Ganeti @ GRNET

George Kargiotakis

kargig@grnet.gr

whoami

Supervising "servers and services team" @ GRNET

Working at GRNET for 5+ years

What is GRNET?

GRNET manages fiber & IP networks, datacenters, servers and services

Provides services to Universities, Research Institutions and Government

- >100 Points of Presence in Greece
- ~ 10.000km private fiber
- 2 DCs
- > 300 Servers
- > 8000 VMs

NOC+Dev teams @ GRNET

- NOC Servers team manages Ganeti clusters + ~700VMs with various services
  - From DNS servers to Virtualization platforms and web-applications
- Developers team... develops :)
  - Multiple projects, synnefo/~okeanos is just one of them
  - Open-source (GPL-licensing)

GRNET Ganeti clusters

- Currently running clusters in 5 locations
- 2 large DCs and 3 smaller locations
- 2 distinct Virtualization platforms based on Ganeti
  - ViMa
  - ~okeanos

ViMa - clusters

- ~1600 VMs
- ~130 Nodes
  - ~90 x Fujitsu PRIMERGY RX200 S5
  - ~20 x Dell PowerEdge R430/R630
  - 12 x HP ProLiant BL460c G1/G6
  - 8 x HP ProLiant DL380 G7
  - 5 x Dell PowerEdge R710/R720
  - 5 x Dell PowerEdge 1950/2950
  - 2 x IBM ThinkServer RD350
  - 2 x IBM System x3550 -[7978B1G]-
- ~20 Node Groups (>12 non-default Node Groups)
- 14 clusters (from 1 to 35 hardware nodes)
- 5 locations

~okeanos - clusters

- ~7000 VMs
- 180+ Nodes
  - ~180 x HP ProLiant DL385 G7
  - 2 x Dell PowerEdge R72
- 13 clusters on 14 full racks
- 1 location

Ganeti storage backends

...it's complicated

#Clusters	Ganeti	Storage
1	2.12	Shared block over FC (NetApp)
3	2.12	NFS (EMC)
13	2.10	DRBD + Archipelago (RADOS)
2	2.12	DRBD
1	2.12	DRBD + NFS (EMC)
1	2.15	iSCSI ExtStorage (NetApp)
1	2.12	DRBD + iSCSI
5	2.12	special purpose clusters (single machine or dual machine clusters, even cross-DC)

