

# Ganeti

## The Cluster-based Virtualization Management Software

Helga Velroyen (helgav@google.com)

Klaus Aehlig (aehlig@google.com)

August 24, 2013





## Virtualization

To build your VMs (“instances”),  
you would take ...

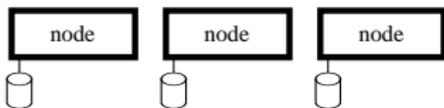




## Virtualization

To build your VMs (“instances”),  
you would take ...

- a bunch of physical machines (“nodes”)

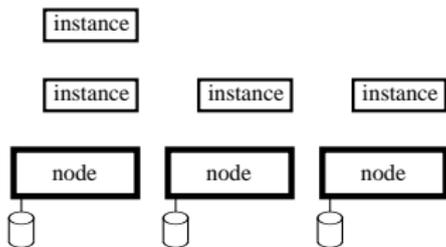




## Virtualization

To build your VMs (“instances”),  
you would take ...

- a bunch of physical machines (“nodes”)
- some hypervisor, say Xen

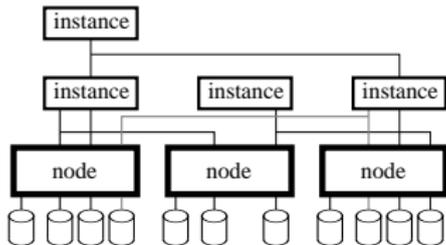




## Virtualization

To build your VMs (“instances”), you would take ...

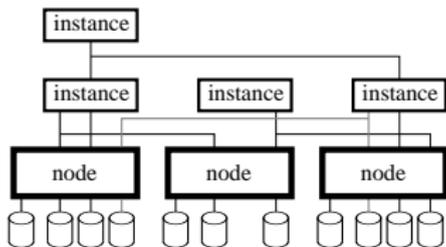
- a bunch of physical machines (“nodes”)
- some hypervisor, say Xen
- some way to replicate storage, say DRBD





## Enter Ganeti

While all this works on its own,  
Ganeti helps

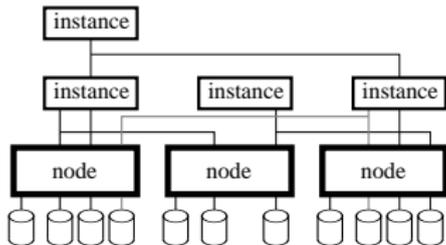




## Enter Ganeti

While all this works on its own,  
Ganeti helps

- to get there
  - uniform interface

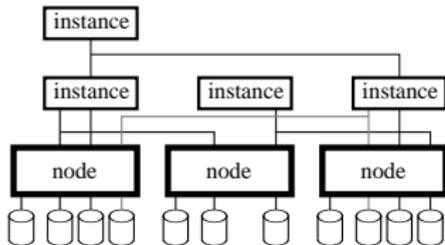




## Enter Ganeti

While all this works on its own,  
Ganeti helps

- to get there
  - uniform interface



- Hypervisors: Xen, kvm, ...
- Storage: drbd, lvm, file, ...
- Network

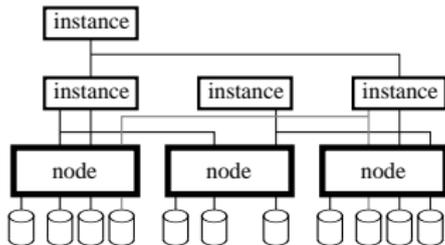




## Enter Ganeti

While all this works on its own,  
Ganeti helps

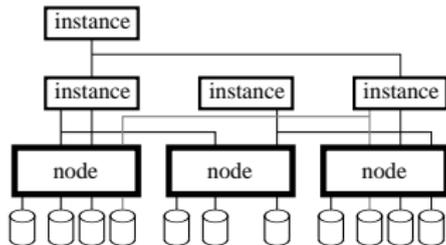
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation



## Enter Ganeti

While all this works on its own,  
Ganeti helps

- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation



- *Instance memory/disk size*
- *CPU oversubscription*
- *tag-exclusion*  
*"Don't put both name servers on the same node!"*

