Ganeti - Overview

Brian Candler
Network Startup Resource Center
brian@nsrc.org

What is Ganeti?

- Cluster virtualization software by Google
- Used for all their in-house (office) infrastructure
- Very actively developed
  - preferred platform is Debian, others can be used
- Good user community
- Totally free and open source

Ganeti features

- Manages hypervisors: Xen, KVM, LXC ...
- Manages storage
- Killer feature: can configure DRBD replication
  - Fault-tolerance and live migration on commodity hardware without shared storage!
- VM image provisioning
- Remote API for integration with other systems

Ganeti caveats

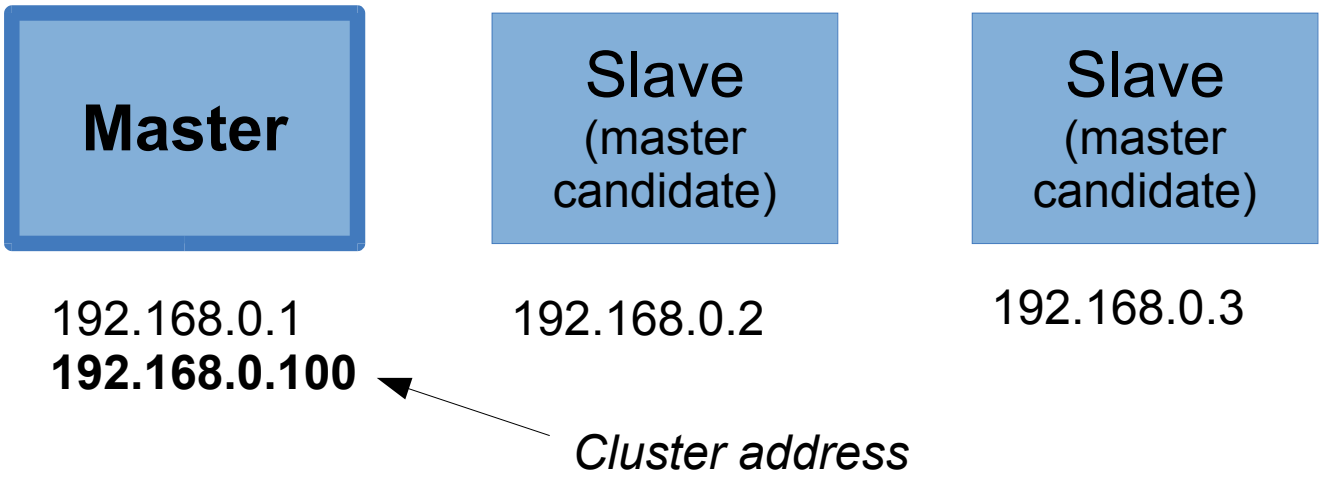
- Command-line based
  - Has a learning curve
  - Lots of features
  - Still much simpler than OpenStack etc
- No web interface included
- Doesn't natively support the idea of different users and roles
  - but Ganeti Web Manager adds this

Terminology

- *instance* = a virtual machine (guest)
- *node* = a physical server (host)
- *cluster* = all nodes

