

# LinuxConf Europe 2007



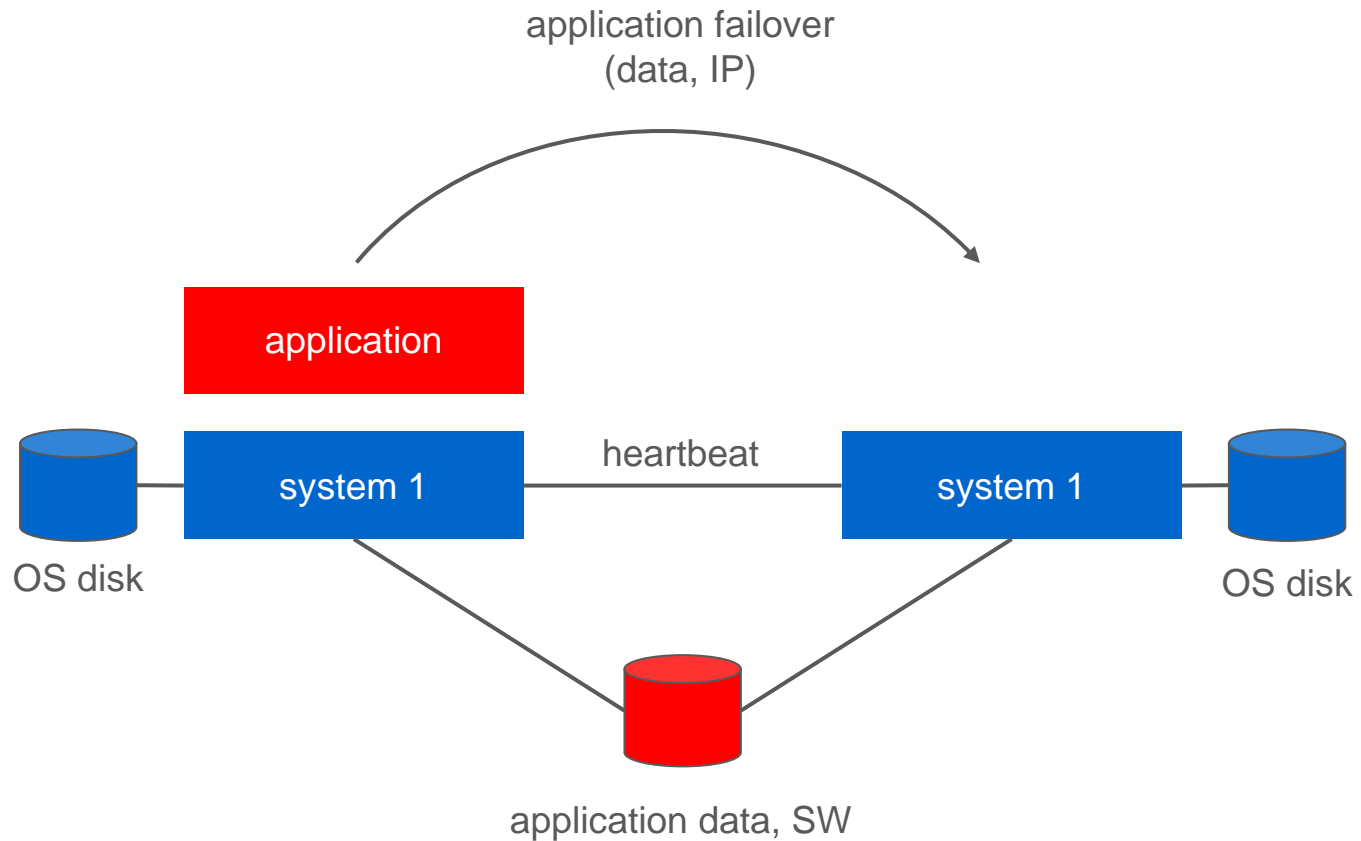
Ganeti

an open source multi-node high-availability cluster based on Xen

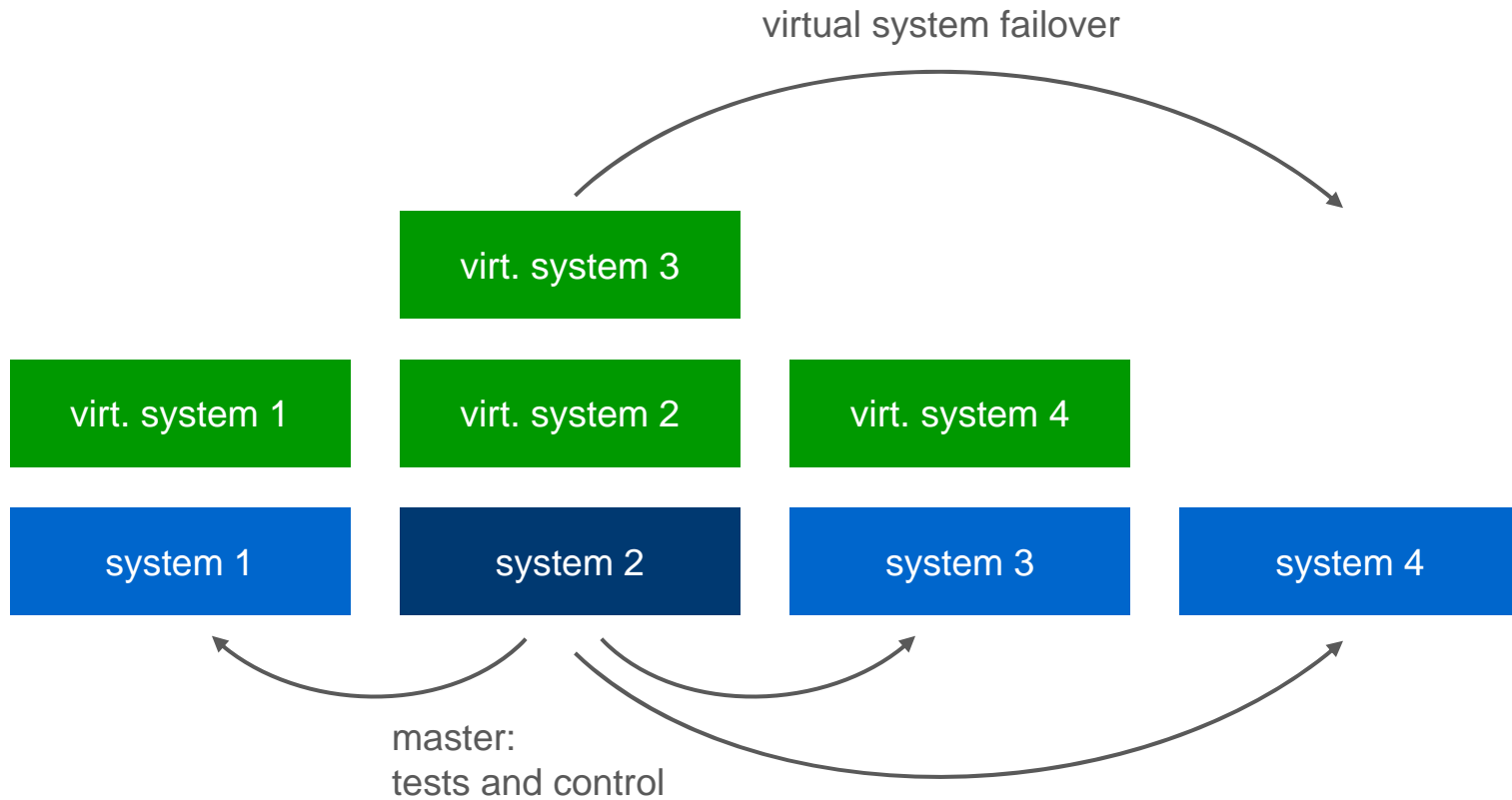
Roman Marxer

- Traditional vs. virtualization-based high-availability clusters
- Ganeti overview and administration
- Ganeti disk details and recovery
- Design goals and principles
- Ganeti usage in Google
- Ganeti code
- Roadmap

