LISA 2007

# Ganeti

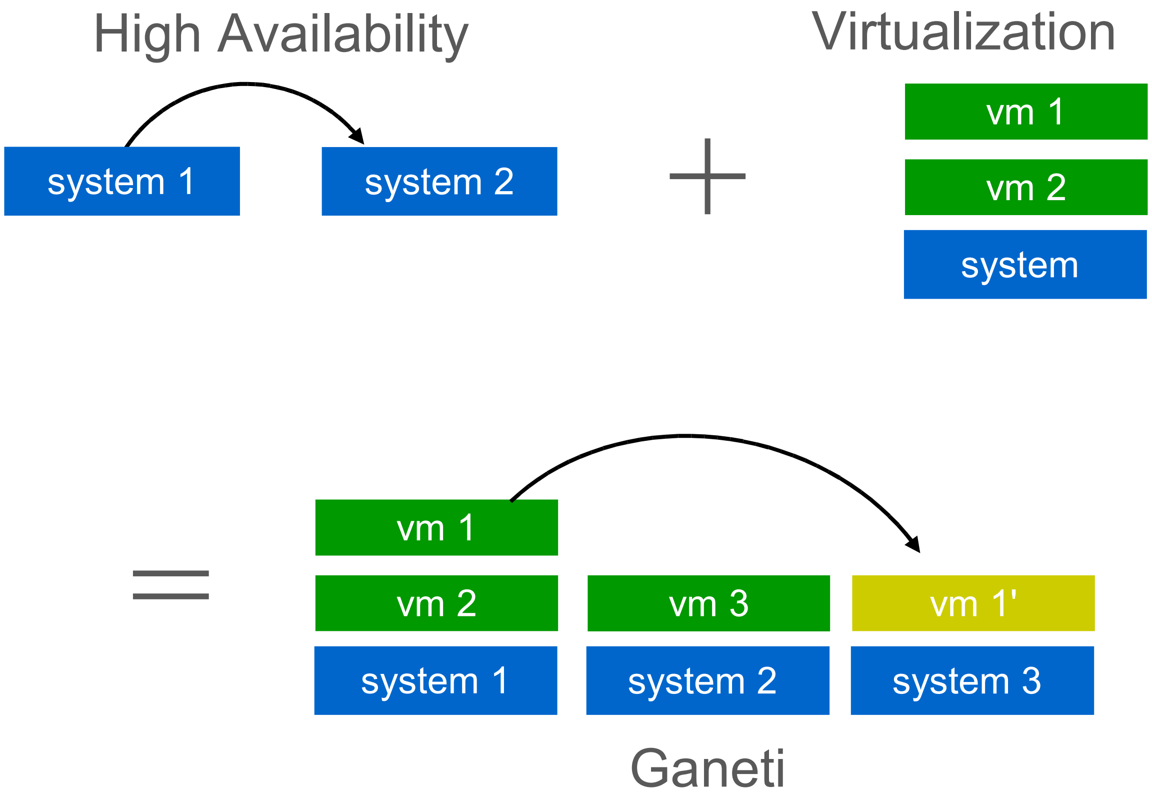
an open source high-availability cluster based on Xen

**Guido Trotter**
**Google Ganeti Team**

# Content

- Design goals and principles
- Ganeti overview and administration
- Ganeti failover details
- Ganeti usage in Google
- Ganeti roadmap
- Live Demo

# Ganeti at a glance



# Design goals and principles

- goals
  - increase availability
  - reduce hardware cost
  - increase flexibility
  - transparency

- principles
  - not dependent on specific hardware (e.g. SAN)
  - support different host systems
  - scales linearly with the number of systems
  - small, iterative development

# Ganeti overview (1/3)

Ganeti is a software to manage clusters of virtual servers

- Based on Xen (but not strictly dependent on it)
- n-node high-availability cluster (future)
- makes it simple to manage 10s of nodes and 100s of instances
- software used
  - language: Python
  - virtualization: Xen
  - disk management: LVM / DRBD / MD
  - RPC: Twisted, ssh

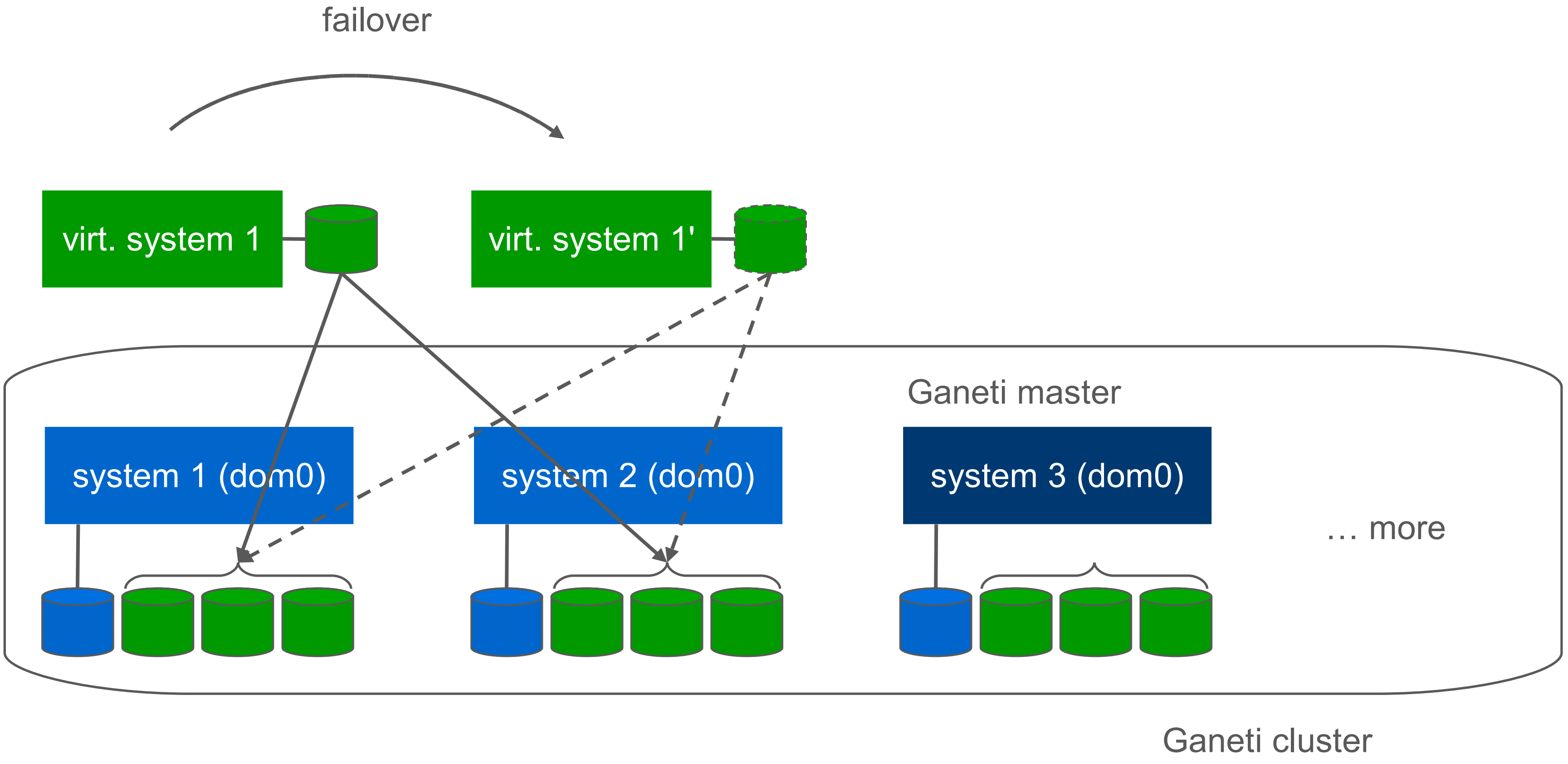
# Ganeti overview (2/3)