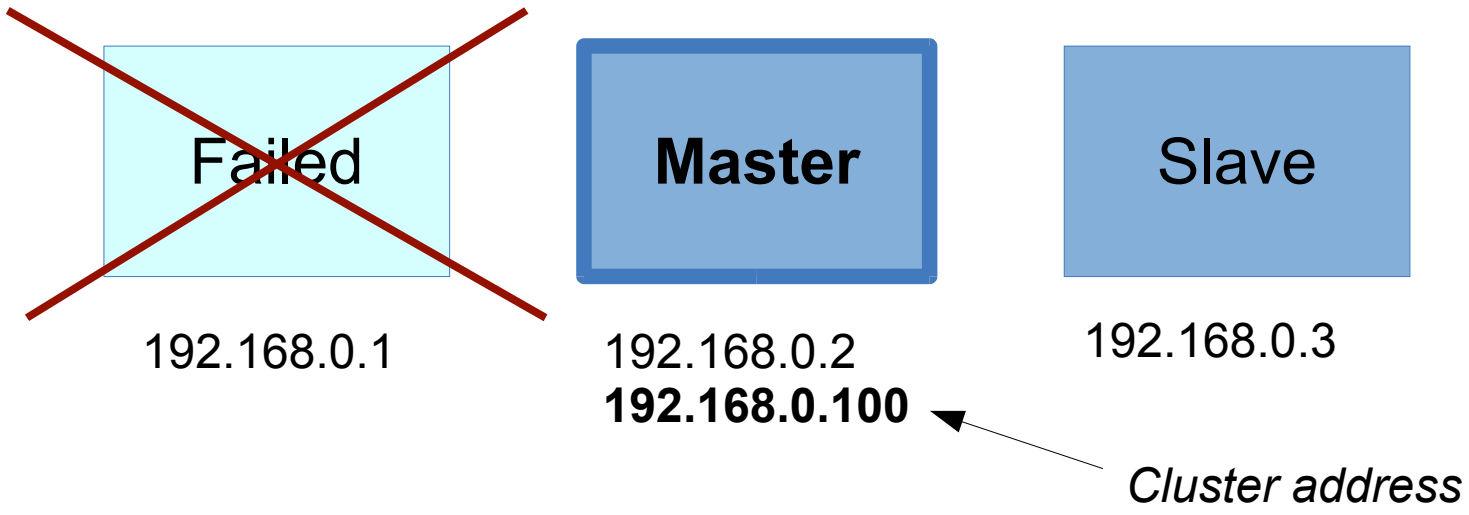
Cluster management

- One of the nodes is the *master*
  - It maintains all the cluster state and copies it to other nodes
  - It sends commands to the other nodes
- It has an extra IP alias (the cluster address)