# Traditional high-availability cluster



# Virtualization-based multi-node high-availability cluster



# Ganeti overview (1/2)

---

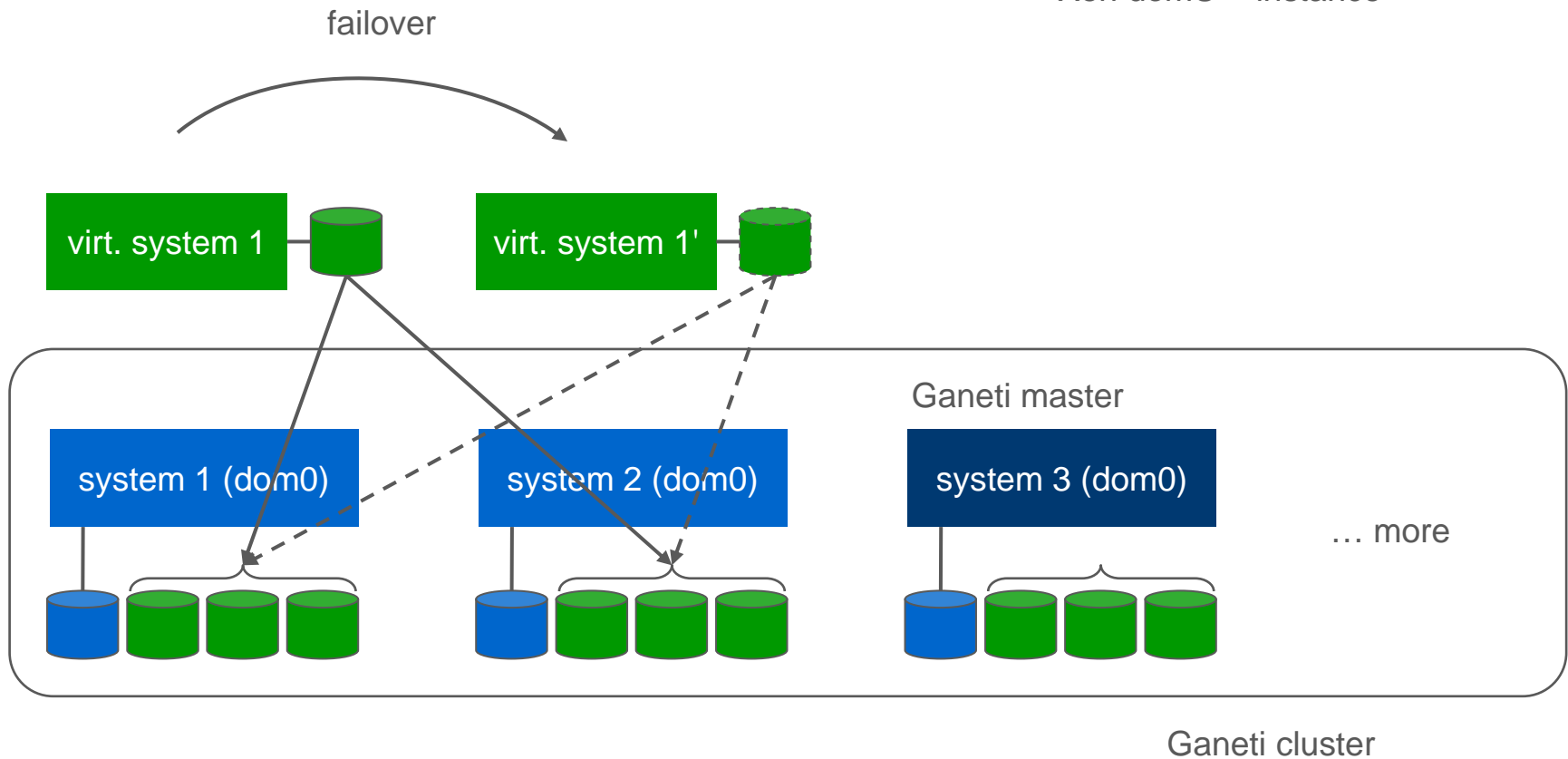


- Xen cluster manager
- n-node high-availability cluster (future)
- software used
  - virtualization: Xen
  - disk management: LVM / DRBD / MD
  - language and RPC: Python / Twisted

# Ganeti overview (2/2)



Xen dom0 = node  
Xen domU = instance



- Ganeti master (special role of a 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-instance add` (instance creation)
  - parameters: name, disk size, RAM size, OS type, disk layout, primary / secondary node
- OS creation script
  - script to install an OS in a partition
    - e.g. debootstrap and additional scripts
  - set IP address and hostname of the image
  - copy image and change