Terminology:

- Cluster
- Node
- Master Node
- Instance
- Pool
- Meta-Cluster

# Ganeti overview (3/3)

Xen dom0 = node
Xen domU = instance



# Ganeti administration (1/4)

The commands are run on the master node

- `gnt-node`: add / remove / list cluster nodes
- `gnt-instance`:
  - add / remove instance
  - failover instance, change secondary
  - stop / start instance, change parameters
- `gnt-os`: instance OS definitions
- `gnt-cluster`: cluster commands
- `gnt-backup`: instance export and import

All commands have man pages and support interactive help.

# Ganeti administration (2/4)

Cluster Setup:

```
node0# gnt-cluster init mycluster
node0# gnt-node add node1
node0# gnt-node add node2
node0# gnt-node add node3
node0# gnt-cluster command \
> apt-get install ganeti-instance-etch
```

# Ganeti administration (3/4)

Creation of an instance:

```
node0# gnt-instance add \
> -n node2:node1 \
> -t drbd8 \
> instance0
```

Migration after a node crash:

```
node0# gnt-instance failover --ignore-consistency instance0
node0# gnt-instance replace-disks -s \
> --new-secondary=node3 instance0
```

# Ganeti administration (4/4)

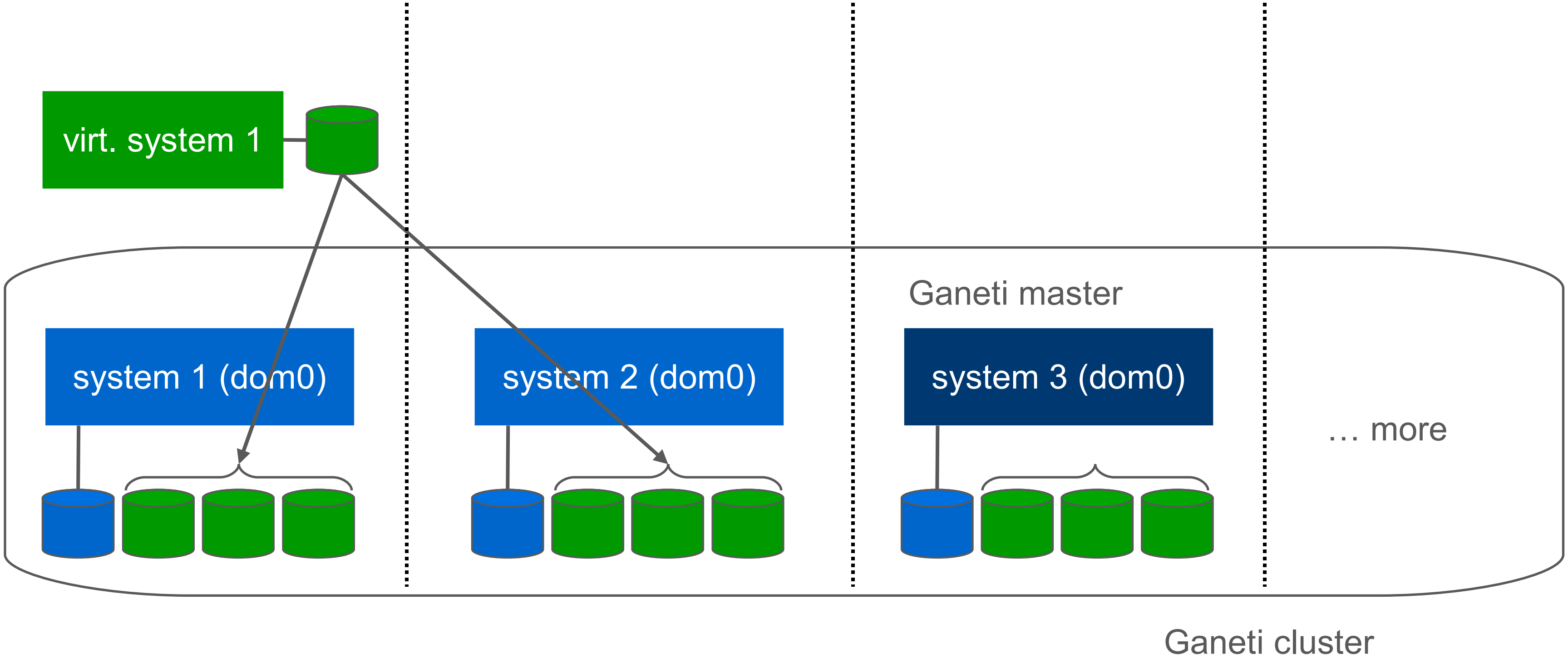
## Cluster status:

```
# gnt-instance list
Instance              OS   Primary_node      Autostart Status  Memory
instance1.example.com etch node1.example.com yes       running    128
instance2.example.com etch node3.example.com yes       running    512
instance3.example.com etch node3.example.com yes       running   1024
instance4.example.com etch node2.example.com yes       running    128
instance5.example.com etch node4.example.com yes       running    512

# gnt-node list
Node                  DTotal  DFree MTotal MNode MFree Pinst Sinst
node1.example.com     858240 442752   4095   511  3456     1     2
node2.example.com     572160 567296   4095   511  3456     1     2
node3.example.com     858240 858240   4095   511  2048     2     1
node4.example.com     356032 356032   4095   511  3072     1     0
```