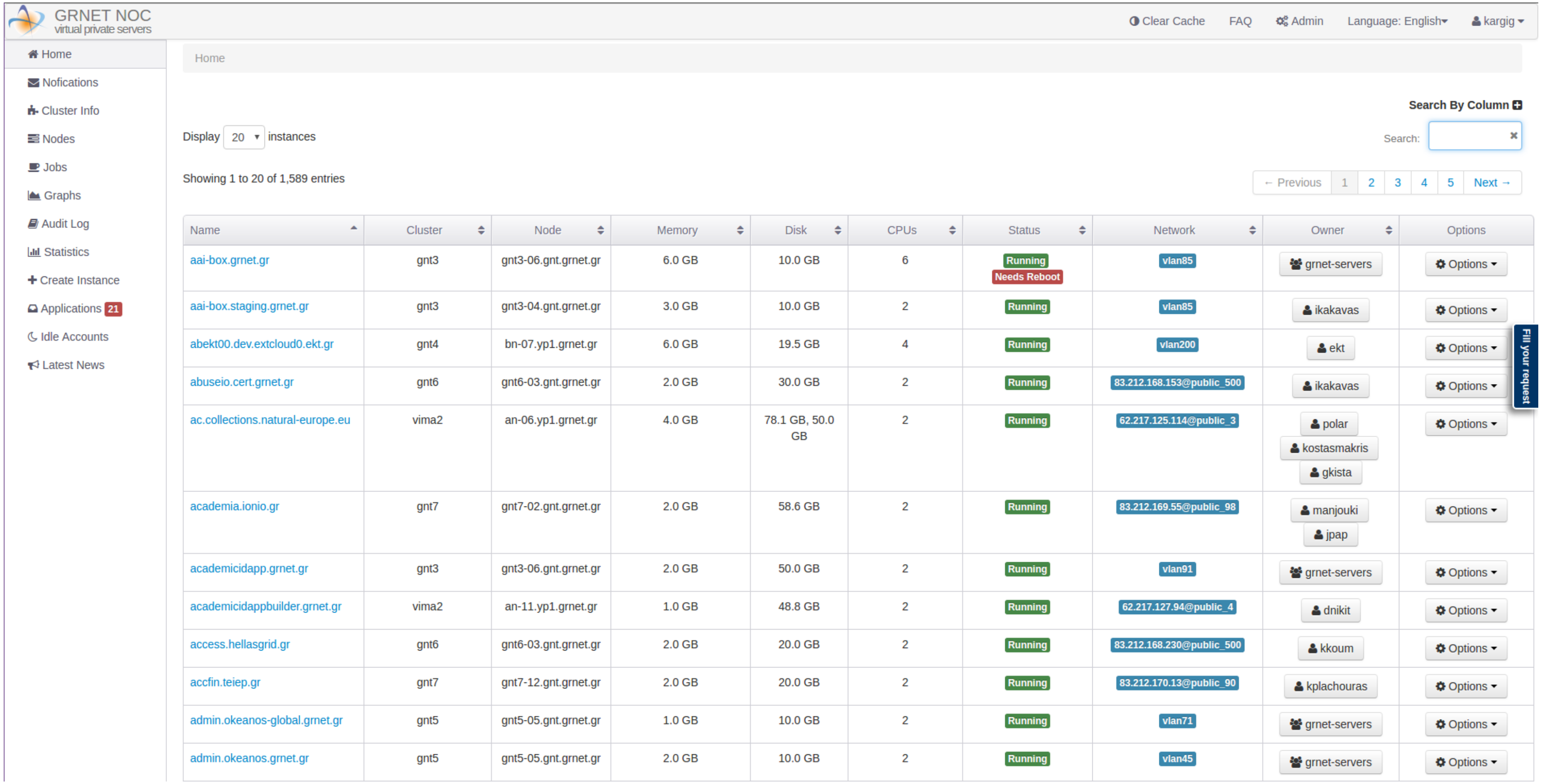
Some clusters have >2 VGs for DRBD w/ hardware or software raid

- SSD/15k/10k RPM disks

Versions

Debian Version	Kernel	Ganeti	qemu-kvm
Wheezy	linux-image-3.2	snf-ganeti 2.10 (heavily patched)	2.1 (bpo)
Wheezy	linux-image-3.16 (bpo)	ganeti 2.12 (bpo)	1.1.2
Jessie	linux-image-3.16	ganeti 2.12	2.1
Jessie	linux-image-3.16	ganeti 2.15 (bpo)	2.1

ViMa - ganetimgr



Home

Search By Column

Display 20 instances

Search:

Showing 1 to 20 of 1,589 entries

← Previous 1 2 3 4 5 Next →

Name	Cluster	Node	Memory	Disk	CPUs	Status	Network	Owner	Options
aai-box.grnet.gr	gnt3	gnt3-06.gnt.grnet.gr	6.0 GB	10.0 GB	6	Running Needs Reboot	vlan85	grnet-servers	Options
aai-box.staging.grnet.gr	gnt3	gnt3-04.gnt.grnet.gr	3.0 GB	10.0 GB	2	Running	vlan85	ikakavas	Options
abekt00.dev.extcloud0.ekt.gr	gnt4	bn-07.yp1.grnet.gr	6.0 GB	19.5 GB	4	Running	vlan200	ekt	Options
abuseio.cert.grnet.gr	gnt6	gnt6-03.gnt.grnet.gr	2.0 GB	30.0 GB	2	Running	83.212.168.153@public_500	ikakavas	Options
ac.collections.natural-europe.eu	vima2	an-06.yp1.grnet.gr	4.0 GB	78.1 GB, 50.0 GB	2	Running	62.217.125.114@public_3	polar kostasmakris gkista	Options
academia.ionio.gr	gnt7	gnt7-02.gnt.grnet.gr	2.0 GB	58.6 GB	2	Running	83.212.169.55@public_98	manjouki jpap	Options
academicidapp.grnet.gr	gnt3	gnt3-06.gnt.grnet.gr	2.0 GB	50.0 GB	2	Running	vlan91	grnet-servers	Options
academicidappbuilder.grnet.gr	vima2	an-11.yp1.grnet.gr	1.0 GB	48.8 GB	2	Running	62.217.127.94@public_4	dnikit	Options
access.hellasgrid.gr	gnt6	gnt6-03.gnt.grnet.gr	2.0 GB	20.0 GB	2	Running	83.212.168.230@public_500	kkoum	Options
accfin.teiep.gr	gnt7	gnt7-12.gnt.grnet.gr	2.0 GB	20.0 GB	2	Running	83.212.170.13@public_90	kplachouras	Options
admin.okeanos-global.grnet.gr	gnt5	gnt5-05.gnt.grnet.gr	1.0 GB	10.0 GB	2	Running	vlan71	grnet-servers	Options
admin.okeanos.grnet.gr	gnt5	gnt5-05.gnt.grnet.gr	2.0 GB	10.0 GB	2	Running	vlan45	grnet-servers	Options

ViMa - Virtual Machines

- GRNET VPS platform
- Moderated instance applications (no quotas)
- Used by GRNET NOC + other knowledgable users (university NOCs, government, research)
- Manual cluster selection for instance creation:
  - Multiple clusters: satisfy different needs using different hardware
  - No billing/accounting → every user asks for max resources
    - Some consulting w/ clients needed
  - NOC approves applications and places instances to the appropriate cluster
- Communicates with all ganeti clusters except ~okeanos

ganetimgr - the software

- Django 1.4 application*
- No database of VM information - as "stateless" as possible regarding instances/networks/nodes
  - Database knows users/groups/clusters
  - Link users with VMs using tags
- Communicates with Ganeti over RAPI
  - No ConfD support yet (patches welcome!)
- Lots of caching using redis
- Asynchronous jobs using beanstalk
- Admin-oriented UI

*Django 1.7 patches almost ready for merging...

ganetimgr notable features

Users can

- Boot instance from CD image over HTTP + Boot device selection
- Change network adapter/hard disk type (Paravirtualized or not)
- Add others as co-admins of an instance
- See actions log
- Use VNC over websockets to manage instance
- See resource usage statistics (management use mostly)