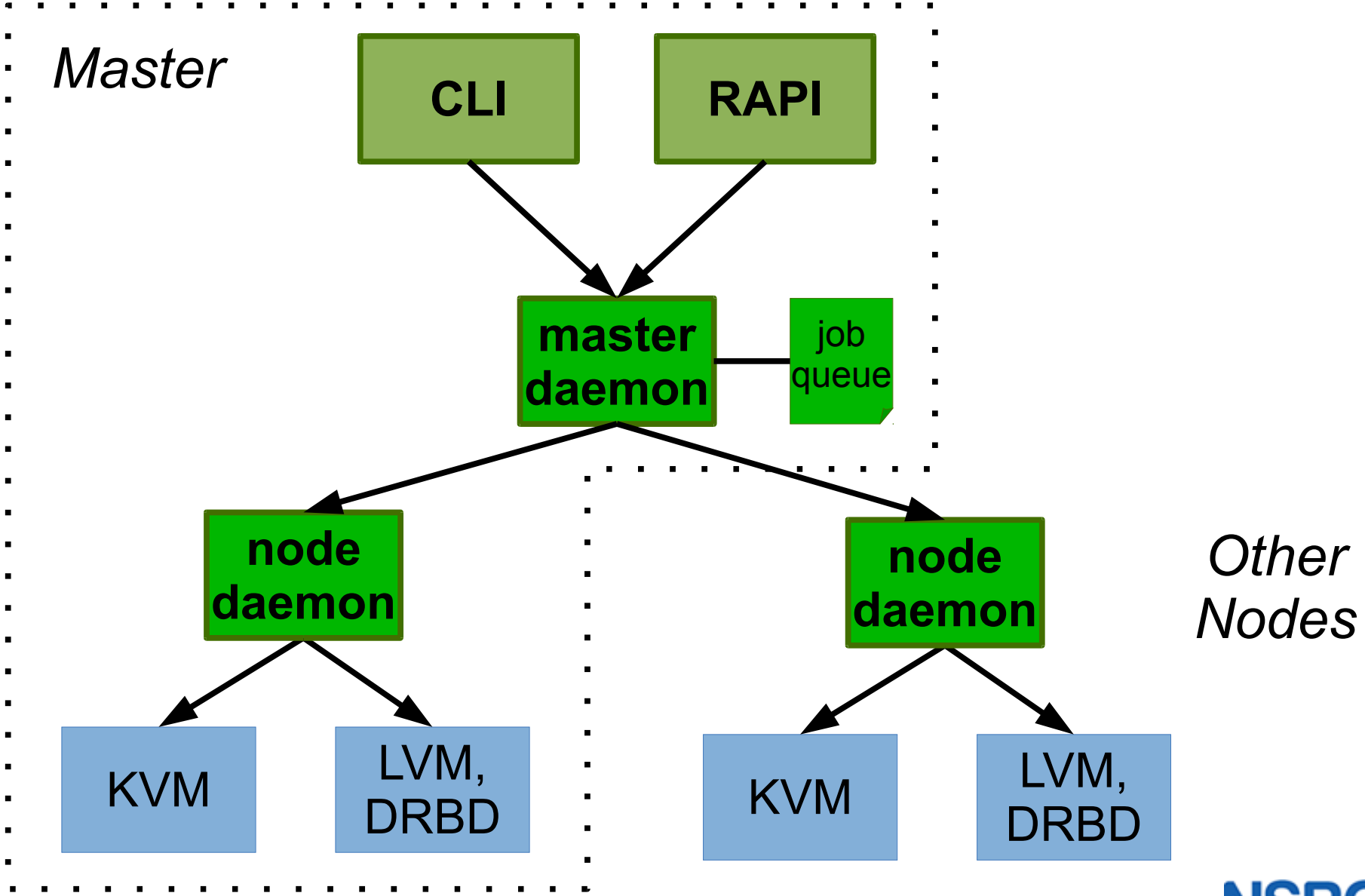


Cluster failover

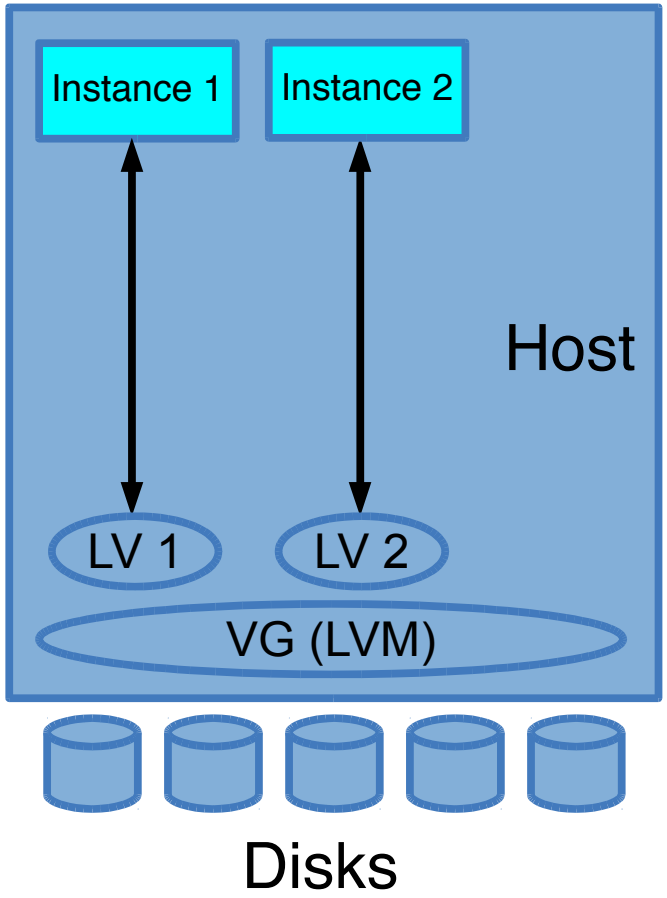
- All commands are executed on the master
- If the master fails, you can promote another node to be master
  - This is a manual event



