Ganeti @ skroutz

Apollon Oikonomopoulos

apollon@skroutz.gr

GanetiCon 2015
15-17 Sep 2015 — Prague, CZ

Brief introduction

skroutz

Skroutz — what we do

Main product: price comparison engine

- ∼ 1k e-shops
- 6M products
- 5M unique visits / month
- 500k 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

- 80 physical servers
- 180 virtual machines
- 4 physical locations
  - production site
  - backup 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
- side-project servers
- Analytics infrastructure
- Mail relays …

Ganeti helped seamlessly migrate our infrastructure to a new site (twice!).

skroutz

Ganeti at skroutz (2)

A *single* Ganeti cluster with…

- 25 nodes
- 3 nodegroups (one per location)
- 180+ KVM instances
- DRBD (using secondary IPs)
- ganeti-instance-image
- ganeti-os-d-i (more later)

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

Interesting bits

skroutz

Staging instances

- We run a single Ganeti cluster for all our needs.
- Staging ”cluster”: 3-4 instances (app server, ES server, DB server), with iSCSI-backed disks.
- Automated instance creation and cleanup via RAPI.
- Modified Ganeti Manager to allow cluster creation using a multi-alloc RAPI call.
- Wrapped RAPI to restrict instance operations via Ganeti Manager’s interface to specific domain suffixes only.

skroutz

systemd integration

- Almost all nodes run Jessie with systemd.
- Normally KVM instances appear in `ganeti.service` (or `cron.service`…) cgroup
  - Not especially pretty
  - Does not allow setting individual process limits easily, e.g. using `systemctl set-property`

skroutz

systemd integration (2)

- systemd allows creating *scope* units corresponding to externally managed processes and optionally placing them under a different *slice*.
- Idea: use `systemd-machined`'s DBUS interface to create scope units for VMs.
- Implemented as a post-{create,startup,migrate,failover} hook, but code is minimal enough to include directly in `hv_kvm`.
- All KVM instances happily reside in `ganeti.slice`, `machinectl` shows instances running on the node.

skroutz

ganeti-os-di

- `ganeti-instance-image`: fast and reliable, but...

skroutz

ganeti-os-di

- `ganeti-instance-image`: fast and reliable, but...
- ...managing OS images is *hard*:
  - needs regular updating (point release updates)
  - needs careful cleaning before use (logs, puppet certificates, SSH keys)
  - error prone: e.g. hooks did not clean up ECDSA SSH host keys properly

skroutz

ganeti-os-di

- `ganeti-instance-image`: fast and reliable, but...
- ...managing OS images is *hard*:
  - needs regular updating (point release updates)
  - needs careful cleaning before use (logs, puppet certificates, SSH keys)
  - error prone: e.g. hooks did not clean up ECDSA SSH host keys properly
- Idea: run the Debian Installer in a small KVM instance for the setup:
  - unsafe writeback caching for speed
  - integrates well with our preseeding config
  - instances are created up to date, no need for dist-upgrades
  - no installer code runs in the node's context
  - small speed penalty: 2min 30s instead of 1min
  - implemented as a "traditional" OS provider

skroutz

Small things we miss

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

skroutz

Thank you!

Q&A

skroutz