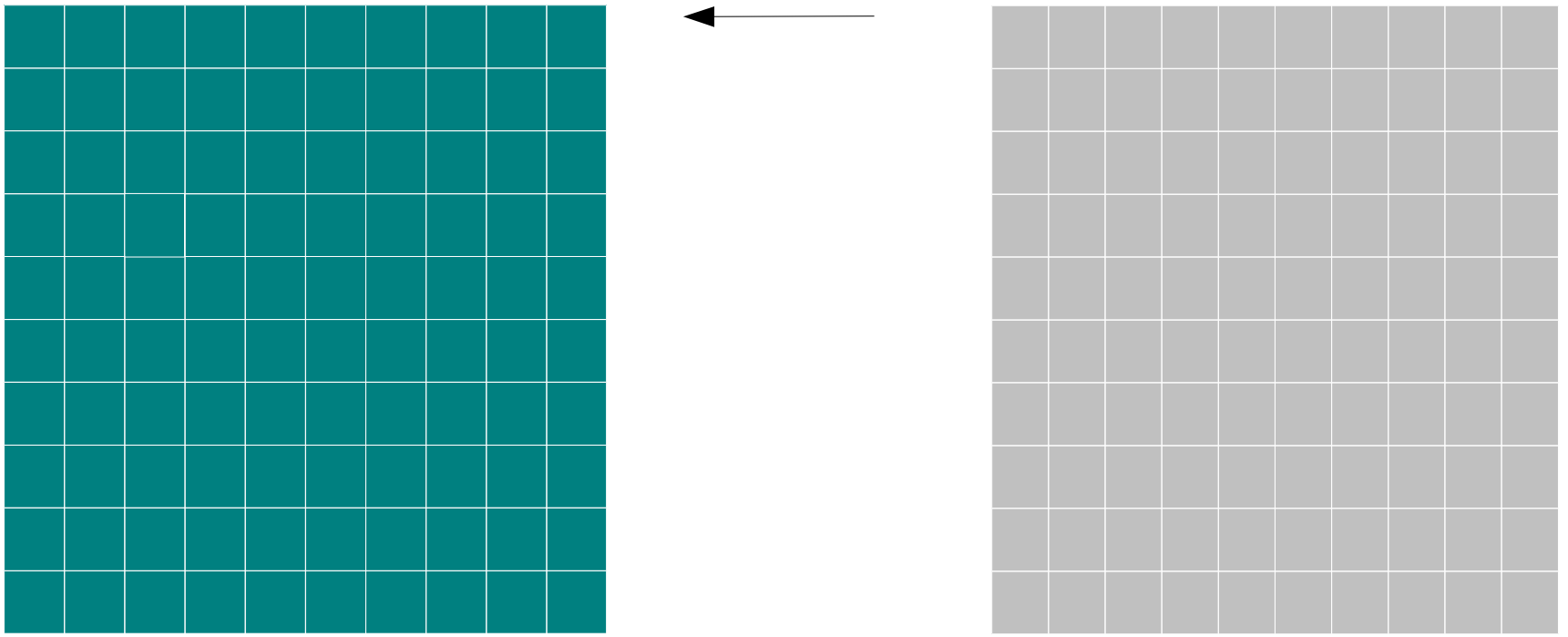


Image streaming

- Background copy. With COR that means reading entire image.

Image streaming: QED patches

- Implemented by IBM.
- COR logic in image format implementation.
- Generic interface for streaming entire image.