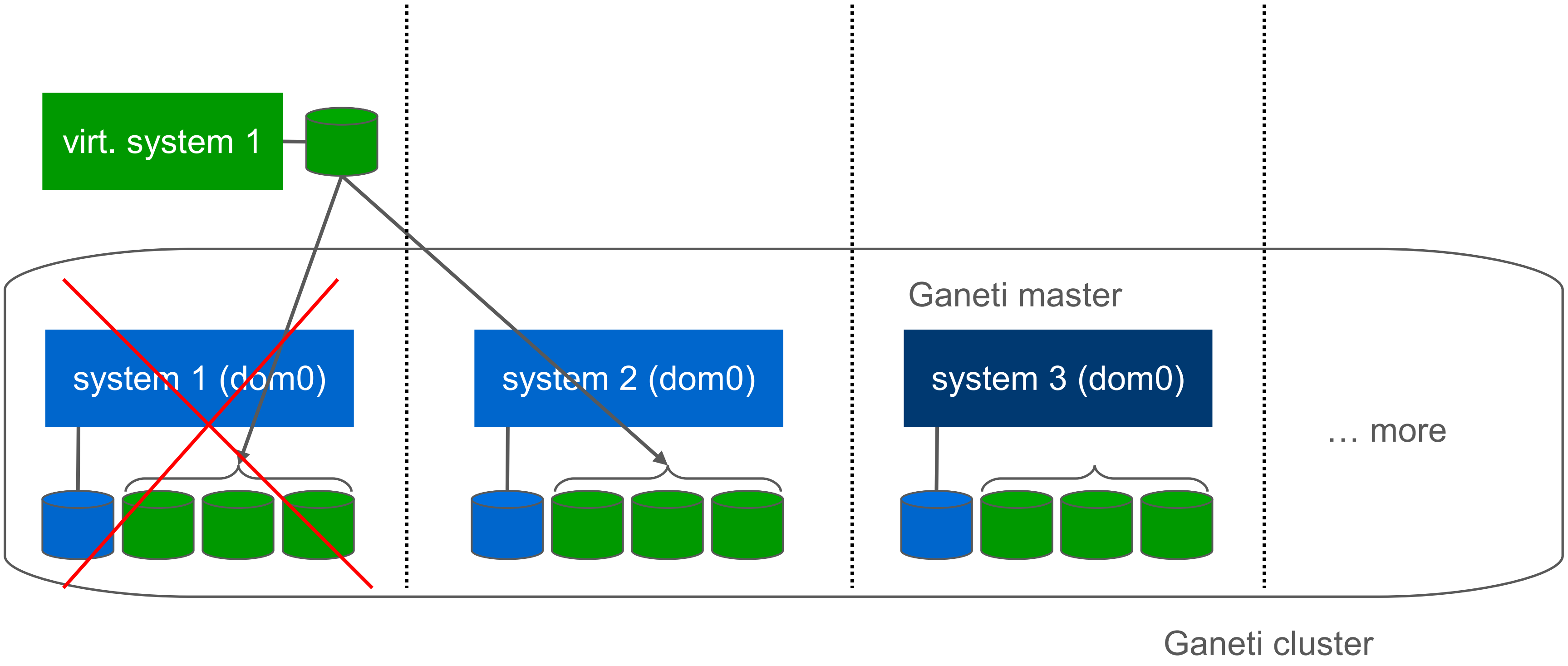
# Instance failover (1/4)