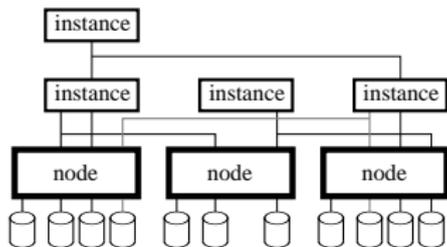




## Enter Ganeti

While all this works on its own,  
Ganeti helps

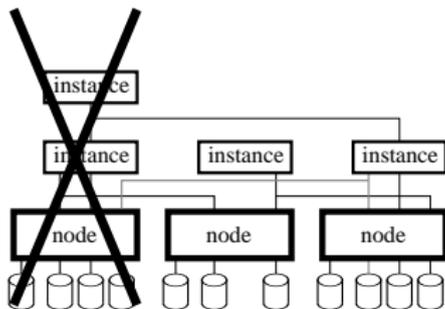
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation
- and to stay there



## Enter Ganeti

While all this works on its own,  
Ganeti helps

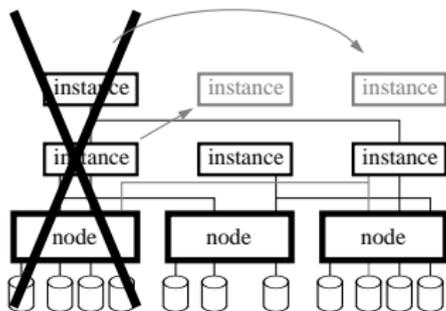
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation
- and to stay there



## Enter Ganeti

While all this works on its own,  
Ganeti helps

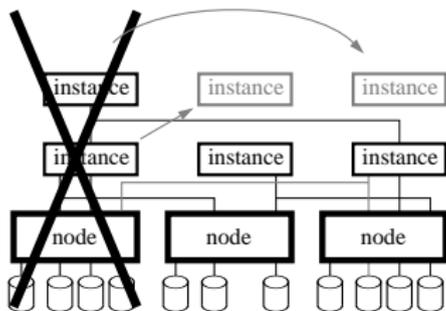
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation
- and to stay there
  - failover instances



## Enter Ganeti

While all this works on its own,  
Ganeti helps

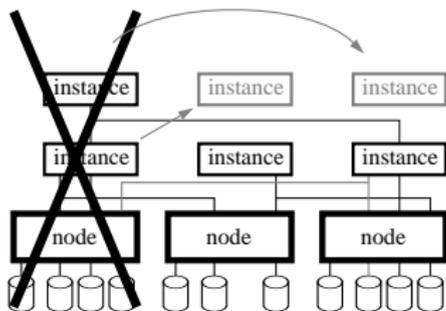
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation  
*keeping  $N + 1$  redundancy*
- and to stay there
  - failover instances



## Enter Ganeti

While all this works on its own,  
Ganeti helps

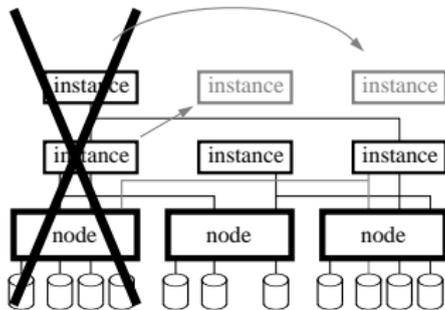
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation  
*keeping  $N + 1$  redundancy*
- and to stay there
  - failover instances  
and evacuate nodes



## Enter Ganeti

While all this works on its own,  
Ganeti helps

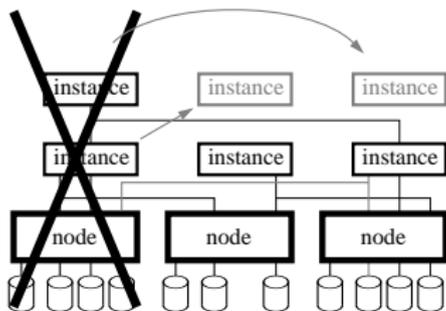
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation  
*keeping  $N + 1$  redundancy*
- and to stay there
  - failover instances  
and evacuate nodes
  - rebalance