Admins can

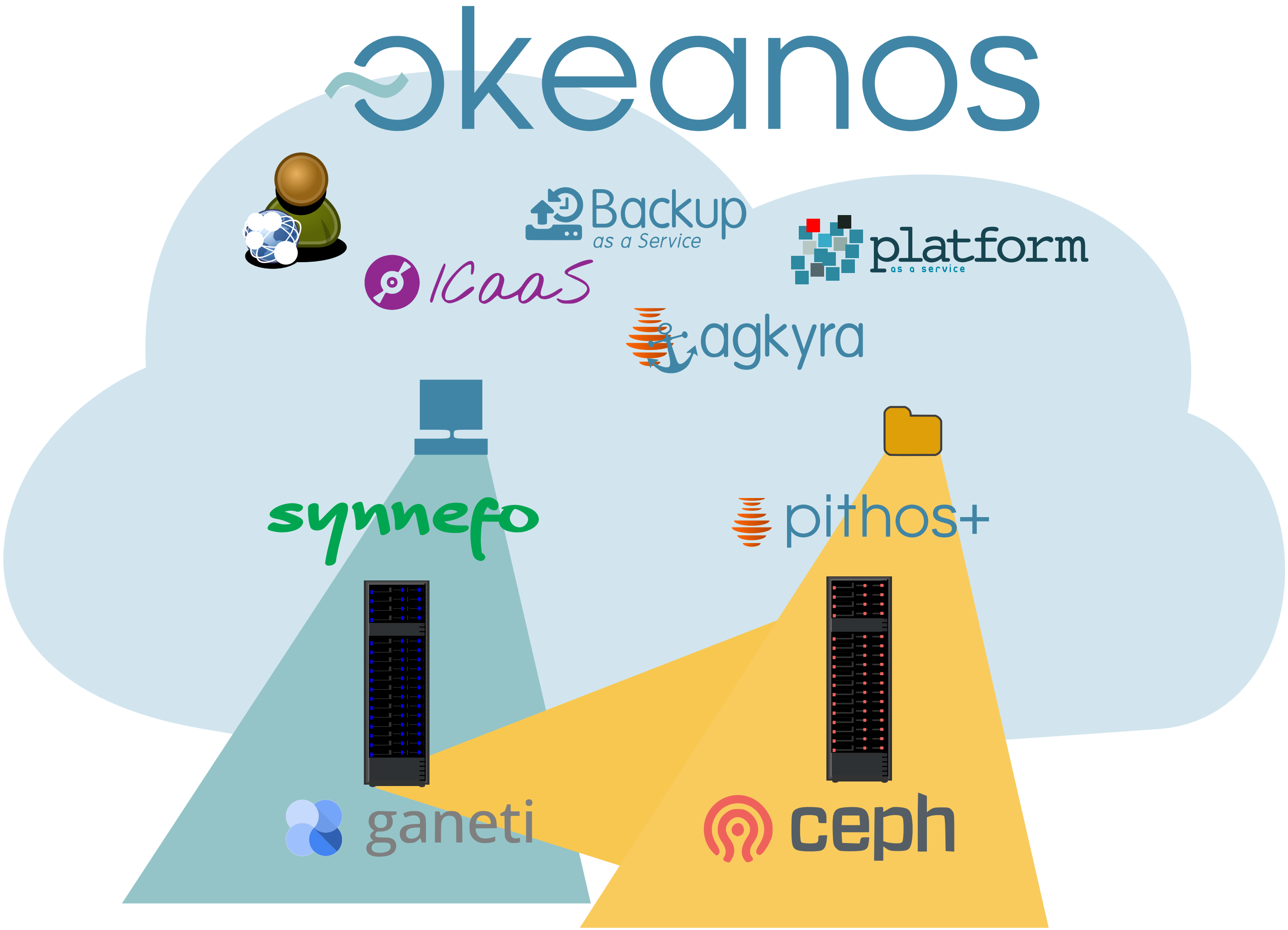
- Email owners of VMs using template syntax
- See per node instance CPU/Network graphs
- (network) Isolate + lock instances from modifications (handle abuse)
- See all users action log

ganetimgr recent changes

since Ganeticon 2015 (v1.5→v1.6)

- snf-image integration (thanks to Brian Candler)
- NoVNC transfer commands from text area
- Improved search filter (CPU, RAM, Cluster, Network, etc)
- Admins can now create instances without having to review their own applications
- OAuth2.0 API providing a user's list of VMs
  - Used by external application (Archiving As A Service - TBA)
- Email notifications archive
- Easier branding
- fabric deployment script

~okeanos - synnefo



~okeanos

- IaaS / cloud service (Compute + Storage)
- PaaS: e.g. Hadoop cluster deployment
- Import images from Bitnami (ICaaS - Image Creation aaS)
- Object Store service with block-based deduplication (Pithos)
  - Backup As a Service for client sync (Agkyra)
- Resource management via Projects
- Fancy UI geared towards users
- Used by thousand end-users for both personal servers and lab-scale infrastructure

synnefo

- Applications based on Django 1.4*
- OpenStack-inspired VM/Volume/ObjectStore API + GRNET extensions
- CLI and Web UI interface
- Multiple authentication backends (local password, shibboleth, LDAP, more)
- Supports multiple ganeti backends
- Ganeti queue monitor agent
- Admin interface