Xen dom0 = node
Xen domU = instance

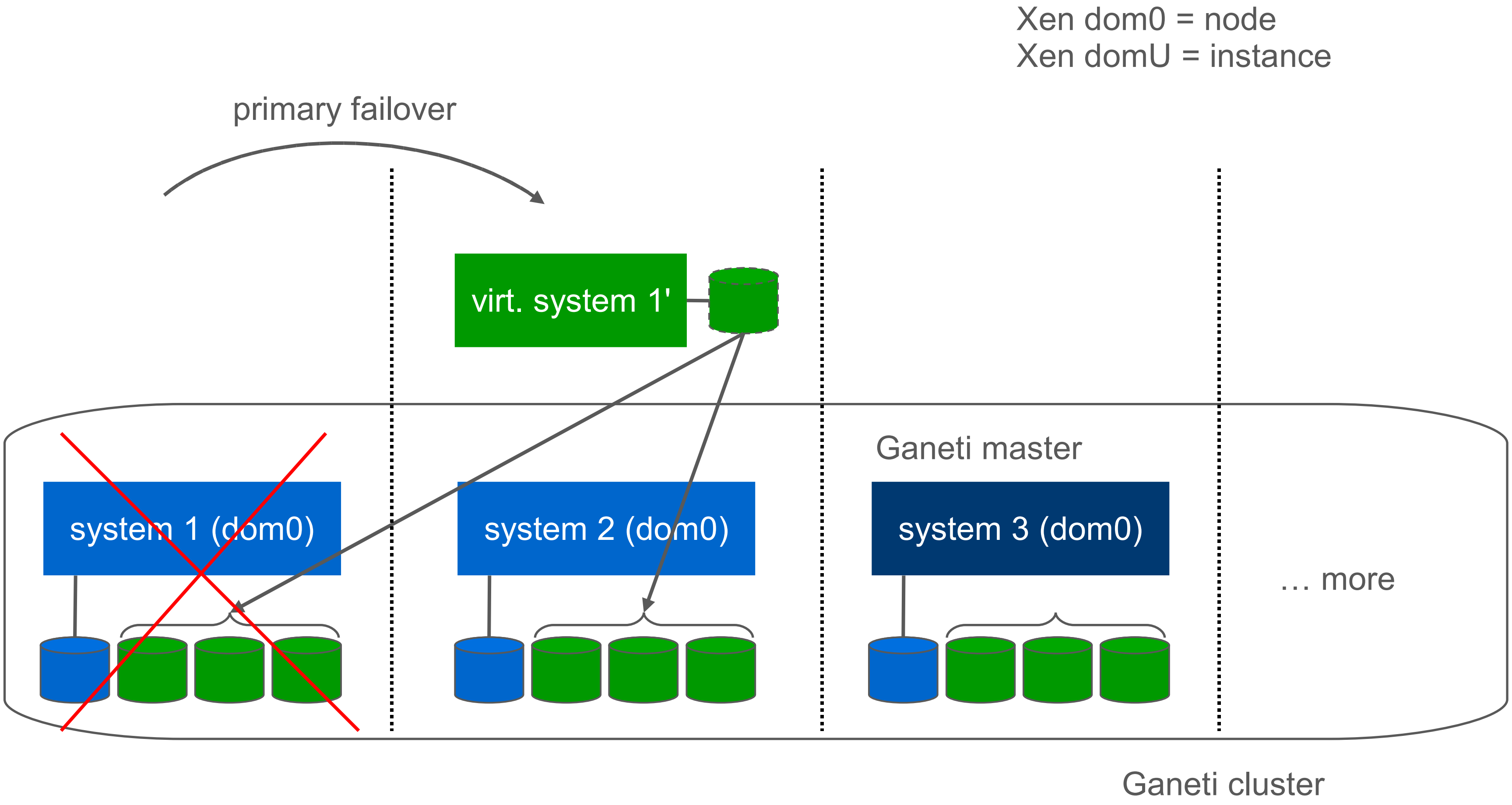


# Instance failover (2/4)

Xen dom0 = node
Xen domU = instance

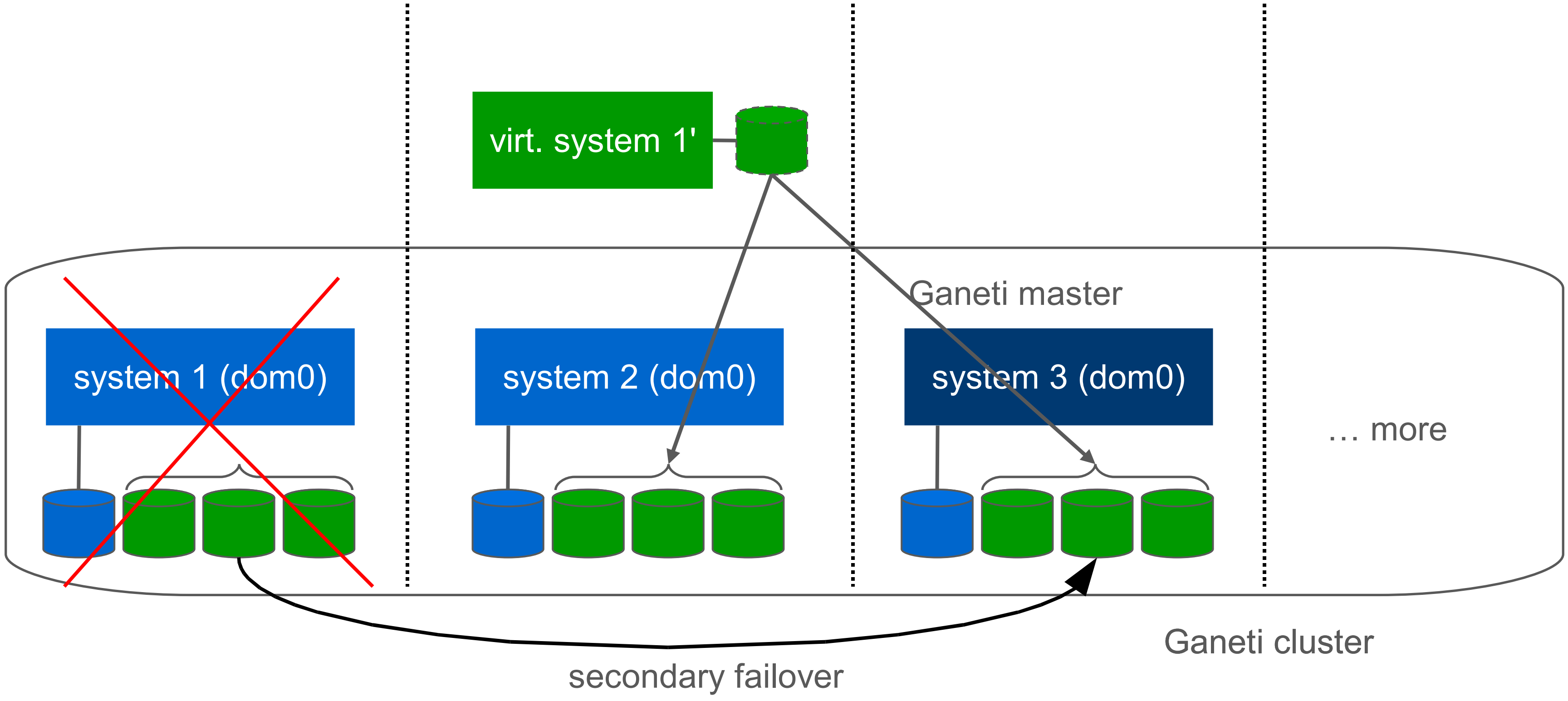


# Instance failover (3/4)



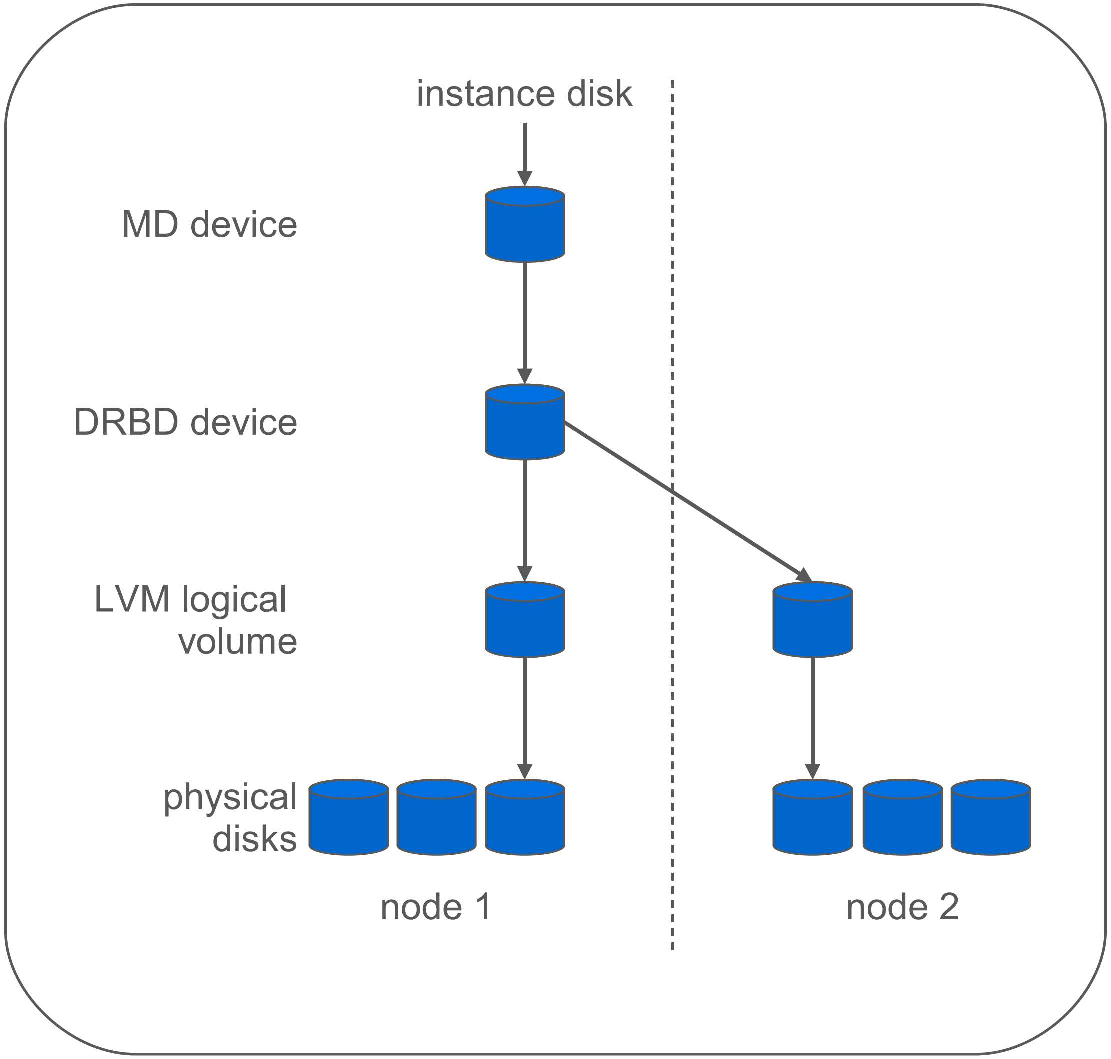
# Instance failover (4/4)