## Enter Ganeti

While all this works on its own,  
Ganeti helps

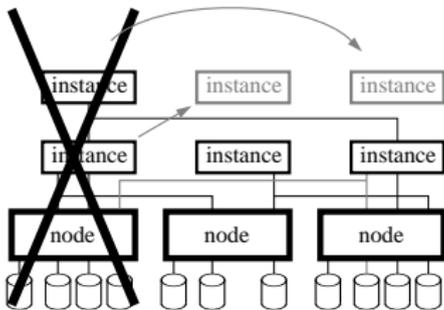
- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation  
*keeping  $N + 1$  redundancy*
- and to stay there
  - failover instances  
and evacuate nodes
  - rebalance
  - Restart instances after power outage



## Enter Ganeti

While all this works on its own,  
Ganeti helps

- to get there
  - uniform interface  
*hypervisors/storage/...*
  - policies, balanced allocation  
*keeping  $N + 1$  redundancy*
- and to stay there
  - failover instances  
and evacuate nodes
  - rebalance
  - Restart instances after power  
outage
  - ...





## Basic Interaction—Cluster creation

- `gnt-cluster init -s 192.0.2.1  
clusterA.example.com`





## Basic Interaction—Cluster creation

- `gnt-cluster init -s 192.0.2.1 clusterA.example.com`
- `gnt-node add -s 192.0.2.2 node2.example.com`





## Basic Interaction—Cluster creation

- `gnt-cluster init -s 192.0.2.1 clusterA.example.com`
- `gnt-node add -s 192.0.2.2 node2.example.com`
- ...





## Basic Interaction—Cluster creation

- `gnt-cluster init -s 192.0.2.1 clusterA.example.com`
- `gnt-node add -s 192.0.2.2 node2.example.com`
- ...
- `gnt-instance add -t drbd -o debootstrap -s 2G --tags=foo,bar instance1.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`
- `gnt-node migrate -f node2.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`
- `gnt-node migrate -f node2.example.com`
- `gnt-node evacuate -f -s node2.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`
- `gnt-node migrate -f node2.example.com`
- `gnt-node evacuate -f -s node2.example.com`
- `gnt-node modify --offline=yes node2.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`
- `gnt-node migrate -f node2.example.com`
- `gnt-node evacuate -f -s node2.example.com`
- `gnt-node modify --offline=yes node2.example.com`

### Using the node again

- `gnt-node modify --online=yes node2.example.com`





## Basic Interaction—Node maintenance

### Evacuating a node

- `gnt-node modify --drained=yes node2.example.com`
- `gnt-node migrate -f node2.example.com`
- `gnt-node evacuate -f -s node2.example.com`
- `gnt-node modify --offline=yes node2.example.com`

### Using the node again

- `gnt-node modify --online=yes node2.example.com`
- `hbal -L -X`





# Jobs

cli





# Jobs

cli

- gnt-cluster
- gnt-node
- gnt-instance
- ...