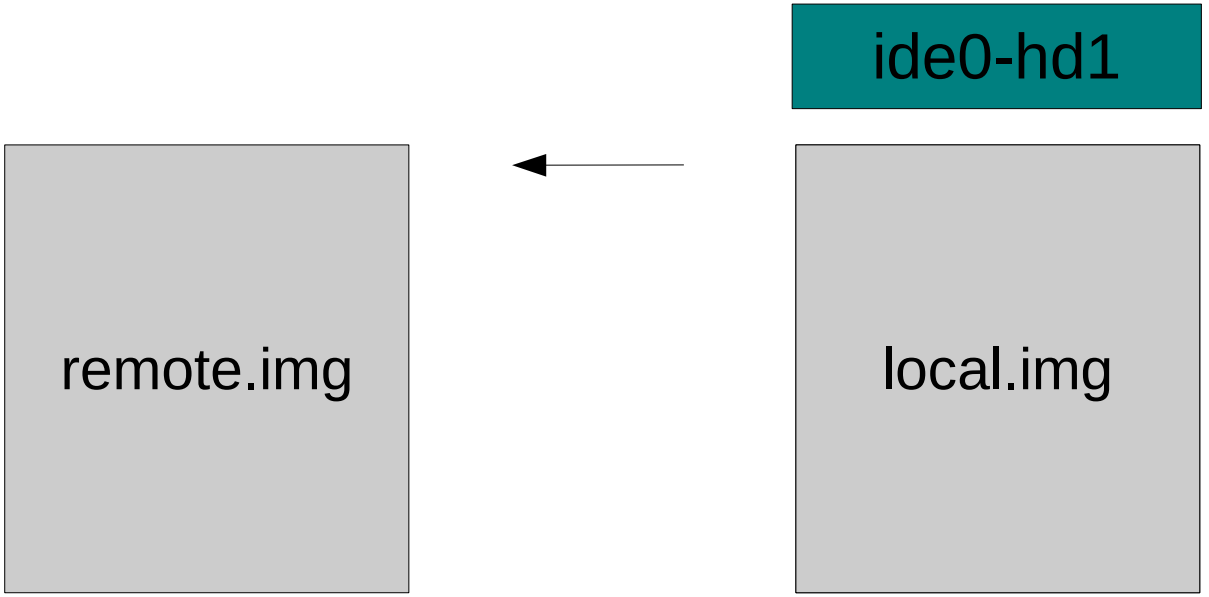
Streaming

- Observation: streaming and live block copy are essentially the same: copy guest disk image while its being accessed.
- Difference is that live block copy copies to an image, and image streaming copies from and image.
- Kevin suggests one implementation to address both requirements.

Blkstream: unified stream/live copy

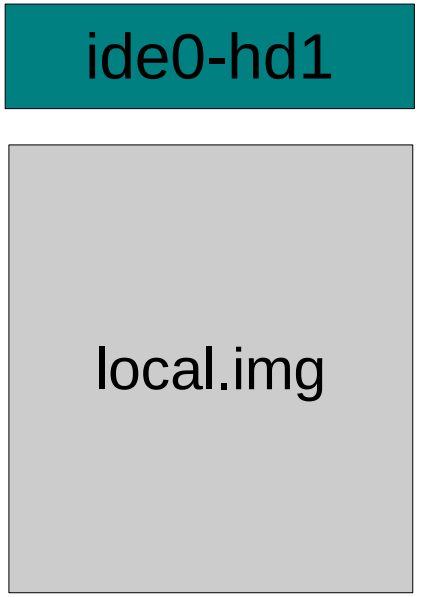
- Block driver that implements COR.
- Works with any format that supports backing files.
- Interface to sequentially read entire image.