# Ganeti administration (3/3)



```
# gnt-instance list
```

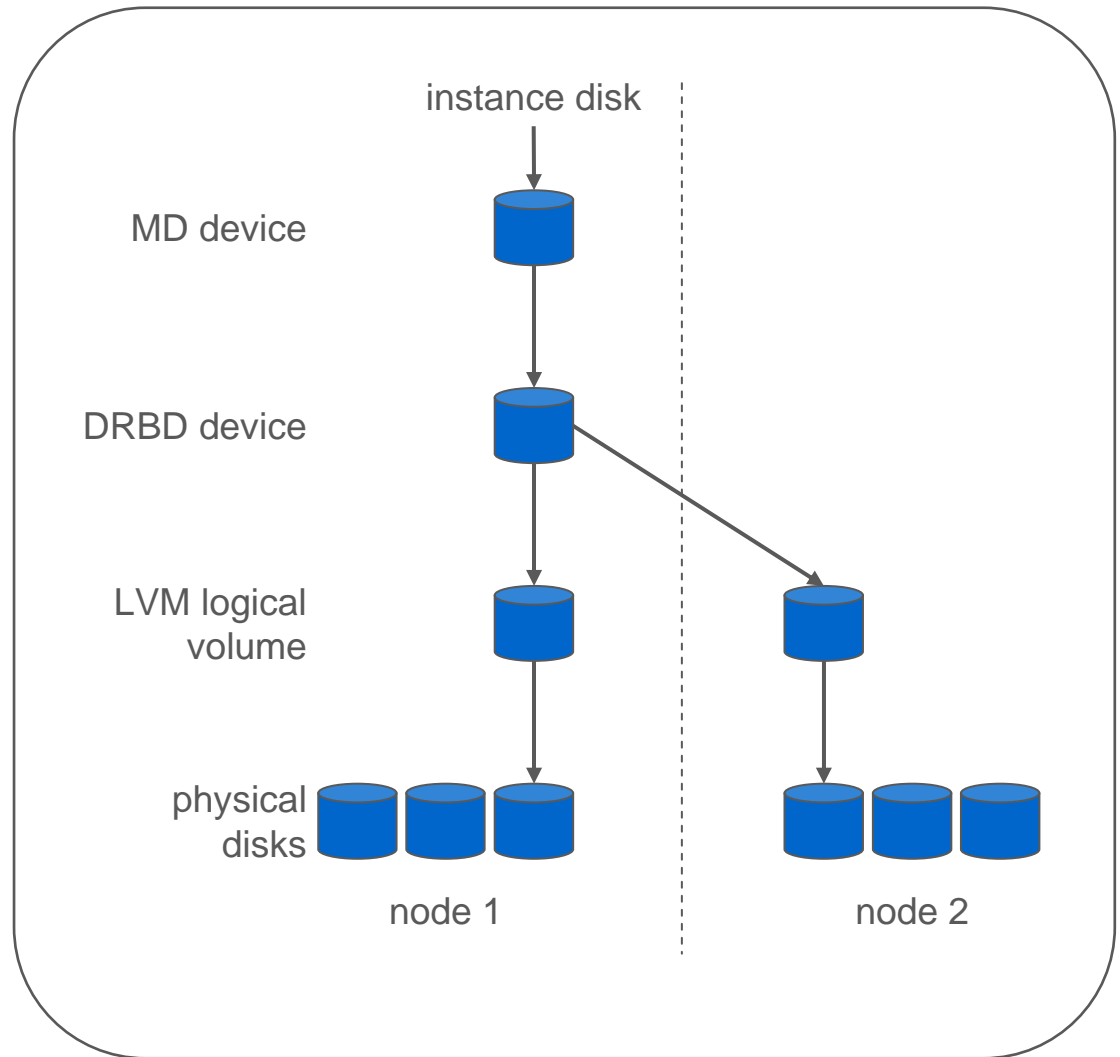
Instance	OS	Primary_node	Autostart	Status	Memory
instance1.example.com	etch	node1.example.com	yes	running	128
instance2.example.com	etch	node3.example.com	yes	running	512
instance3.example.com	etch	node3.example.com	yes	running	1024
instance4.example.com	etch	node2.example.com	yes	running	128
instance5.example.com	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