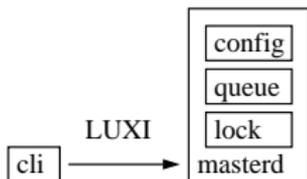
# Jobs

cli



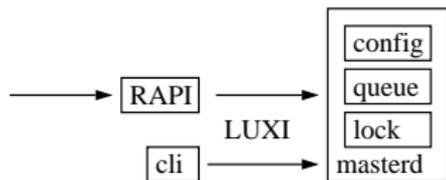


# Jobs



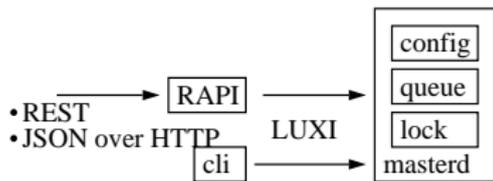


# Jobs



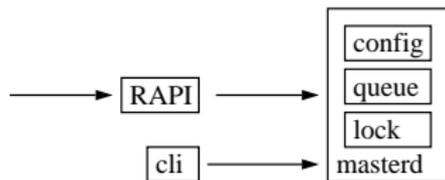


# Jobs



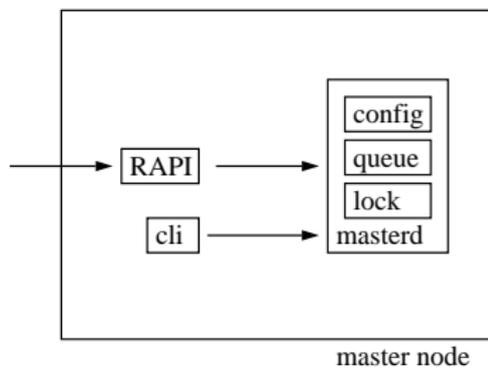


# Jobs





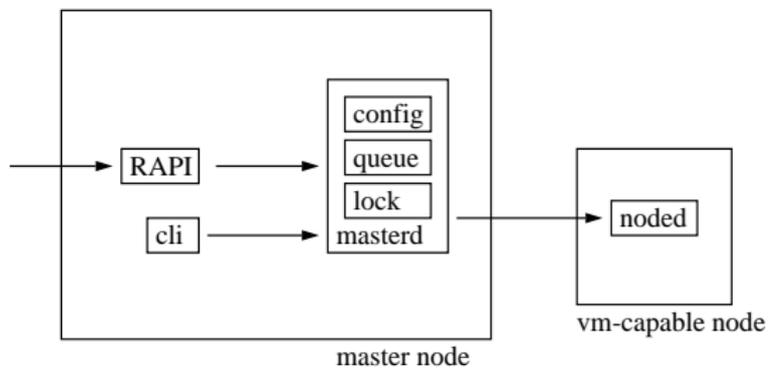
# RPC



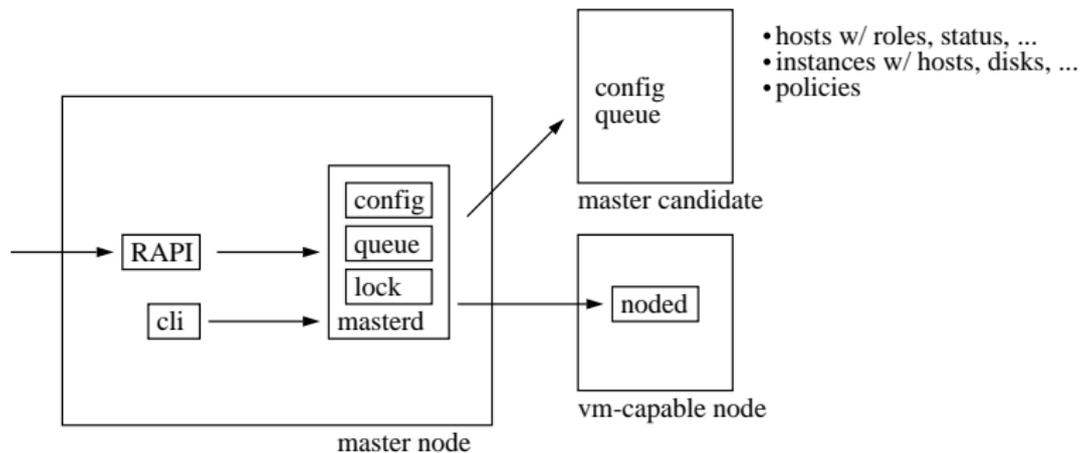




