Build your own cloud
using ganeti, (kvm, drbd) salt and zfs

Dobrica Pavlinušić
Luka Blašković
DORS/CLUC 2014
What are we going to talk about?

- Which cloud IaaS or PaaS
- FFZG legacy infrastructure overview
- Ganeti - Open Source cloud solution
- SaltStack - deploy ganet nodes
- ZFS - storage server (nfs)
- our migration to cloud
Cloud: is it IaaS or PaaS?

**Infrastructure as a service**
- reliable, persistent VMs
- legacy consolidation

**Platform as a service**
- deploy applications using custom config

- VMWare
- Amazon EC2 (persistent?)
- oVirt (libvirt)
- Ganeti
- OpenStack

- heroku
- Google App Engine
- Azure
- Docker (kubernetes, DEIS)
Motivation for building a cloud

- 10+ aging Debian GNU/Linux machines installed in last 15 years on three locations
- upgraded memory (FB DIMM DDR2, from ebay, cheap)
- upgraded disks (SAS and SATA)
- better resource usage
- **high availability**
  - resilient to failure of machines
  - maintenance during working hours
- VMs are not cattle, they are pets
- Every VM configured like real snowflake
SaltStack

- Automation for installation of ganeti nodes
- ZeroMQ and declarative rules
- Deployment of new node under an hour

[https://github.com/ffzg/ganeti-salt](https://github.com/ffzg/ganeti-salt)
Ganeti integrates known tools

- kvm (or xen) virtualization
- drbd (w/ LVM) for disk replication (no SAN!)
- kvm+drbd = HA with live migration

Terminology:
- node - physical hardware
- instance - virtual machine
- cluster - combination of above components

gnt-* command-line interface for sysadmins
Ganeti hints

What you wanted to know about cloud but were too afraid to ask it…. 
ganeti nodes and instances

root@vmh02:~# gnt-node list

<table>
<thead>
<tr>
<th>Node</th>
<th>DTot</th>
<th>DFre</th>
<th>MTot</th>
<th>MNode</th>
<th>MFree</th>
<th>Pinst</th>
<th>Sinst</th>
</tr>
</thead>
<tbody>
<tr>
<td>arh-hw.gnt.ffzg.hr</td>
<td>?</td>
<td>?</td>
<td>7.8G</td>
<td>173M</td>
<td>1.3G</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>blade05.gnt.ffzg.hr</td>
<td>123.7G</td>
<td>1.4G</td>
<td>7.8G</td>
<td>5.0G</td>
<td>2.5G</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>box02.gnt.ffzg.hr</td>
<td>1.3T</td>
<td>1005.6G</td>
<td>15.7G</td>
<td>10.0G</td>
<td>6.7G</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>lib10.gnt.ffzg.hr</td>
<td>3.6T</td>
<td>2.5T</td>
<td>19.6G</td>
<td>12.1G</td>
<td>10.6G</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>lib15.gnt.ffzg.hr</td>
<td>543.7G</td>
<td>279.5G</td>
<td>15.7G</td>
<td>8.4G</td>
<td>10.6G</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>lib20.gnt.ffzg.hr</td>
<td>822.6G</td>
<td>516.4G</td>
<td>15.7G</td>
<td>10.7G</td>
<td>4.2G</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>vmh01.gnt.ffzg.hr</td>
<td>917.0G</td>
<td>583.3G</td>
<td>11.7G</td>
<td>7.6G</td>
<td>4.6G</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>vmh02.gnt.ffzg.hr</td>
<td>917.0G</td>
<td>569.7G</td>
<td>15.7G</td>
<td>10.0G</td>
<td>6.5G</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>vmh03.gnt.ffzg.hr</td>
<td>917.0G</td>
<td>592.9G</td>
<td>15.7G</td>
<td>8.9G</td>
<td>9.5G</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>vmh11.gnt.ffzg.hr</td>
<td>264.9G</td>
<td>38.6G</td>
<td>7.8G</td>
<td>5.2G</td>
<td>1.7G</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>vmh12.gnt.ffzg.hr</td>
<td>917.0G</td>
<td>566.6G</td>
<td>15.7G</td>
<td>9.7G</td>
<td>7.7G</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

root@vmh02:~# gnt-instance list --no-headers -o status,hv/kernel_path | sort | uniq -c

  2 ADMIN_down
  4 ADMIN_down /boot/vmlinuz-3.2-kvmU
34 running
33 running /boot/vmlinuz-3.2-kvmU
disks

● two LVs as disks for instance (root, swap)
● boot via grub or from host kernel
● liberal use of nfs (from zfs pool) to provide shares to VMs (backups, archives…)
● gnt-instance modify -t drbd
● gnt-backup assumes 1 partition per disk
  ○ create LV snapshot (without shutdown)
  ○ transfer dump of file system to some node
  ○ remove snapshot
● plan to modify into incremental backup
  ○ lv snapshot => rsync => zfs snap
PERC SAS/SATA controllers