* Django 1.6 & 1.7 upcoming

synnefo notable features

- Batch instance creation/deletion via API
- {Physical,Virtual}-to-virtual snf-image-creator tool
- CLI (kamaki) and Web interface
- VM customization at boot (disk resize, ssh-keys, passwords, network) via snf-image
- Thin provisioning over Ceph/RADOS (Archipelago)
- User-creatable private networks via snf-network+nfdhcpd
- Swap disks between VMs (hotplugging)
- Floating IP(v4) for VMs
- Console support by proxying VNC
- Helpdesk can manage users/Projects

synnefo

Major software changes since last year v0.17

Released: Thu Apr 28 12:35:46 EEST 2016

- Cyclades shared resources among members of a project.
- Cyclades support for detachable volumes
- Brand new pithos UI web application
- Support LDAP authentication in Astakos service

synnefo

Major software changes since last year v0.18

Released: Wed 7 Sep 16:50:30 EEST 2016

- Improved project management and quota policy enforcement
- Performance optimizations of Pithos object listing queries
- Support for modifying user e-mails from the Admin Panel
- Various admin panel enhancements
- Support for multiple eventd instances and automatic ganeti master failover detection
- Support for Sentry

Operations

Installation and Management (or coping, or surviving)

- Debian packages (thanks Apollon!)
- Puppet + Hiera
  - Puppet ENC tells nodes in which cluster they belong
  - Separate Puppet classes per cluster
  - Networks/NFS backend information in hiera
  - DC awareness through API calls to Servermon

Day to Day

- CLI
- hbal
- Mcollective
- evac-gnt-node
- Clustertool

Monitoring 1/3

Icinga plugins

Plugin name	Comment
check_ganeti	check gnt-cluster verify output for errors
check_ganeti_balance	check hbal dry-run improvement score
check_ganeti_freemem	check for memory starving nodes in gnt-node list output
check_ganeti_ippool	check number of free IPs in public pools
check_ganeti_joblist	check number of queued jobs
check_ganeti_nodes	check for DRAINED or OFFLINE nodes w/o special maintenance tags
check_ganeti_queue	check for failed jobs in queue
check_ganeti_watcher	check whether watcher is left paused for too long

