Ganeti @ skroutz

Apollon Oikonomopoulos

apollon@skroutz.gr

GanetiCon 2014
2-4 Sep 2014 — Portland, OR

Brief introduction

skroutz

Skroutz — what we do

Main product: price comparison engine

- ∼ 1k e-shops
- 6M products
- 3.5M unique visits / month
- 320k visits / day
- 2½ countries (GR, TR & UK)

skroutz

What we use

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

skroutz

Infrastructure

- 50 physical servers
- 150 virtual machines
- 3 physical locations
  - production site
  - HQ
  - old DC

skroutz

Ganeti at skroutz

skroutz

Ganeti at skroutz

Runs production, development and staging instances. Production:

- ElasticSearch cluster
- alve.com and scrooge.co.uk appservers
- Redis
- Skroutzstore & MyBill servers
- Analytics infrastructure MongoDB
- Mail relays …

Ganeti helped seamlessly migrate our infrastructure to a new site, and will also help our upcoming migration this fall.
We *do* trust and value Ganeti a lot!

skroutz

Ganeti at skroutz (2)

A *single* Ganeti cluster with…

- 32 nodes
- 3 nodegroups (one per location)
- 150+ KVM instances
- DRBD (using secondary IPs)
- ganeti-instance-image

skroutz

Ganeti + puppet

class `ganeti::node` {

- Install ganeti, g-i-m, qemu
- 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 → 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)

skroutz

Challenges

skroutz

Staging instances

- We prefer running a single cluster for (almost) all our needs.
- Staging cluster: 3-4 instances (app server, ES server, DB server), with iSCSI-backed disks.
- Our staging environment requires automated instance creation and cleanup via RAPI $\rightarrow$ should access only part of the cluster.
- RAPI has a two-level authz control, RO or RW $\rightarrow$ coarse granularity.
- An additional authnz application is not trivial to implement in many cases.

skroutz

RAPI access control

- Minimal, future-proof™ 100 LOC wrapper around ganeti-rapid, providing a custom request handler at HTTP level.
- Use additional access levels in the RAPI credentials file: `user password staging`
- Filter request *payload* according to the user account tags: staging accounts can only manipulate instances with specific name patterns.
- Authorization could be better performed at the OpCode level, but this requires extensive modifications.

skroutz

Small things we miss

- Ability to specify the *nodegroup* at `gnt-instance add` time (requires changes to the iallocator protocol)
- A more expressive, CLI-friendly query language `gnt-instance list -F …`
- Location awareness

skroutz

Thank you!

Q&A

skroutz