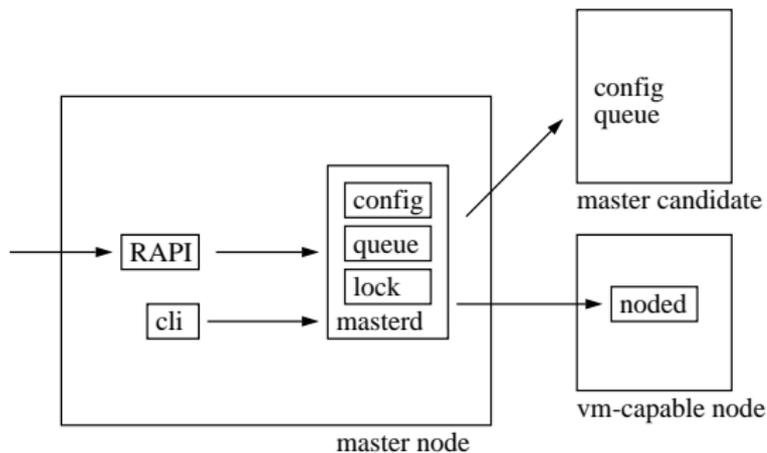
# RPC



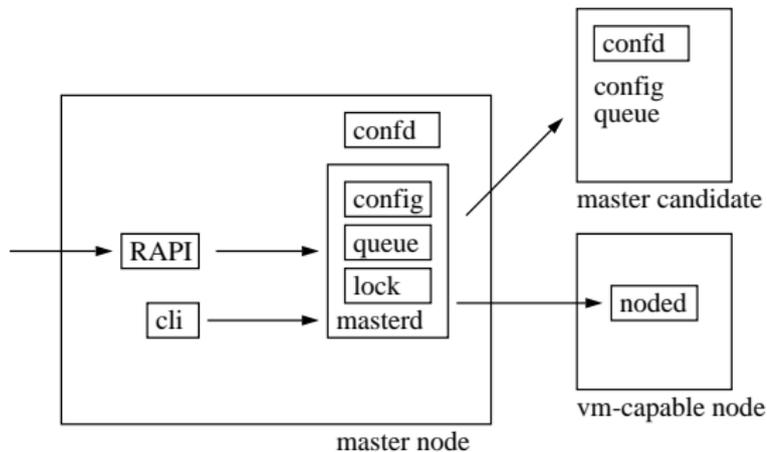
# Configuration



# Configuration



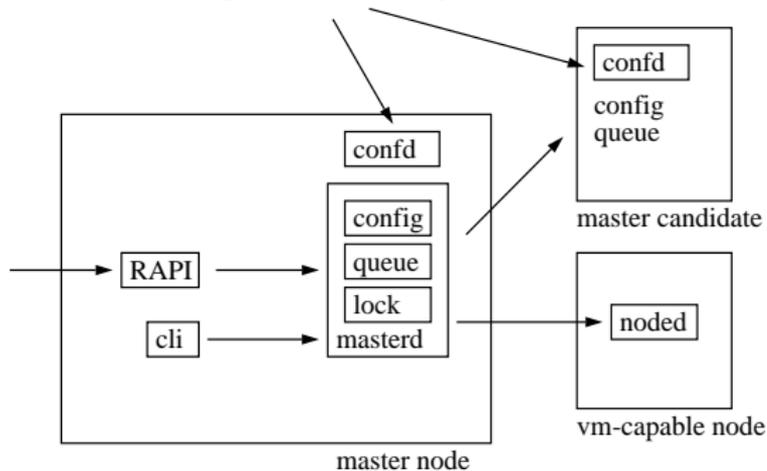
# Configuration



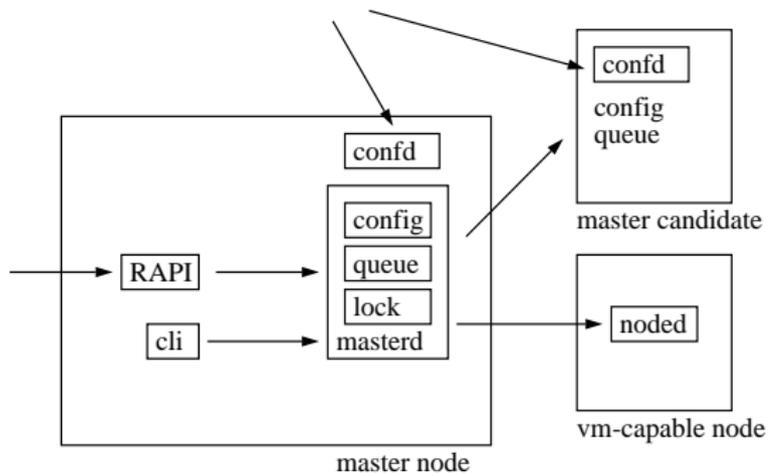
# Configuration

confd protocol

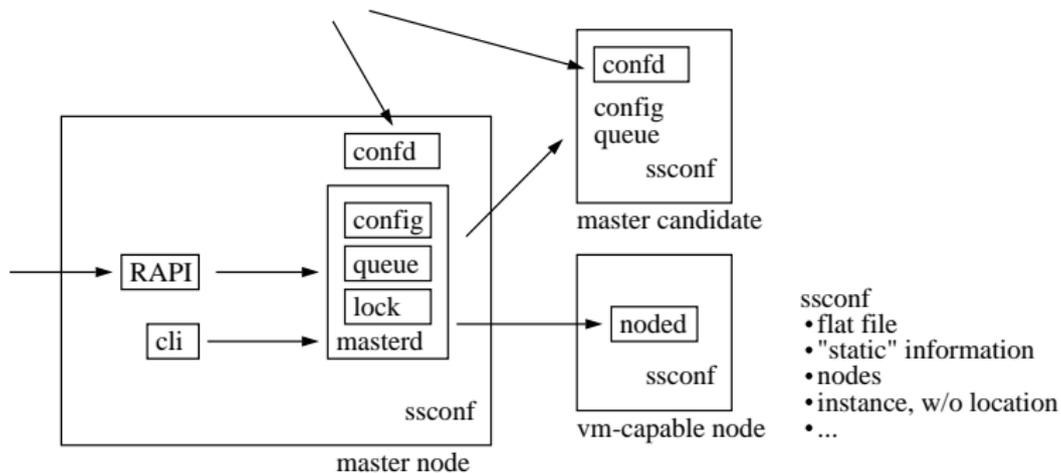
- upd
- ask all, take best answer
- configurations time stamped



