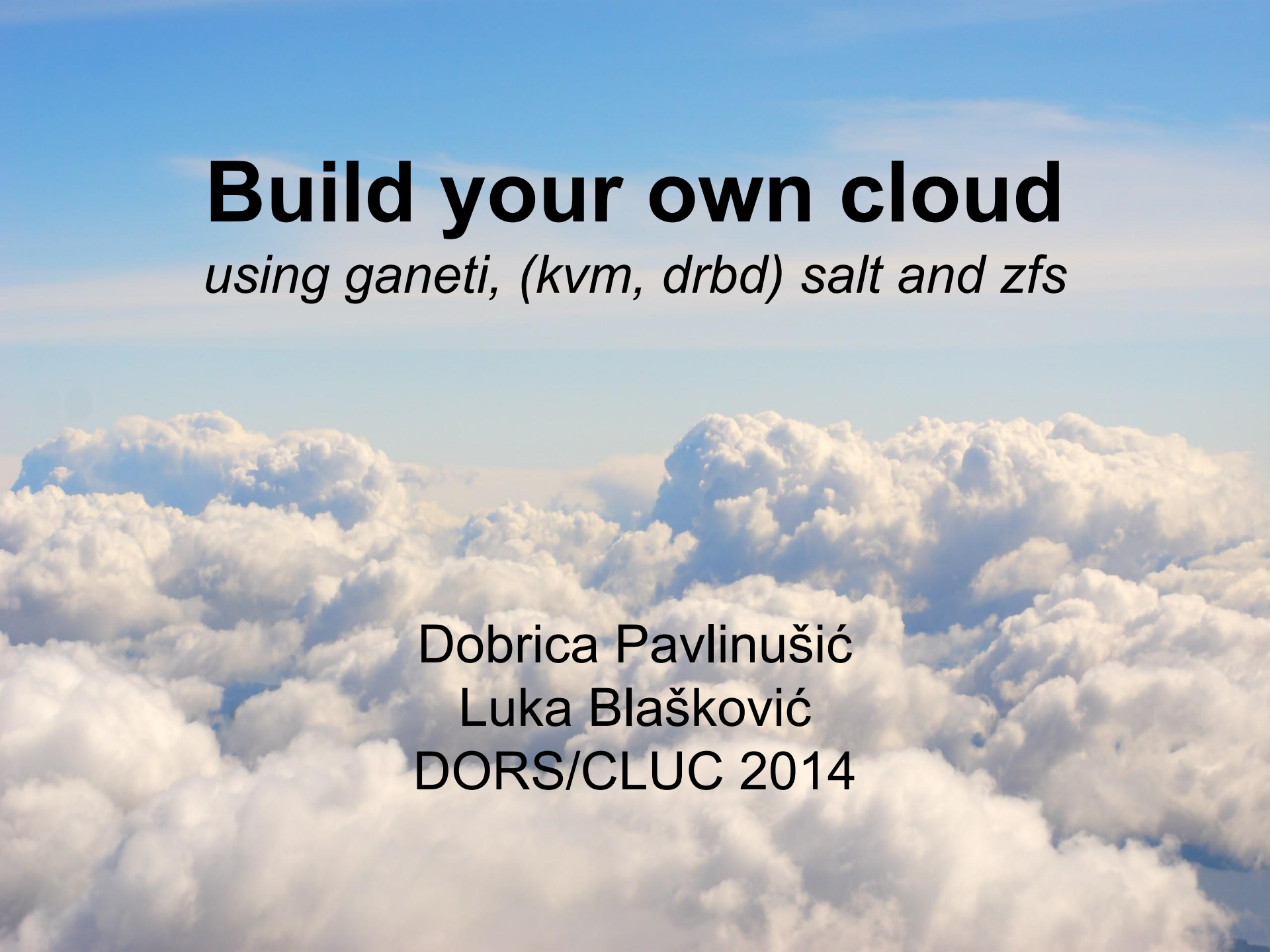


Build your own cloud

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

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

Platform as a service

deploy applications using custom config

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

- http://www.saltstack.com/
- automation for installation of ganeti nodes
- ZeroMQ and declarative rules
- deployment of new node under an hour
  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
Node                  DTotal   DFree MTotal MNode MFree Pinst Sinst
arh-hw.gnt.ffzg.hr         ?       ?   7.8G  173M  1.3G     0     0
blade05.gnt.ffzg.hr 123.7G    1.4G   7.8G  5.0G  2.5G     8     2
box02.gnt.ffzg.hr     1.3T 1005.6G  15.7G 10.0G  6.7G    14     0
lib10.gnt.ffzg.hr     3.6T    2.5T  19.6G 12.1G 10.6G     8     7
lib15.gnt.ffzg.hr   543.7G  279.5G  15.7G  8.4G 10.6G     3     2
lib20.gnt.ffzg.hr   822.6G  516.4G  15.7G 10.7G  4.2G     3     3
vmh01.gnt.ffzg.hr   917.0G  583.3G  11.7G  7.6G  4.6G     8     8
vmh02.gnt.ffzg.hr   917.0G  569.7G  15.7G 10.0G  6.5G     8     7
vmh03.gnt.ffzg.hr   917.0G  592.9G  15.7G  8.9G  9.5G     8     7
vmh11.gnt.ffzg.hr   264.9G   38.6G   7.8G  5.2G  1.7G     8     7
vmh12.gnt.ffzg.hr   917.0G  566.6G  15.7G  9.7G  7.7G     5    10
```