Simplified architecture

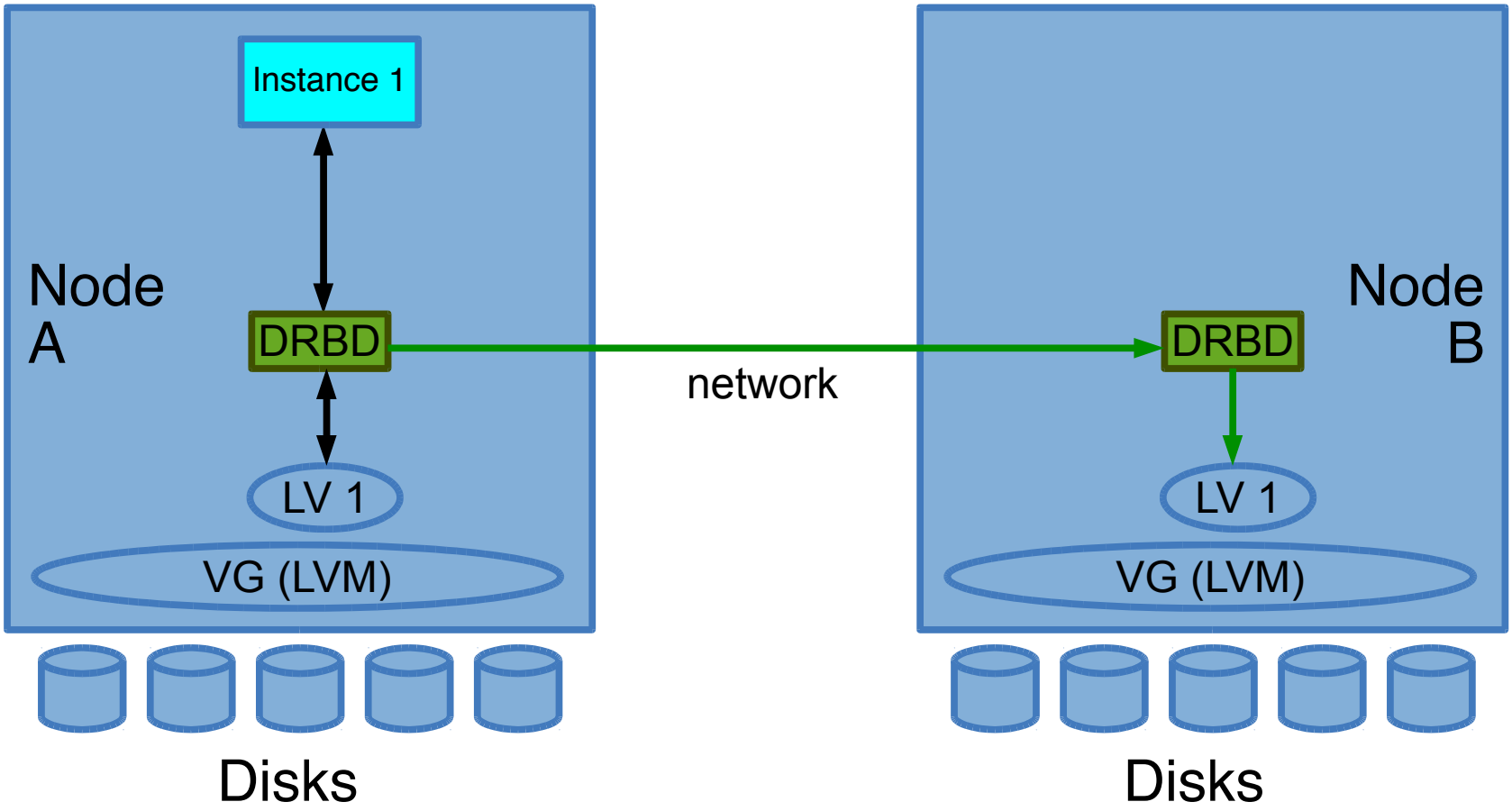


“Plain” instance



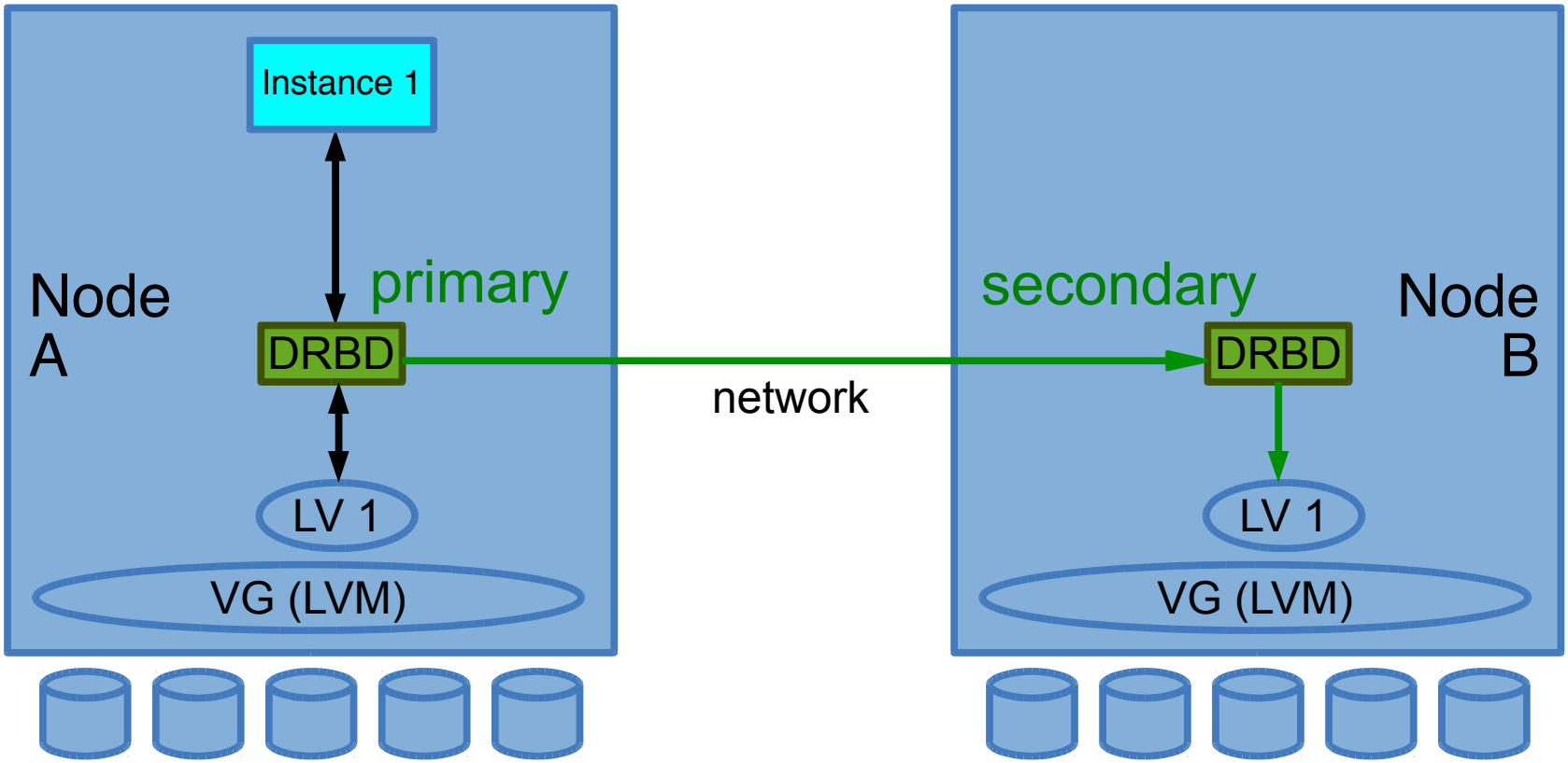
- Instance disks are Logical Volumes, stored on a Volume Group
- Can be expanded as required

