Ganeti @ skroutz

Apollon Oikonomopoulos

apollon@skroutz.gr

GanetiCon 2013
3-5 Sep 2013 — Athens, Greece

Brief introduction

skroutz

Skroutz — what we do

Main product: price comparison engine

- ∼ 1k e-shops
- 5M products
- 2.5M unique visits / month
- 200k visits / day
- 2 countries (GR & TR)

skroutz

Skroutz — what we do

Main product: price comparison engine

- ∼ 1k e-shops
- 5M products
- 2.5M unique visits / month
- 200k visits / day
- 2 countries (GR & TR)

Side-projects:

- SkroutzStore: e-shop as-a-service
- Skroutz MyBill: mobile phone contract comparison

skroutz

What we use

100% FOSS stack

- Debian
- Ruby on Rails
- Percona/MariaDB
- ElasticSearch
- MongoDB
- Redis
- …

skroutz

Infrastructure

- 45 physical servers
- 90 virtual machines
- 3 physical locations
  - production site
  - HQ
  - old DC

skroutz

Ganeti at skroutz

skroutz

Ganeti at skroutz

Runs production and development instances. Production:

- ElasticSearch cluster
- Alve.com appservers
- Redis
- Skroutzstore & MyBill servers
- Analytics MongoDB (“Wanderer”)
- Mail relays …

Ganeti helped seamlessly migrate our infrastructure to a new site.
We *do* trust and value Ganeti a lot!

skroutz

Ganeti at skroutz (2)

```
# hspace -L
...
The cluster has 20 nodes and the following resources:
  MEM 456173, DSK 9656860, CPU 168, VCPU 672.
There are 86 initial instances on the cluster.
```

A *single* Ganeti cluster with...

- 20 nodes
- 3 nodegroups (at different locations)
- 85+ KVM instances
- DRBD (using secondary IPs)
- ganeti-instance-image

skroutz

Ganeti + puppet

class `ganeti::node` {

- Install ganeti, g-i-m, qemu-kvm
- Create /etc/ganeti/hooks and install custom hooks
- Turn on KSM for KVM memory deduplication
- Make sure `drbd` and `vhost_net` modules are loaded
- Permit root SSH access
- ”Orphan” nodes only: populate /root/.ssh/authorized_keys with all known cluster keys
- Install firewall rules
- Install additional Icinga/Check-MK checks

}

skroutz

Firewall configuration

- Firewall on each node, using `ferm`
- 2 distinct configurations, distinguished by `ssconf_*`
  1. ”Orphan” node (not part of a cluster): allow pubkey-only SSH from everywhere (limited by edge firewall)
  2. Normal node: permit SSH, RPC, confd, KVM migration and DRBD from nodes only (+ RAPI on the cluster IP)

```
@def $CLUSTER = `cat /var/lib/ganeti/ssconf_cluster_name 2>/dev/null || true`;
@def $PRIMARY_NODE_IPS = `cat /var/lib/ganeti/ssconf_node_primary_ips 2>/dev/null \
                          | awk '{ print $2 }'`;
@if $CLUSTER {
    domain ip table filter chain accept_ganeti_nodes {
        saddr $PRIMARY_NODE_IPS ACCEPT;
        saddr $SECONDARY_NODE_IPS ACCEPT;
    }
...
```

- `node-{add,remove}-post.d` hook triggers fw reload on *all* nodes

skroutz

Node monitoring

Icinga + Check-MK $\rightarrow$ easy-to-write local checks. Standard checks +

- Is /dev/kvm present? (bitten by this once...)
- Are there instances running with older KVM binary versions?

Puppet ENC querying RAPI, automatically setting

- icinga hostgroup (“ganeti-vms” or “ganeti-nodes”)
- parent node to the host node (VMs will appear as unreachable if node down/unreachable)

Attempts at injecting and checking `gnt-cluster verify output`, but waiting for 2.8 & `ganeti-mond` for further integration.

skroutz

Challenges

skroutz

Secondary IPs and group moves

- Our nodes have *public* primary IP addresses (IPv4 + IPv6)
- We use *unroutable* private subnets for secondary IPs
- Instance group moves happen over secondary IPs if these are configured

skroutz

Secondary IPs and group moves

- Our nodes have *public* primary IP addresses (IPv4 + IPv6)
- We use *unroutable* private subnets for secondary IPs
- Instance group moves happen over secondary IPs if these are configured
- … so we had to route the private subnets between different sites over the internet (OpenVPN + pain)

skroutz

Secondary IPs and group moves

- Our nodes have *public* primary IP addresses (IPv4 + IPv6)
- We use *unroutable* private subnets for secondary IPs
- Instance group moves happen over secondary IPs if these are configured
- … so we had to route the private subnets between different sites over the internet (OpenVPN + pain)

It would be better if

- group moves could use (or fall back to) the primary IPs.

skroutz

Instance placement restrictions

`htools` currently supports placement restrictions using “exclusion tags”. Great for

- making sure both your mail relays won’t end up on the same node
- maintaining your ElasticSearch cluster quorum in case of node failure

However, *not* great for

- making sure primary and secondary are not on the same blade chassis

Proposal: generic “location awareness” (separate design discussion)

skroutz

Plain instance support enhancements

Why use plain instances?

- fsync() is painfully slow over DRBD
- application-level redundancy provides fault-tolerant behaviour and data-loss is no issue
  - mail relays
  - ElasticSearch nodes
  - “stateless” application servers
- you don’t care about data loss/availability at all
  - test/experimental machines

Ganeti’s plain instance support is a bit “rough” around the edges. *htools* could deal with them in a more graceful manner (node migrations, rolling restarts, etc). Ideally *instance groups* could be used to define availability zones using tags.

skroutz

squeeze + 2.5 → wheezy + 2.7 migration

squeeze + 2.5

- linux 2.6.32 + qemu-kvm 0.12.5
- wrapper around `/usr/bin/kvm.real` to pass custom flags (e.g. `-cpu qemu64,+ssse3`)

wheezy + 2.7

- linux 3.2 + qemu-kvm 1.1
- no need for the wrapper, ganeti now supports `cpu_type` hvp
  - but what about already-running instances?
    → modify all `.runtime` files with a script to add the extra arguments to already-running instances.

Proposal: no idea (!)

skroutz

Thank you!

Q&A

skroutz