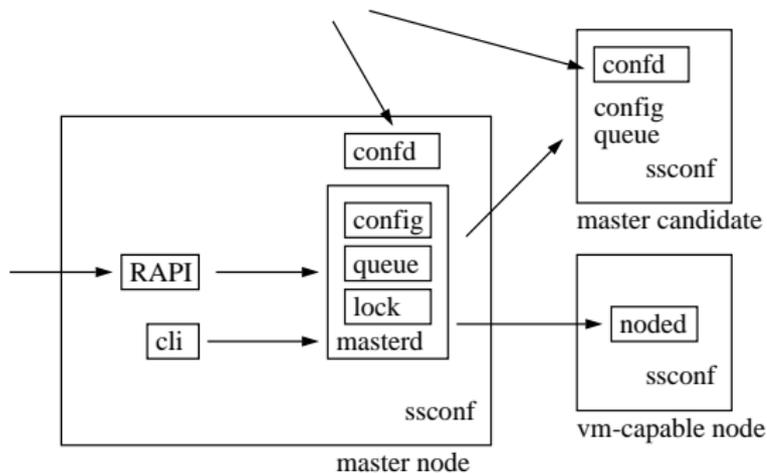
# Configuration



# Configuration



# Configuration





## Roles and Statuses

Nodes can serve different roles.

*(Nodes can, and usually do, take both roles.)*





## Roles and Statuses

Nodes can serve different roles.

*(Nodes can, and usually do, take both roles.)*

- VM-hosting nodes
  - VM-capable
  - grouped in “node groups”



## Roles and Statuses

Nodes can serve different roles.

*(Nodes can, and usually do, take both roles.)*

- VM-hosting nodes
  - VM-capable
  - grouped in “node groups”
- Administrative nodes
  - master capable (*policy decision*)
  - master candidate (*have a full copy of the live configuration*)
  - master (*manages all operations on the cluster*)



## Roles and Statuses

Nodes can serve different roles.

*(Nodes can, and usually do, take both roles.)*

- VM-hosting nodes
  - VM-capable
  - grouped in “node groups”
- Administrative nodes
  - master capable (*policy decision*)
  - master candidate (*have a full copy of the live configuration*)
  - master (*manages all operations on the cluster*)

Independently of its role, nodes can be in a different statuses:  
online, drained, offline



## Guest OS Interface

Ganeti is agnostic about the guest OSes;  
it just expects information to be provided.

*(on directory per guest OS)*

- executables: create, import, export, rename, verify
- text files: ganeti\_api\_version, variants.list

Executables are provided with information via the environment.

- OS\_VARIANT
- HYPERVISOR
- DISK\_COUNT, DISK\_0\_PATH, DISK\_1\_PATH, ...

- ...





## Available OS Definitions

There exist quite a few implementations of the guest OS interface.



## Available OS Definitions

There exist quite a few implementations of the guest OS interface.

- **debootstrap** (`git://git.ganeti.org/instance-debootstrap.git`)  
glorified call of `debootstrap(8)`  
`sfdisk, mkswap, mke2fs, ...; /etc/{hostname, ...}`









## Ways to customize Ganeti

- Hooks
- Allocator
- ...