```
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

- ganetimgr - KISS web interface https://code.grnet.gr/projects/ganetimgr/
- Synnefo - AWS like compute, network, storage 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:

https://raw.githubusercontent.com/lblasc/dorscluc2014-ganeti/master/salted_variant.sh

Initial salting

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

```
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

- used states: https://github.com/lblasc/dorscluc2014-ganeti
- check boring stuff (apt_sources, dhcp hostname, locales, timezone, ssh)
- install xen kernel and tools
- leave hard work to workshoppers

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

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

Go go http://bit.ly/dc14-ganeti

- open: https://github.com/lblasc/dorscluc2014-ganeti#dorscluc-ganeti-workshop

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

- 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:

echo “***hostname***.dors.cluc” > /etc/hostname

hostname ***hostname***.dors.cluc

/etc/hosts

- should have valid hosts file:
- run:

echo “172.16.1.**XXX** ***hostname***.dors.cluc ***hostname***” >> /etc/hosts

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

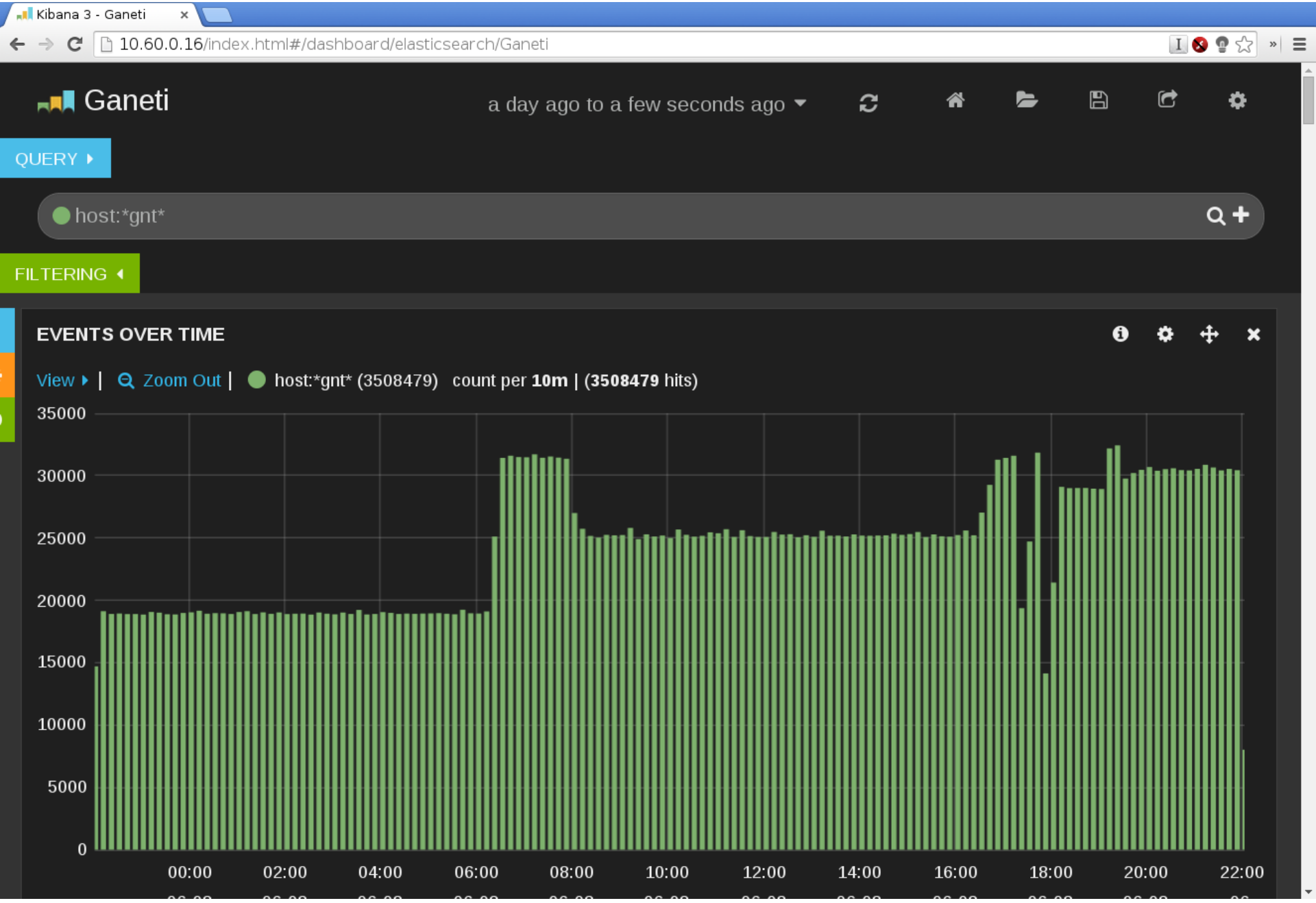
hspace

-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

curl http://localhost:9200/_cat/indices | sort -k 2 | grep
logstash- | head -1 | awk '{ print $2 }' | xargs -i curl -
XDELETE 'http://localhost:9200/{}'

done
```


Kibana 3 - Ganeti
10.60.0.16/index.html#/dashboard/elasticsearch/Ganeti
Ganeti
a day ago to a few seconds ago
QUERY
host:*gnt*
FILTERING
EVENTS OVER TIME
View | Zoom Out | host:*gnt* (3508479) count per 10m | (3508479 hits)
35000
30000
25000
20000
15000
10000
5000
0
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00

Thx!