# Ganeti disk details

- disk types
  - plain
  - local\_raid1
  - remote\_raid1

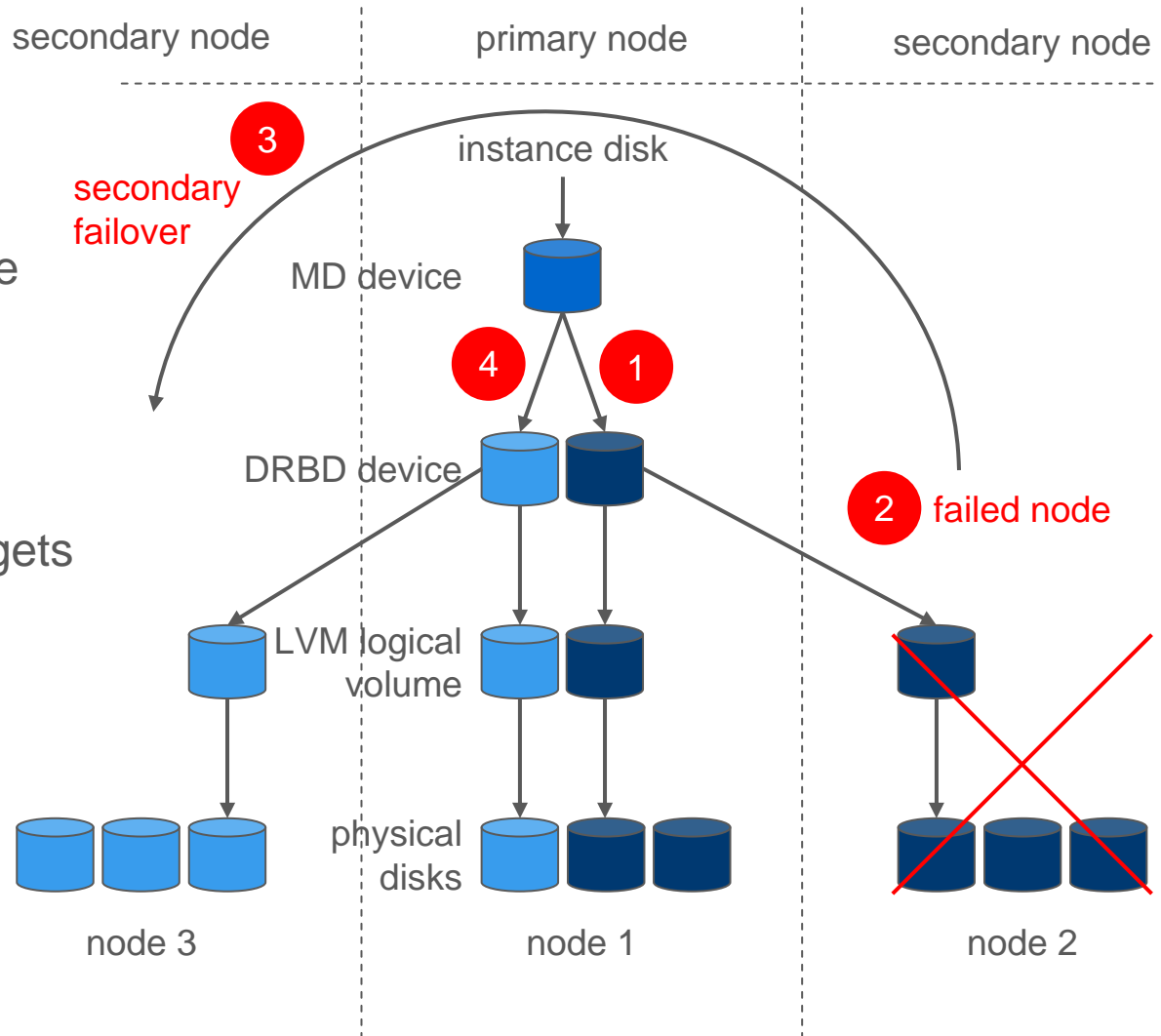


remote\_raid1 details

# Ganeti 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



- goals
  - increase availability
  - reduce hardware cost
  - increase flexibility
  - transparency
  