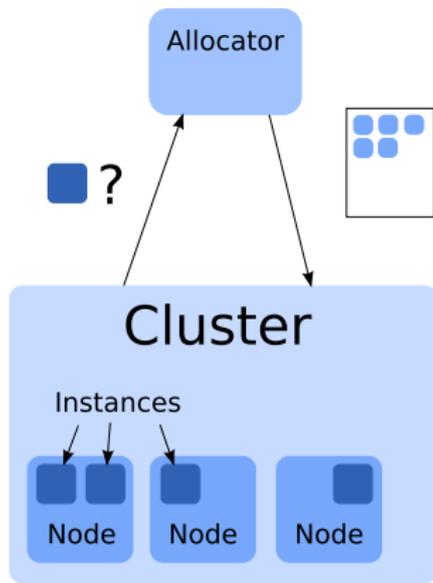
# Hooks

- hook scripts to customize cluster operations
- useful for syncing with external systems
- pre phase: e.g. for authorization
- post phase: e.g. for logging, billing, setting passwords
- examples: `cluster-verify-post.d`, `node-add-pre.d`



# Allocation

- Where to put an instance?
- protocol:
  - JSON over pipes
  - input: cluster's state + request-specific info
  - output: suggestions where to place which instance
- supported requests: allocate, relocate, change-group, node-evacuate, multi-allocate



# Ganeti in Production

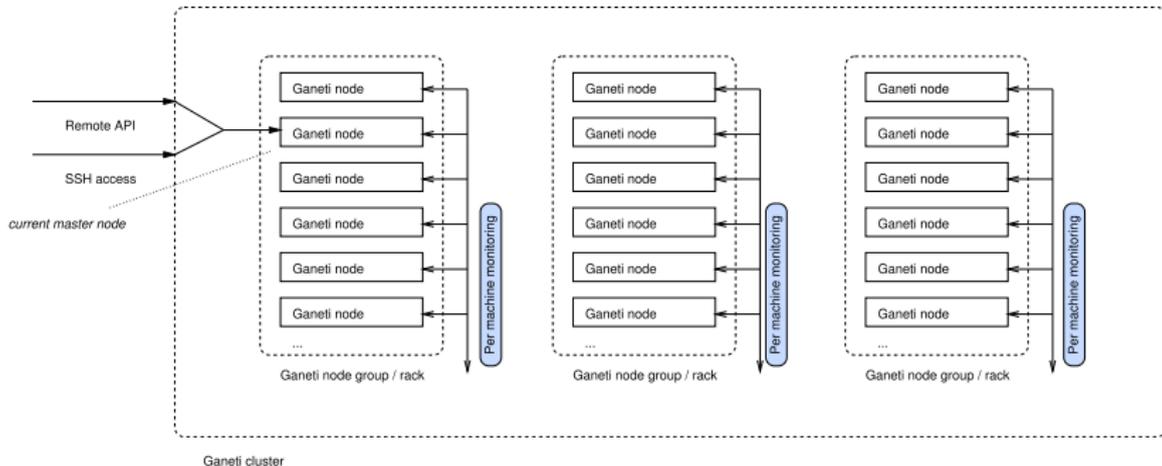
What should you add?

- Monitoring:
  - Check host disks, memory, load
- Automation:
  - Trigger events (evacuate, send to repairs, readd node, rebalance)
- Configuration Management:
  - Automated host installation / setup
- Self service use
  - Graphical interface (e.g. Ganeti Web Manager)  
(<http://ganeti-webmgr.readthedocs.org/en/latest/>)
  - Instance creation and resize
  - Instance console access