Image streaming with blkstream



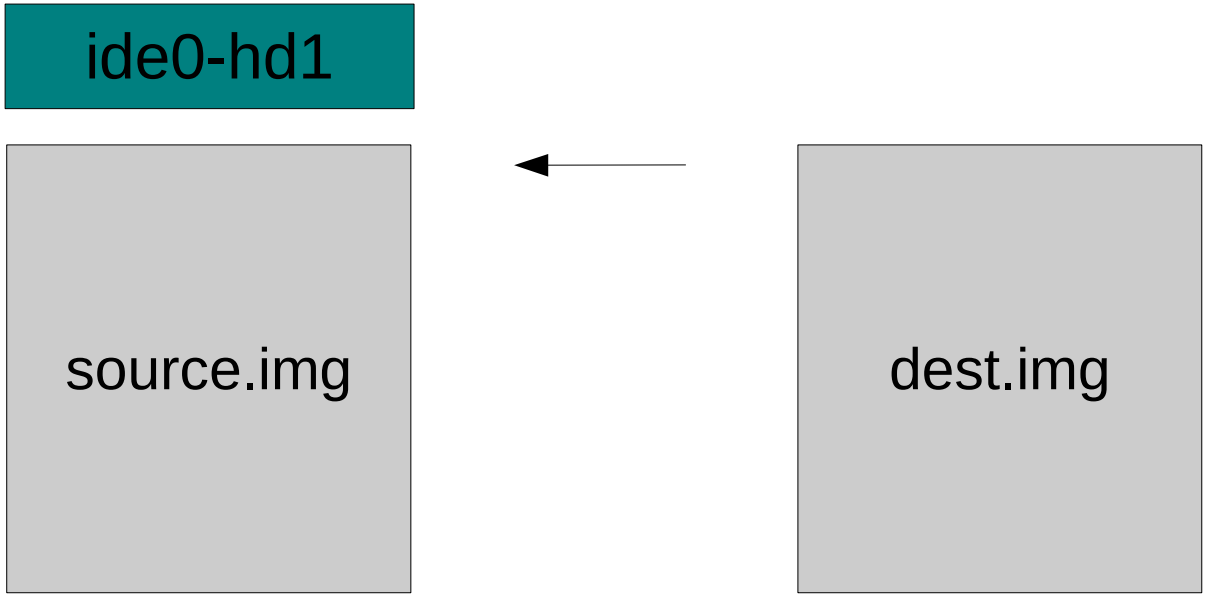
1) start guest with COR enabled.

Image streaming with blkstream



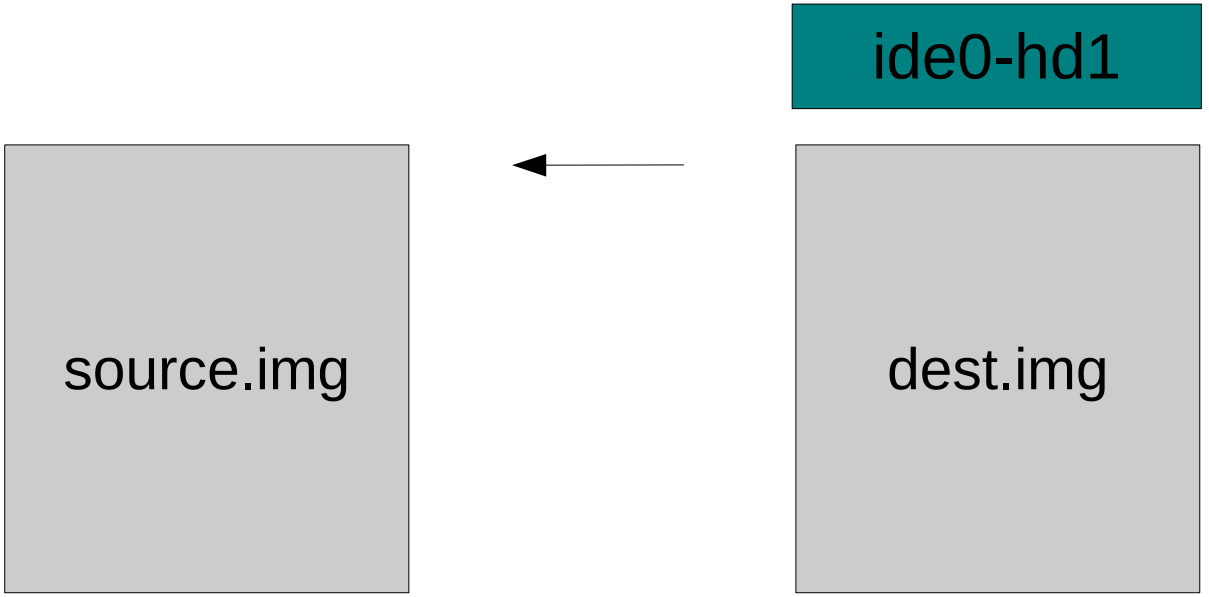
1) start guest with streaming enabled.
2) once streaming is finished, remove backing file reference.

Storage motion with blkstream



1) create destination image with source as backing file.