“DRBD” instance



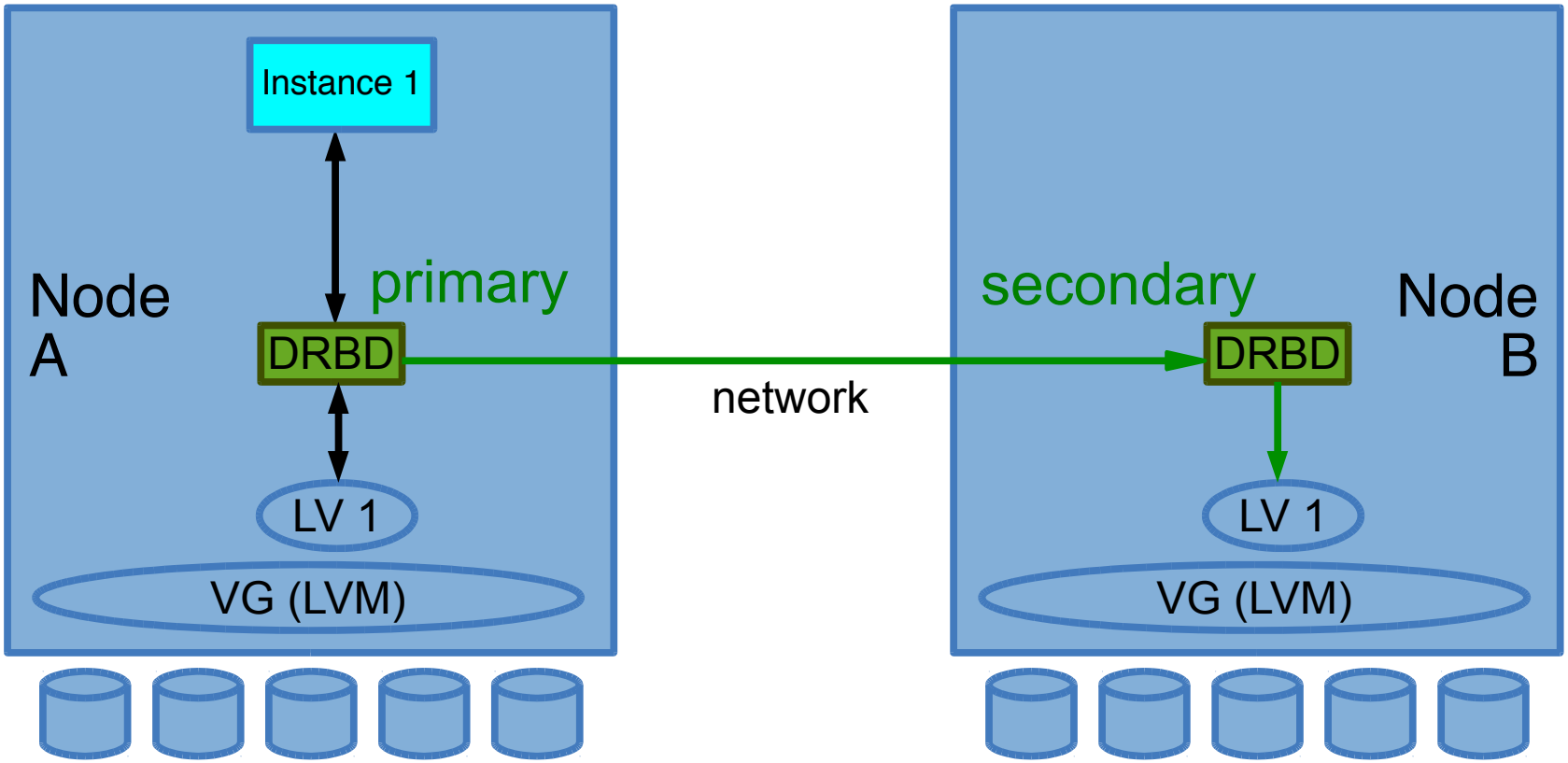
- Instance (Guest) does not use LV directly
- DRBD layer replicates disk to secondary node (B)
- Failover / migration becomes possible