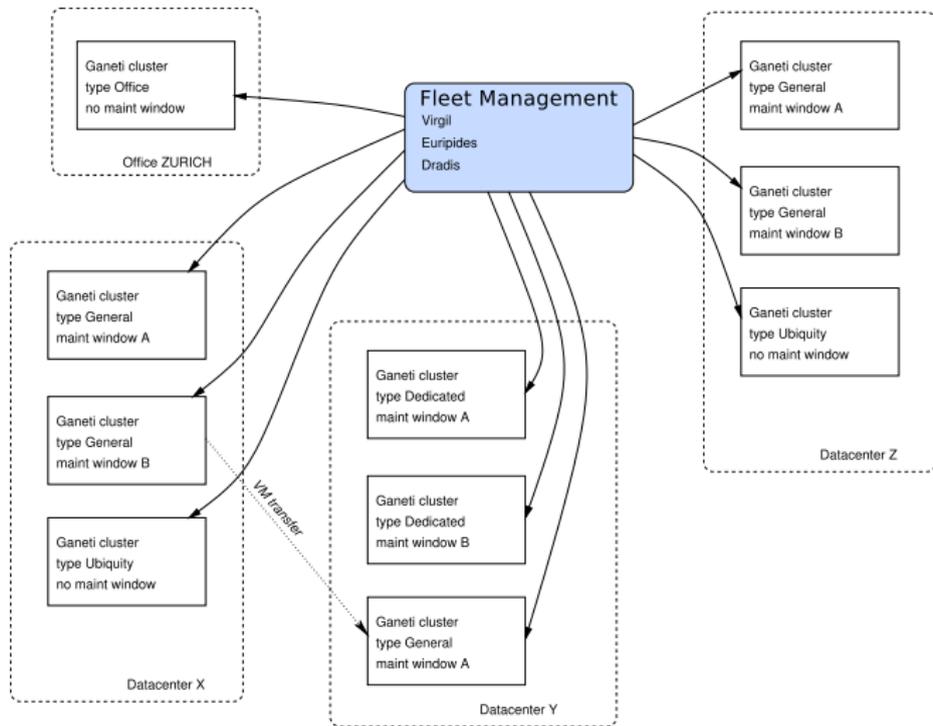
# Production Cluster

As we use it in a Google Datacenter





# Fleet at Google





## 2.7 (Current Release)

- Network management (contributed by grnet.gr)
- Exclusive storage
- Opportunistic locking
- Restricted commands
- Monitoring agent







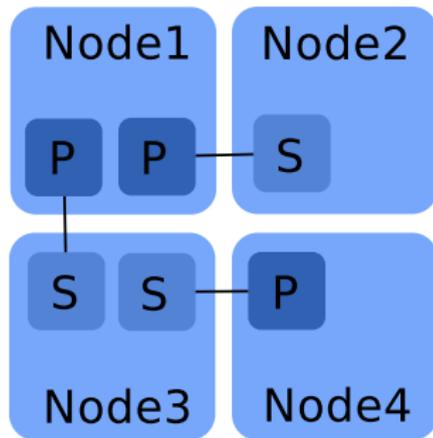
## 2.8 (Beta)

- Improved support of non-lvm storage
- Downgrading
- More work on monitoring daemon
- Autorepair tool
- Hroller



# Hroller

- Scheduler for rolling reboots
- Partiones cluster into groups of nodes that can be rebooted simultaneously
- various modes: default, full evacuation, offline-maintenance
- options for non-redundant instances





## 2.9 (Alpha)

- DRBD 8.4 support
- Improved support of non-lvm storage handling
- Improvements of monitoring agent
- Improvements of hroller





# Future

Just plans, no promises!

- Hot-plugging
- Automatic updates
- More fine-grained job-queue management
- Storage pools



# Open Source Ganeti

- Ganeti has been open source since 2007
- Relatively big community of external users and contributors
- People running Ganeti:
  - Google (Corporate Computing Infrastructure)  
(<https://www.youtube.com/watch?v=TElArK6SmyY>)
  - grnet.gr (Greek Research & Technology Network)
  - osuosl.org (Oregon State University Open Source Lab)
  - fsffrance.org (Free Software Foundation France)



## Ganeti Development Process

- Time-based release process, one freeze every 3 months
- Code reviews over the mailing list
- Discussion of design documents publicly on the mailing list
- Video-conferences with bigger contributors
- Public continuous build system<sup>1</sup>
- QA scripts public to be re-used





## Upcoming Events

- GanetiCon, Athens, Sep 2013

(<https://sites.google.com/site/ganeticon/>)

- LinuxCon North America, New Orleans, Sep 2013, introductory talk

(<http://events.linuxfoundation.org/events/linuxcon-north-america/program/schedule>)

- LinuxCon Europe, Edinburgh, UK, Oct 2013, introductory talk

(<http://events.linuxfoundation.org/events/linuxcon-europe>)

- LISA, Washington D. C., Nov 2013, workshop / class

(<https://www.usenix.org/conference/lisa13>)

A list of publications from previous events (slides, recordings) can be found in our wiki. (<https://code.google.com/p/ganeti/wiki/Publications>)





## Conclusion

- Check us out at <https://code.google.com/p/ganeti/>
- Or just search for "Ganeti"
- We are around on FrOSCon today and tomorrow!



Questions? Feedback? Ideas? Flames?

©2010-2013 Google

Use under GPLv2+ or CC-by-SA

Some images borrowed / modified from Lance Albertson and Guido Trotter



Google