Xen dom0 = node
Xen domU = instance



# Ganeti disk details

- disk types
    - plain
    - local_raid1
    - remote_raid1
    - **drbd8 (new)**

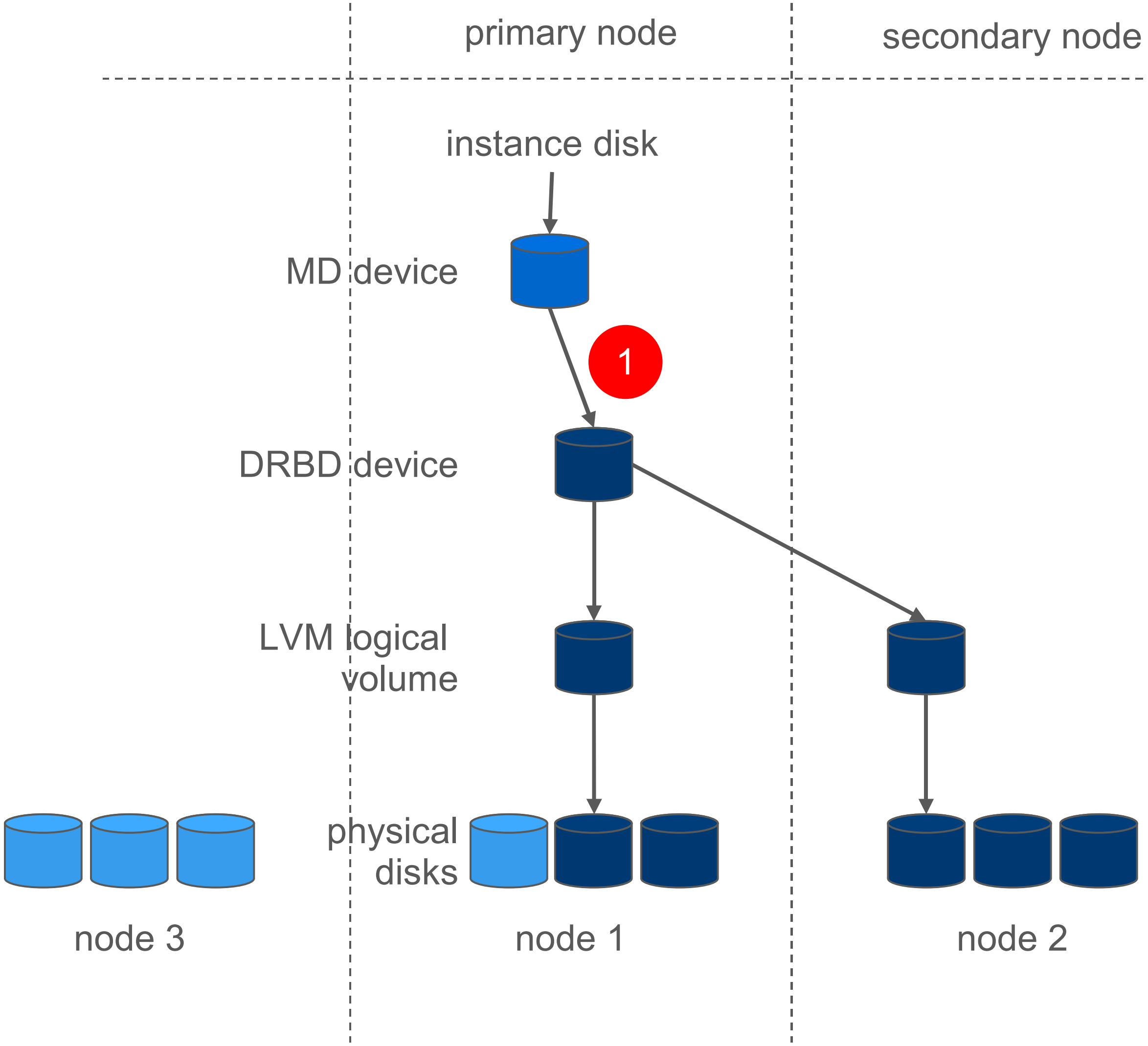


remote_raid1 details

# Ganeti remote_raid1 disk recovery

remote_raid1 failover