DRBD: Primary & secondary



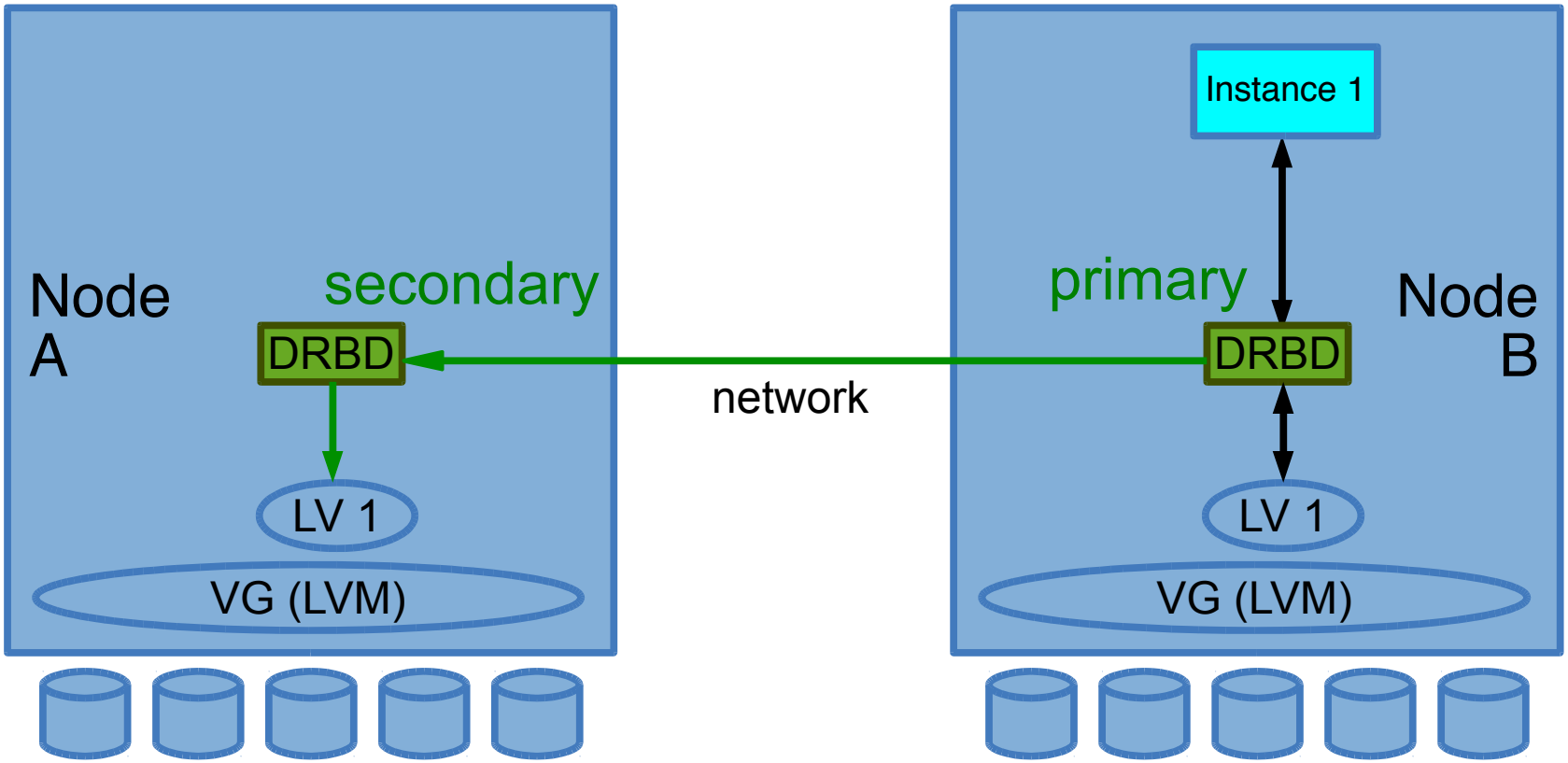
- Node A is said to be *“primary”* for Instance 1
- Node B is said to be *“secondary”* for Instance 1

DRBD: Migration

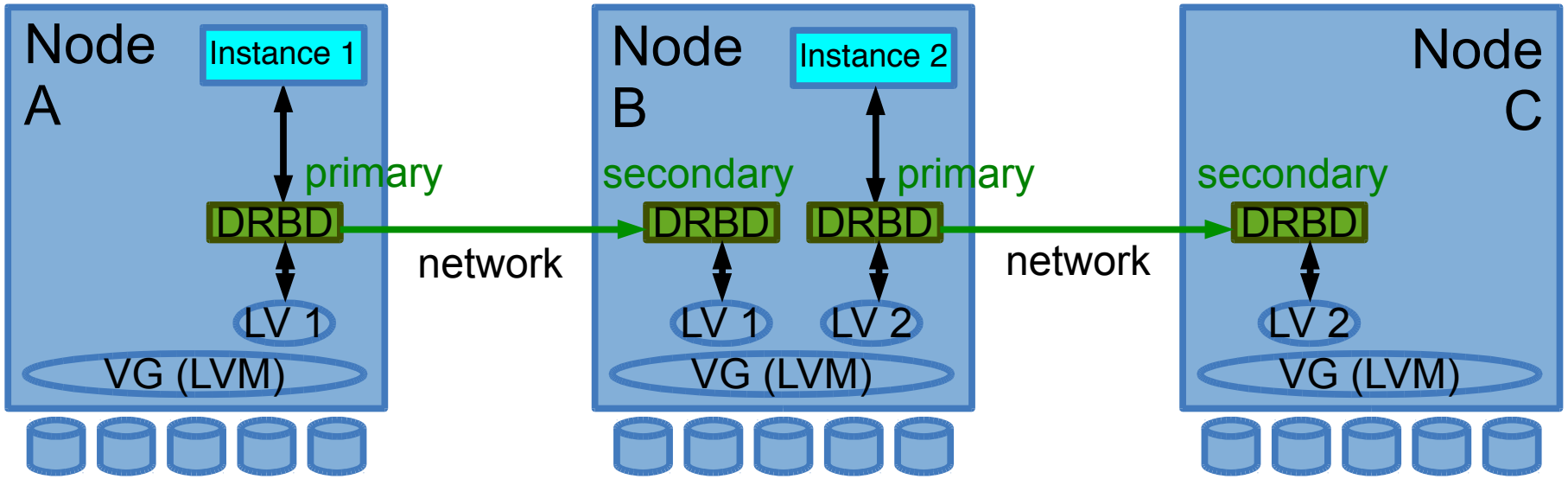


- Copy RAM + state of Instance 1 from A to B
- Pause Instance 1 on A
- Copy RAM again, reverse DRBD roles
- Resume Instance on B – B is now primary for Inst. 1