Storage motion with blkstream

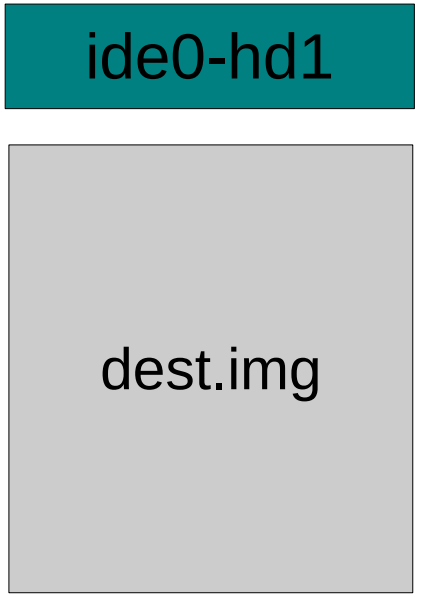


1) create destination image with source as backing file.

2) switch to destination (management must update its record).

3) read all clusters.

Storage motion with blkstream



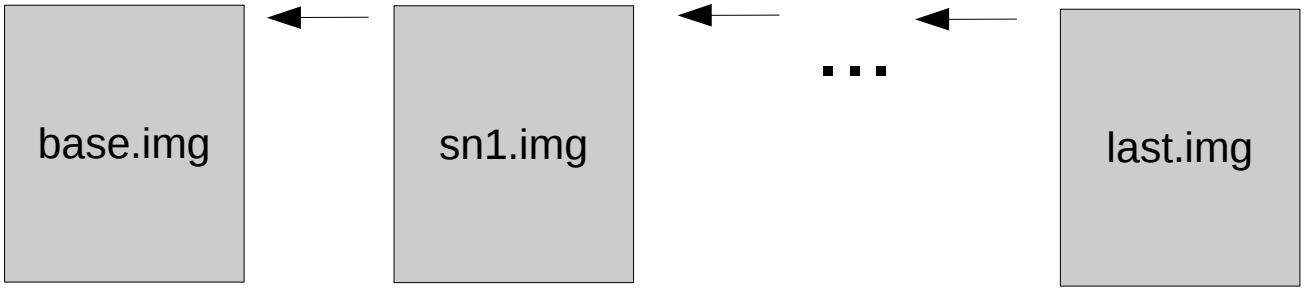
1) create destination image with source as backing file.

2) switch to destination (management must update its record).

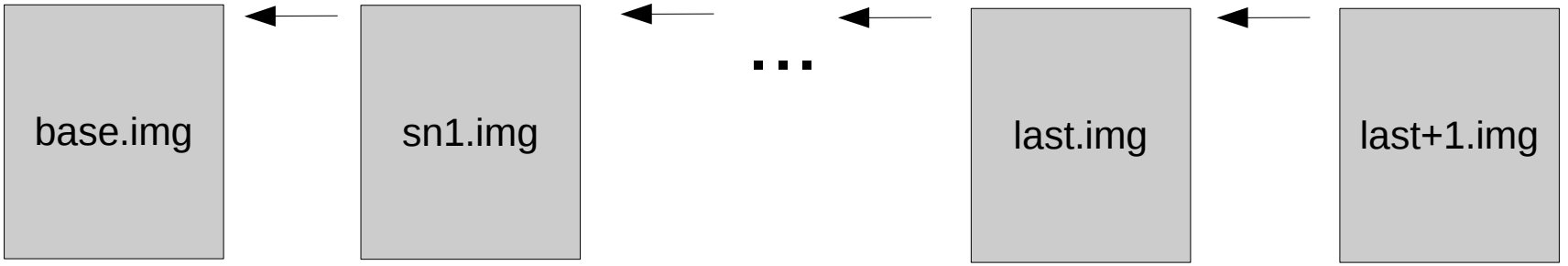
3) read all clusters.

4) remove backing file reference.