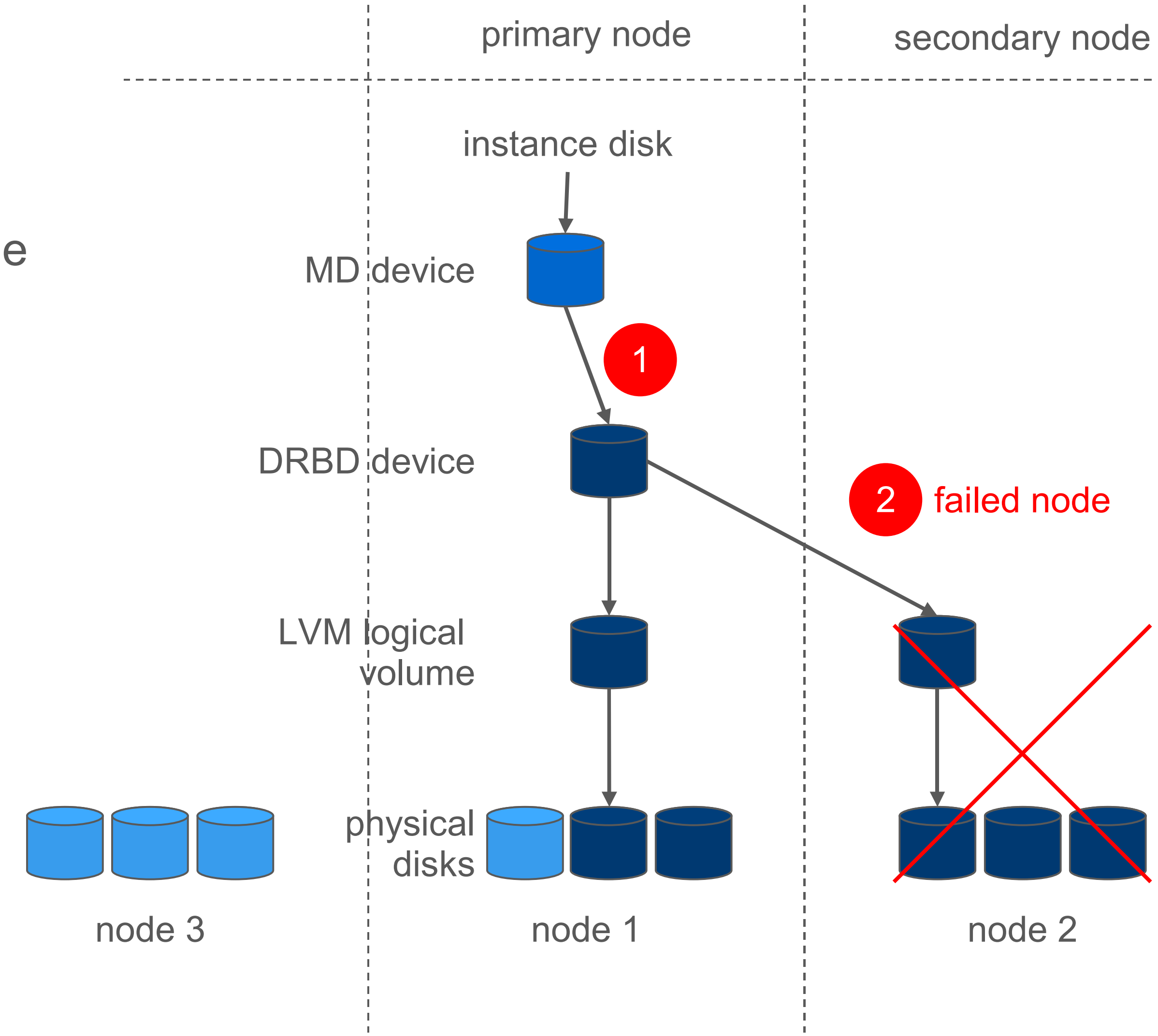
1. dark blue DRDB set serves data



# Ganeti remote_raid1 disk recovery

remote_raid1 failover

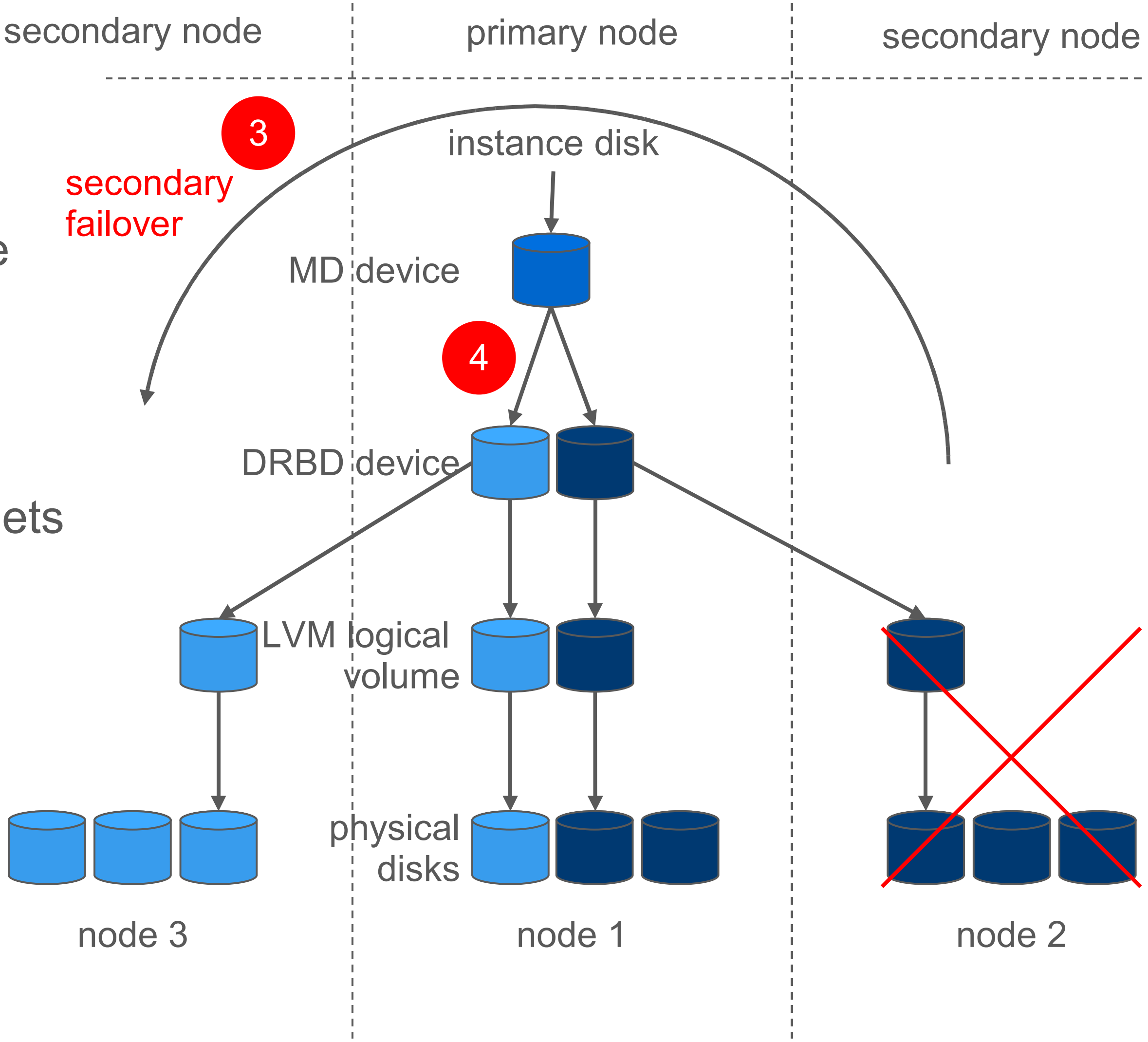
1. dark blue DRDB set serves data
2. node fails in dark blue DRDB set