After migration

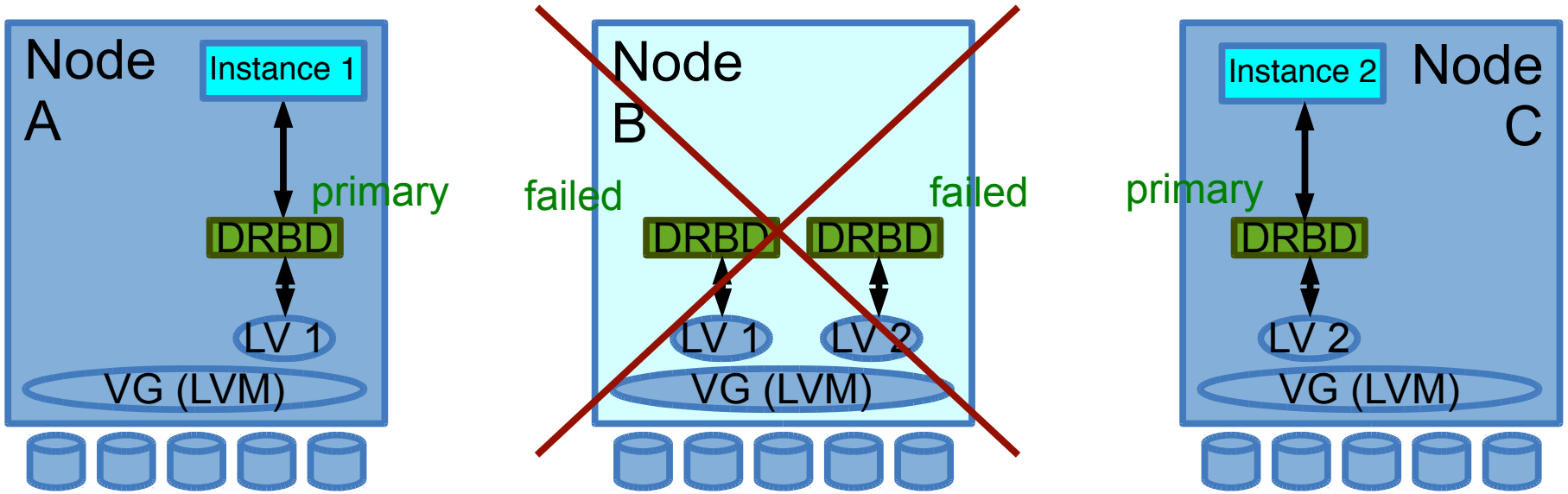


Primary & secondary



- A node can both be primary and secondary for different instances
  - Node A is *"primary"* for Instance 1
  - Node B is *"secondary"* for Instance 1
  - Node B is also *"primary"* for Instance 2
  - Node C is *"secondary"* for Instance 2

Unplanned node failure



- Instance 2 was running on node B
- It can be restarted on its secondary node (instance failover)
- This is a manual event

Available storage templates

- plain: logical volume
- drbd: replicated logical volume
- file: local raw disk image file
- sharedfile: disk image file over NFS etc
- blockdev: any pre-existing block device
- rbd: Ceph (RADOS) distributed storage
- diskless: no disk (e.g. live CD)
- ext: pluggable storage API

Networking

- You need one management IP address per node, plus one cluster management IP
  - all on the same management subnet
- Optional: separate replication network for drbd traffic
- Optional: additional network(s) for guests to connect to
  - so they cannot see the cluster nodes
  - reduces the impact of a DoS attack on a guest VM

Ganeti Scaling

- Start with just one or two nodes!
- Recommended you limit cluster to 40 nodes (approx. 1000 instances)
- Beyond this you can just build more clusters

Exercise

- Build a Ganeti cluster
- Working in groups
- 3 or 4 nodes in each cluster