PERC 4 - bios JBOD mode (SCSI vs RAID)
PERC 5 - no JBOD mode
PERC 6 - LSI IT firmware for JBOD mode (newer IR have JBOD)
SMBus issue on Intel Chipsets with tape fix

http://www.overclock.net/t/359025/perc-5-i-raid-card-tips-and-benchmarks
VCPU

- give more than one VCPU to VMs
  - monitor uptime load of instance < VCPU
- do you want to pin kvm VCPUs to node?
  - might be beneficial for HPC nodes (caches?)
- kernel
  - node: 3.10 based on proxmox rhel7 kernel https://github.com/ffzg/linux-kernel-3.10
  - instance: 3.2-kvmU (3.10-kvmU)
- in mixed nodes environment, use common cpu set for kvm to enable VM migration anywhere
reboot

- It will happen, sooner than you think
- don’t run manually started services!
- acpi-support-base for clean shutdown
- gnt-instance reboot [instance]
  - power-cycle as opposed to reboot within instance
    (ganeti >=2.11 kvmd)
  - required to reload kvm config, hwclock, etc
network

- bonded 1G bridges per vlan
- jumbo frames for drbd traffic (9k mtu)
- disable host nic hardware offloads
- don’t let bridge traffic pass through fw chains
- pay with sysctl setting, switch congestion control algorithm
- Use our virtio-mq patch (ganeti >=2.12, linux kernel >=3.8)
tap challenges

qemu-system-x86_64: -netdev type=tap,id=hotnic-a74f9700-pci-6,fd=8,vhost=on: Device 'tap' could not be initialized

gnt-instance modify -H vhost_net=false pxelator

- mysterious unreported bug when vhost_net=True (network offloading from qemu to separate kernel thread)
- we will fix this, don’t worry :)}
groups

- limit instance drbd replication and migration
  - same top-of-rack switch

```
root@vmh02:~# gnt-instance list --no-headers -o status,pnode.group,snodes.group | sort | uniq -c
  6 ADMIN_down test
  6 running    default
48 running    default default
  5 running    lib    lib
  8 running    test
```
console

- serial console
  - console=ttyS0
  - gnt-instance console [instance]
- VNC for graphic console
  - vnc on local address
  - NoVNC web console
  - https://code.grnet.gr/projects/ganetimgr/
  - https://code.osuosl.org/projects/ganeti-webmgr/
NoVNC web console

https://code.grnet.gr/projects/ganetimgr/
time

- ntp and/or ntpdate inside vms harmful
- ntp should be on node
- make sure that UTC=yes is same on vm/host
htools

A collection of tools to provide auxiliary functionality to Ganeti.

- **hail**: `gnt-instance add -I hail instance #`
  
  Where to put an instance?

- **hbal**: `hbal -G default -L #` cluster balancing

- **hspace**: `hspace -L #` How many more instances can I add to my cluster?

- **harep**: `harep -L #` repair/recreate instances
Migration of LXC into Ganeti VMs

Your (LXC) snowflakes can melt in process

- create LV for root fs
- rsync files over (defragment, ext4 upgrade)
- VMs disk size = used + 10%
- use host 3.2 kernel to run machines
- install modules and acpi support
- modify disk configuration to drbd

http://sysadmin-cookbook.rot13.org/#ganeti_migrate_lxc
Our experience

- We are not creating similar instances
- Performance impact compared to LXC
- Memory usage of VM hit-or-miss game
- Memory upgrade during working hours (evacuate, power off, upgrade, hbal)
- Firmware upgrades become reality
- First time to backup some machines (!)
- Works for us™
- https://code.google.com/p/ganeti/wiki/PerformanceTuning
Ganeti is good cloud core

- Synnefo - AWS like compute, network, storage [https://www.synnefo.org/](https://www.synnefo.org/)
  - OpenStack API (not code!)
  - Archipelago - distributed storage management
    - Ceph - distributed disk store
Questions?

See you at workshop!
Workshop!
Technologies

- Linux and standard utils (iproute2, bridge-utils, ssh)
- socat
- KVM/Xen/LXC
- DRBD, LVM, SAN, Ceph, Gluster (=>2.11)
- Python (plus a few modules)
- Haskell
Ganeti on ganeti

● 6 virtual nodes
● nested virtualization not working (no KVM)
● separate volume group
● so plan is to setup XEN-PVM (paravirtualized), sorry no KVM this time :(
Bootstrap virtual “nodes”

gnt-instance add -t plain \ 
-n node{0..5} 
-B maxmem=3.7G, minmem=1G, vcpus=4 \ 
-o debootstrap+salted \ 
--disk 0:size=20g, vg=dorsvg \ 
--disk 1:size=2g, vg=dorsvg \ 
--disk 2:size=300g, vg=dorsvg \ 
--net 0:mode=bridged, link=br1001 \ 
--net 1:mode=bridged, link=br0080 \ 
--no-name-check --no-ip-check \ 
dors-ganeti{0..5}.dhcp.ffzg.hr # metavg= for drbd
debootstrap+salted