# Ganeti remote_raid1 disk recovery

remote_raid1 failover

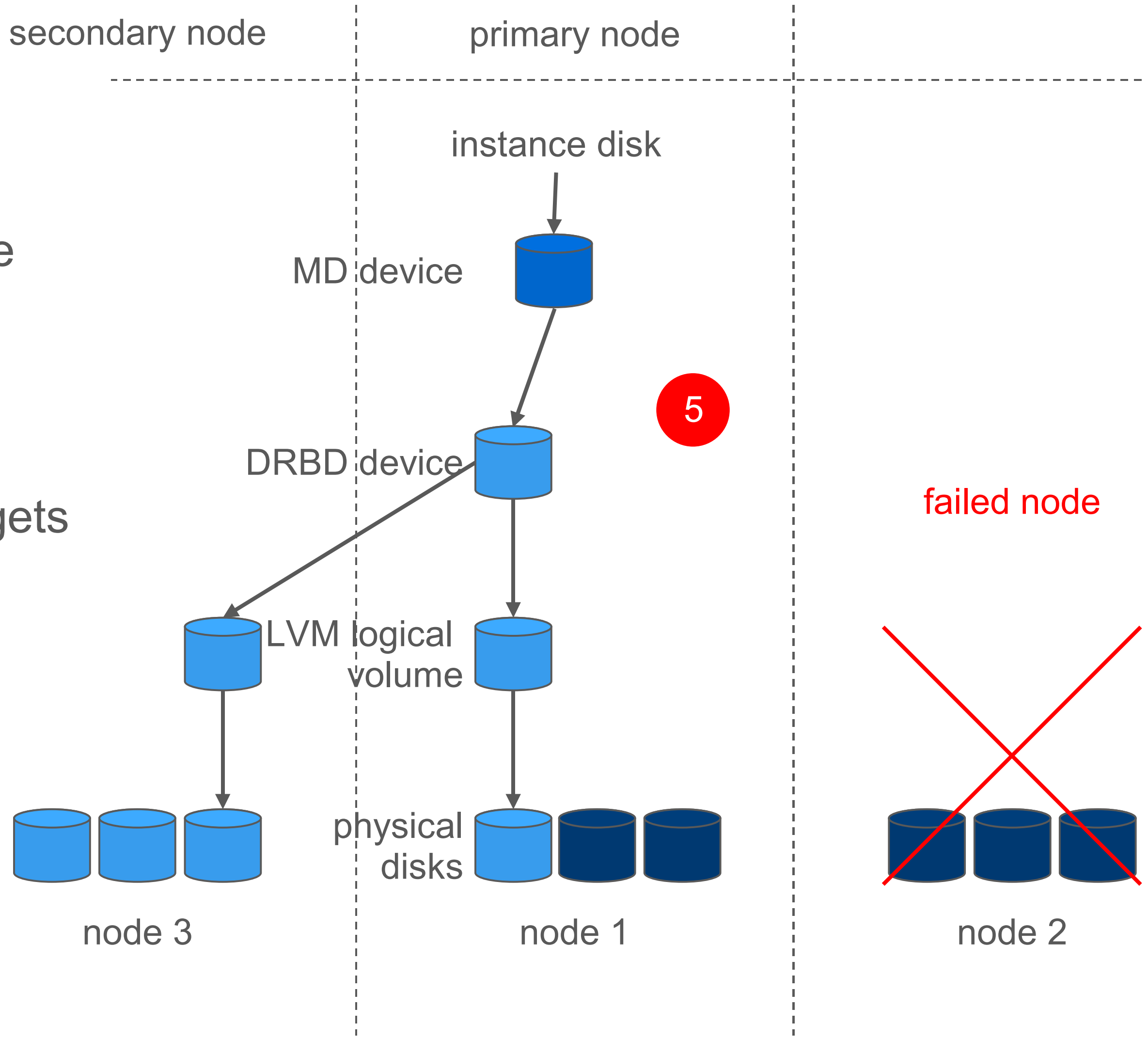
1. dark blue DRDB set serves data
2. node fails in dark blue DRDB set
3. admin: gnt-instance replace-disks
4. light blue DRDB set gets added and is synchronized



# Ganeti remote_raid1 disk recovery

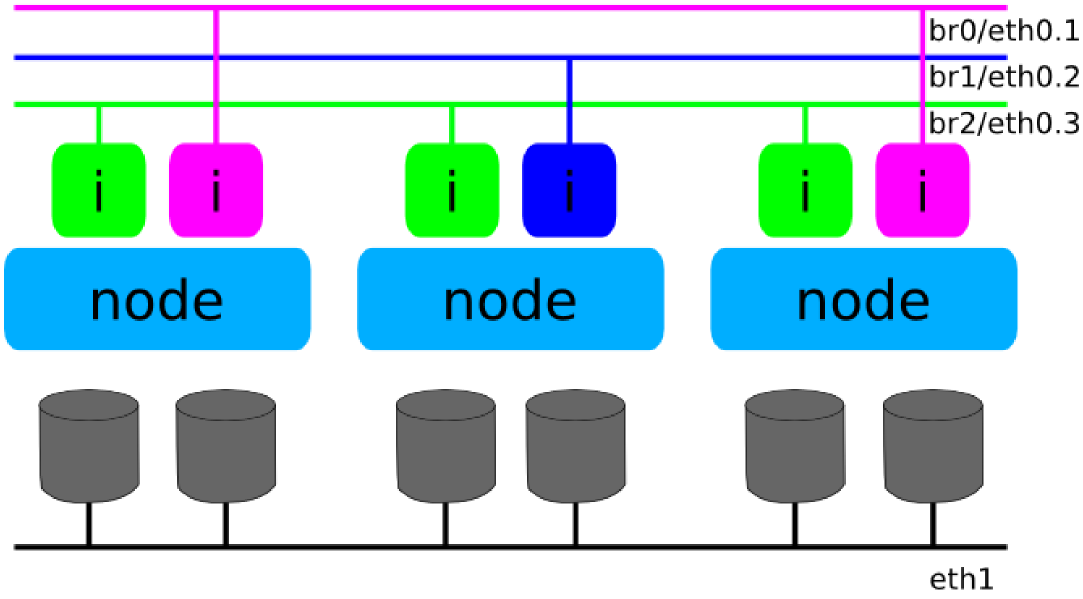
remote_raid1 failover

1. dark blue DRDB set serves data
2. node fails in dark blue DRDB set
3. admin: gnt-instance replace-disks
4. light blue DRDB set gets added and is synchronized
5. dark blue DRDB set gets removed