Storage motion with shared base image

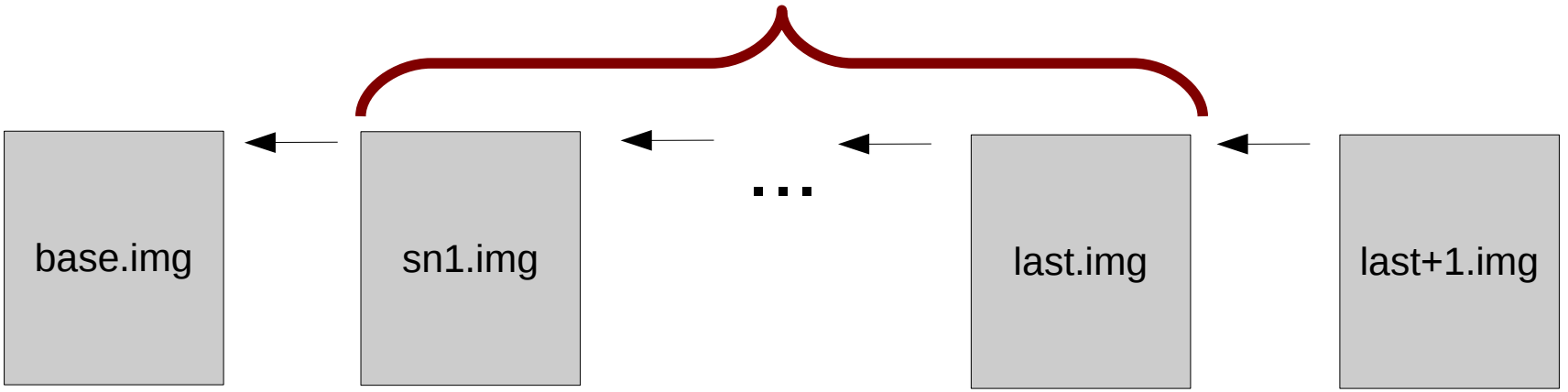


Storage motion with shared base image



1) create new image with last as backing file.

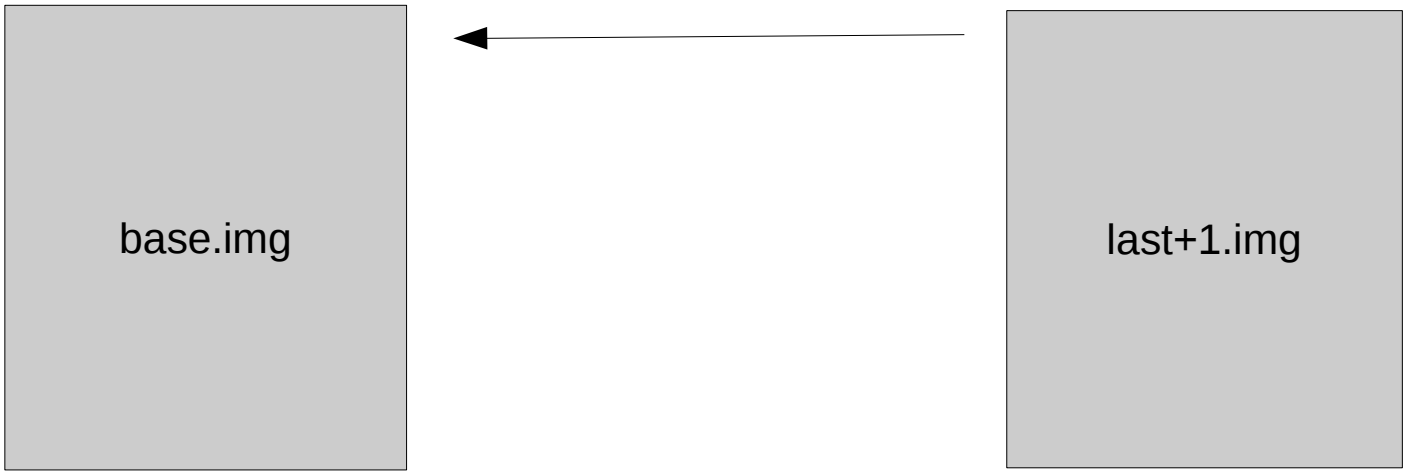
Storage motion with shared base image



1) create new image with last as backing file.

2) only COR if cluster allocated up the chain from shared base.

Shared base image



1) create new image with last as backing file.

2) only COR if cluster allocated up the chain from shared base.

3) read all clusters, write final backing file.

COW emulation

- Image streaming requires backing file support.
- For formats that do not support backing files, external support will be provided.
- Essentially on disk bitmap with allocated information.
- Robert Wang @ IBM working on it.

Questions? Comments?