TODO: many checks must be rewritten to use ConfD

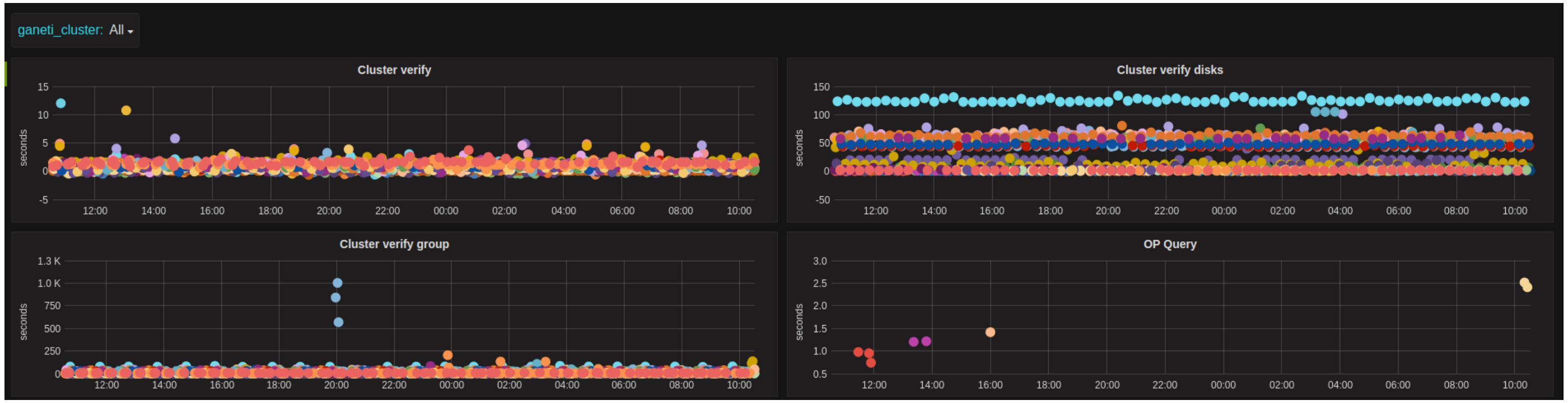
Monitoring 2/3

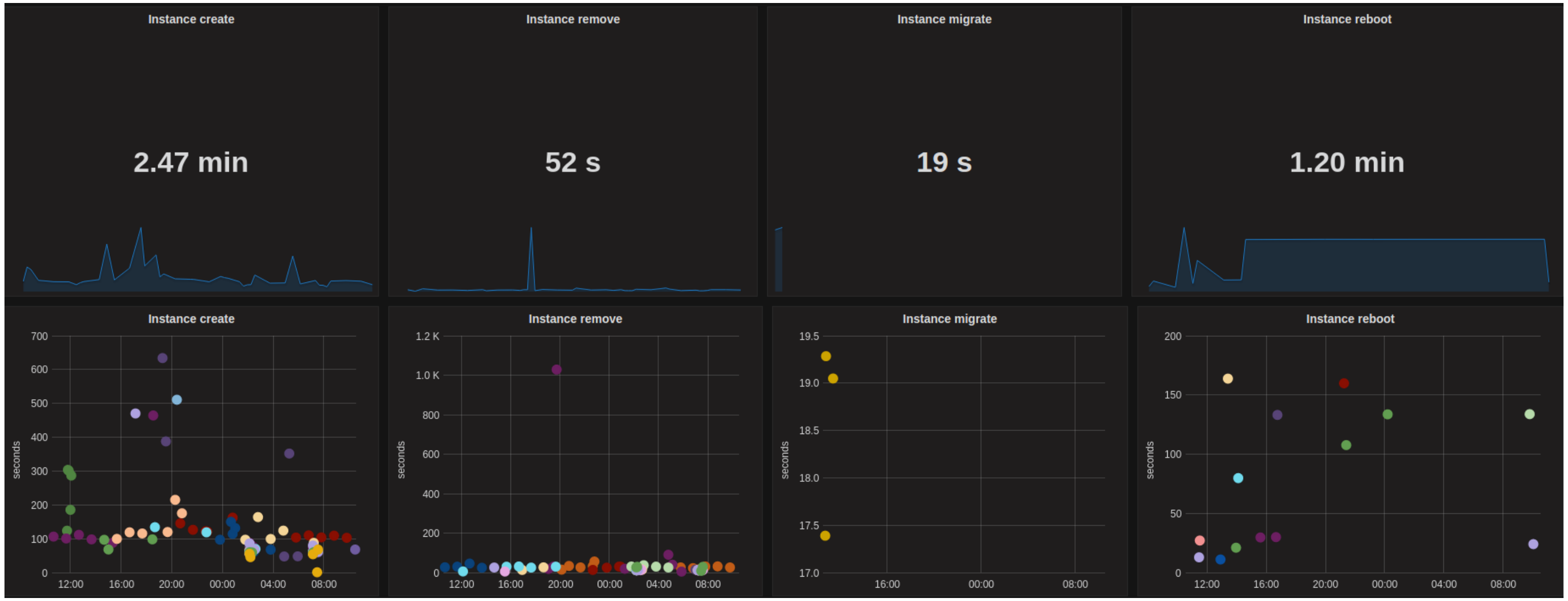
ELK/Graphite/Grafana dashboards

- Log-courier to Logstash
- Logstash parses {jobs, node-daemon, rapi-daemon, wconf-daemon}.log *
- Logstash sends duration and execution times data to Graphite
- Grafana dashboard
  - Time per VM creation/deletion
  - Duration of Cluster verify