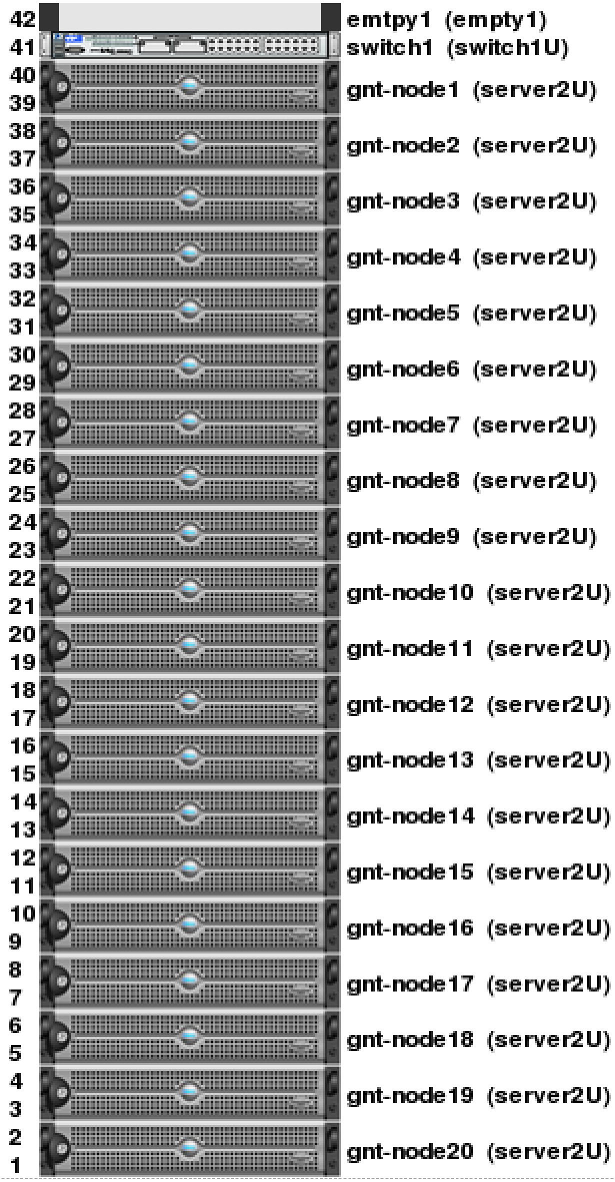


# Optional advanced features

- Separate replication network
- Multiple bridges/VLAN support
- **Tagging (new)**



# Ganeti usage in Google



- 20-node Ganeti cluster
- 64-bit node OS
- 80 virtual instances
- used for internal systems
- **not** used for google.com
- best for non-resource intensive systems

# Ganeti code

- developed at Google
- license: GPLv2
- code location: http://code.google.com/p/ganeti/
- August 2007
  - open source and release 1.2b1
- November 2007
  - release 1.2b2
- December 2007
  - release 1.2
- February 2008
  - release 1.2.1
- Later
  - release 1.3

# 1.2 Roadmap

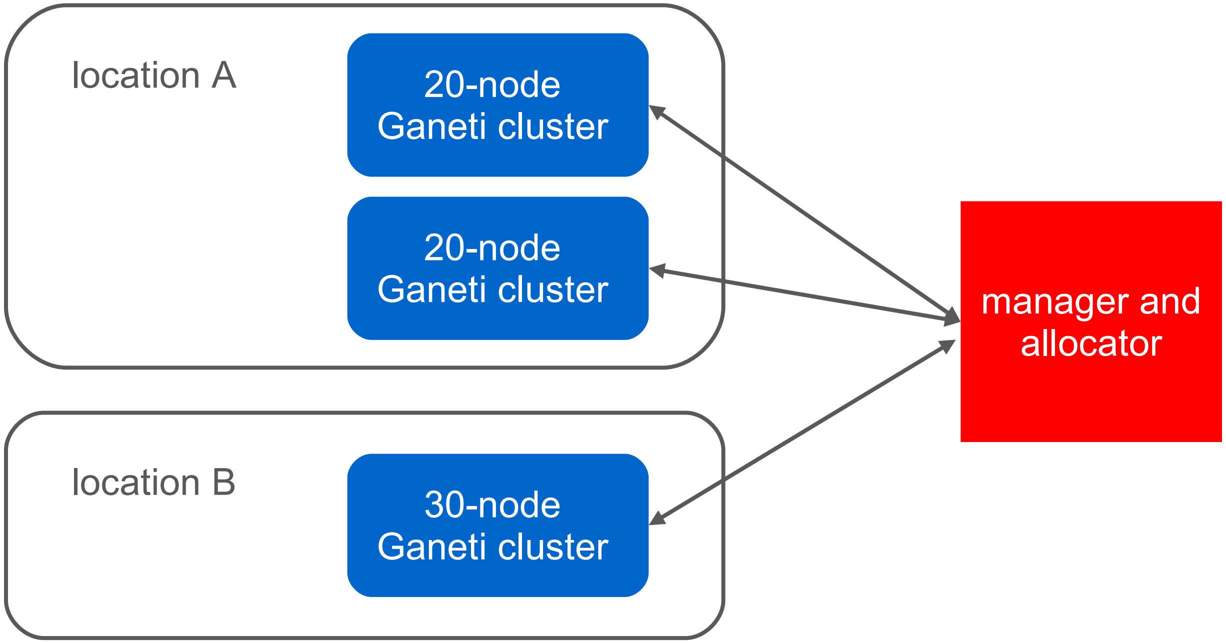
- Release 1.2b2:
  - new cluster configuration format
  - drbd8 disk template
  - simplify common tasks (node evacuation, reboot, tags)
  - ganeti-watcher now reactivates drbd pairs
  - easier packaging experience
  - tags
- Release 1.2:
  - no more new features
  - code cleanup and bugfixes
- Future point releases:
  - only features that do not affect the core code
  - investigate experimental support for KVM and Xen-HVM

# 1.3 Draft Roadmap

- External API
- Transparent failover
- Granular locking
- Job Queuing
- Support for more diverse instances
- Stable support for different virtualization technologies

# The Future

- automatic instance failover
- automatic node allocation
- master node election
- manager GUI / meta-cluster manager



# Demo and Q&A