- principles
  - not dependent on specific hardware (e.g. SAN)
  - scales linearly with the number of systems

# Ganeti usage in Google

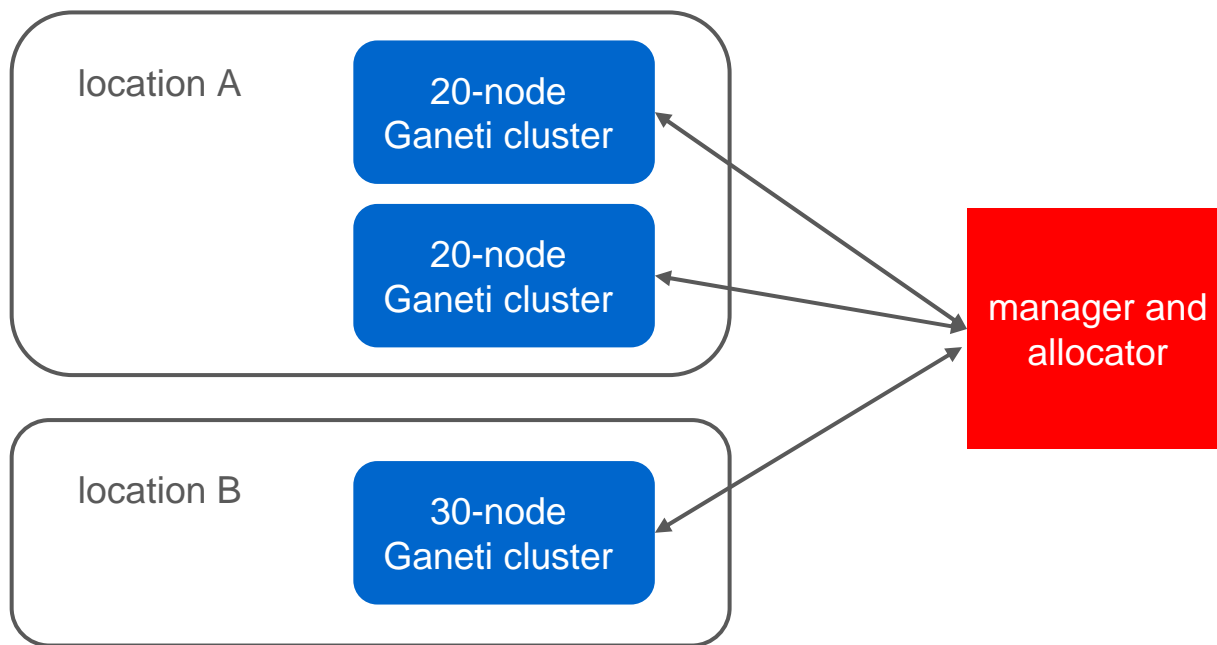


42		empty1 (empty1)
41		switch1 (switch1U)
40		
39		gnt-node1 (server2U)
38		gnt-node2 (server2U)
37		gnt-node3 (server2U)
36		gnt-node4 (server2U)
35		gnt-node5 (server2U)
34		gnt-node6 (server2U)
33		gnt-node7 (server2U)
32		gnt-node8 (server2U)
31		gnt-node9 (server2U)
30		gnt-node10 (server2U)
29		gnt-node11 (server2U)
28		gnt-node12 (server2U)
27		gnt-node13 (server2U)
26		gnt-node14 (server2U)
25		gnt-node15 (server2U)
24		gnt-node16 (server2U)
23		gnt-node17 (server2U)
22		gnt-node18 (server2U)
21		gnt-node19 (server2U)
20		gnt-node20 (server2U)
19		
18		
17		
16		
15		
14		
13		
12		
11		
10		
9		
8		
7		
6		
5		
4		
3		
2		
1		

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

- developed at Google
- license: GPL v2
- code location: <http://code.google.com/p/ganeti/>
- August 2007
  - beta release and open source
- November 2007
  - release v.1.2
  - development contributions possible

- HVM integration (Windows support)
- automatic instance failover / node allocation
- master node election
- manager GUI / instance allocator



# Thank You!

---

Q&A