- debootstrap default variant with saltstack bootstrap script:

Initial salting

- nodes (minions) are automagically connected to master (known as “h”)

```bash
lblask@h:~$ sudo salt-key -L
Accepted Keys:
Unaccepted Keys:
dors-ganeti01.dhcp.ffzg.hr
dors-ganeti02.dhcp.ffzg.hr
dors-ganeti03.dhcp.ffzg.hr
dors-ganeti12.dhcp.ffzg.hr
dors-ganeti20.dhcp.ffzg.hr
dors-ganeti21.dhcp.ffzg.hr
```
Initial salting

lblask@h:~$ sudo salt-key -A
The following keys are going to be accepted:
Unaccepted Keys:
dors-ganeti01.dhcp.ffzg.hr
dors-ganeti02.dhcp.ffzg.hr
dors-ganeti03.dhcp.ffzg.hr
dors-ganeti12.dhcp.ffzg.hr
dors-ganeti20.dhcp.ffzg.hr
dors-ganeti21.dhcp.ffzg.hr
Proceed? [n/Y] y
Initial salting

- check boring stuff (apt_sources, dhcp hostname, locales, timezone, ssh)
- install xen kernel and tools
- leave hard work to workshops
Initial salting

- modify instances to boot from own kernel

    for x in $(gnt-instance list|grep dors|awk '{print $1}'| xargs); do
gnt-instance modify --submit
    -H initrd_path=,kernel_path=,disk_type=scsi,
    nic_type=e1000 $x
    done
Initial salting

- reboot instances

```bash
for x in \
$(gnt-instance list|grep dors|awk '{print $1}'| xargs); \
do gnt-instance reboot --submit $x \
; done
```


- will be using latest Ganeti from wheezy-backports (2.10)

- [http://docs.ganeti.org/ganeti/2.10/html/install.html#ganeti-installation-tutorial](http://docs.ganeti.org/ganeti/2.10/html/install.html#ganeti-installation-tutorial)
SSH to machine

ssh root@hostname.dhcp.ffzg.hr

- password
- change password :D
Hostname

- ganeti needs fqdn in hostname:
- run:

```bash
echo "hostname.dors.cluc" > /etc/hostname
hostname hostname.hostname.dors.cluc
```
/etc/hosts

- should have valid hosts file:
- run:

```bash
echo "172.16.1.XXX hostname.dors.cluc hostname" >> /etc/hosts
```

```bash
echo "172.16.1.1 cluster.dors.cluc" >> /etc/hosts
```
checkpoint

hostname -f # should work
XEN specific settings

- go to: http://docs.ganeti.org/ganeti/2.10/html/install.html#xen-settings

Limit amount of memory dedicated to hypervisor, add to /etc/default/grub:

```
GRUB_CMDLINE_XEN_DEFAULT="dom0_mem=512M"
```
Selecting the instance kernel

$ cd /boot
$ ln -s vmlinuz-3.2.0-4-amd64 vmlinuz-3-xenU
$ ln -s initrd.img-3.2.0-4-amd64 initrd-3-xenU
DRBD setup

echo "drbd minor_count=128
usermode_helper=/bin/true" >> /etc/modules

apt-get install drbd8-utils
Network setup

auto xen-br0
iface xen-br0 inet static
    address YOUR_IP_ADDRESS
    netmask YOUR_NETMASK
bridge_ports eth1
bridge_stp off
bridge_fd 0
up ip link set addr $(cat /sys/class/net/eth1/address) dev $IFACE
Network setup

apt-get install bridge-utils
ifup xen-br0
LVM setup

apt-get install lvm2
pvcreate /dev/sdc
vgcreate xenvg /dev/sdc
Install ganeti & instance-debootstrap

apt-get install -t wheezy-backports ganeti

apt-get install -t wheezy-backports ganeti-instance-debootstrap
Initialize cluster

gnt-cluster init --vg-name xenvg --no-etc-hosts
--master-netdev xen-br0 --enabled-hypervisors
xen-pvm --primary-ip-version 4 cluster.dors.cluc
Initialize cluster

# set default memory and vcpu count

gnt-cluster modify -B vcpus=2,memory=512M
Add a second node

gnt-node add --master-capable=yes dors-ganeti20.dors.cluc
Create the instance

gnt-instance add -n *hostname* -o
debootstrap+default -t plain -s 3G --no-ip-check
--no-name-check myfirstinstance
Lets play

```
  gnt-instance *
  gnt-node *
  hbal
  hspace
  -l hail
  .....```

Kibana, LogStash and ElasticSearch

dpavlin@kibana:/etc/cron.hourly$ cat kibana-drop-index
#!/bin/sh -xe

min_free=`expr 2048 \* 1024` # k

free() {
    df -kP /var/lib/elasticsearch/ | tail -1 | awk '{ print $4 }'
}

while [ $(free) -lt $min_free ] ; do


done
Thx!