TODO: use check_graphite icinga check for outliers

* Ganeti logfile parsing hell (more about this later)


ganeti_cluster: All
Cluster verify
seconds
15
10
5
0
-5
12:00
14:00
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
Cluster verify disks
seconds
150
100
50
0
-50
12:00
14:00
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
Cluster verify group
seconds
1.3 K
1.0 K
750
500
250
0
12:00
14:00
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
OP Query
seconds
3.0
2.5
2.0
1.5
1.0
0.5
12:00
14:00
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00


Instance create
2.47 min
Instance remove
52 s
Instance migrate
19 s
Instance reboot
1.20 min
Instance create
seconds
700
600
500
400
300
200
100
0
12:00
16:00
20:00
00:00
04:00
08:00
Instance remove
seconds
1.2 K
1.0 K
800
600
400
200
0
12:00
16:00
20:00
00:00
04:00
08:00
Instance migrate
seconds
19.5
19.0
18.5
18.0
17.5
17.0
16:00
00:00
08:00
Instance reboot
seconds
200
150
100
50
0
12:00
16:00
20:00
00:00
04:00
08:00

Monitoring 3/3

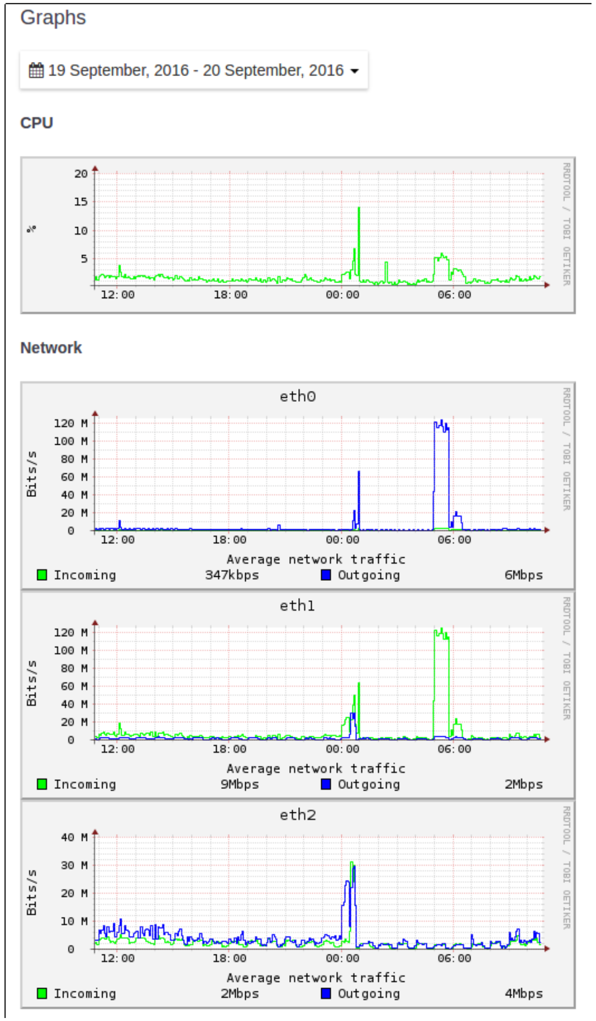
System metrics/graphs

- Munin shows per node statistics
- Ganglia shows cluster-wide metrics

VM metrics/graphs

- vima-grapher
  - collectd python plugin + python wsgi

vima-grapher



Graphs

19 September, 2016 - 20 September, 2016

CPU

Network

eth0

Average network traffic

🟩 Incoming 347kbps 🟦 Outgoing 6Mbps

eth1

Average network traffic

🟩 Incoming 9Mbps 🟦 Outgoing 2Mbps

eth2

Average network traffic

🟩 Incoming 2Mbps 🟦 Outgoing 4Mbps

Ganeti Networking

3+1 Modes

- Bridged
- "Routed"
- Open vSwitch
- MAC-filtered

Public Networking Modes

- Bridged networks (currently only used by GRNET NOC)
- Routed networks with nfdhcpd
  - ARP/ND requests of VMs stay inside the hardware node (arp-proxy, proxy-ndp)
  - Provides DHCP, RAs (SLAAC) and Other Config for DHCPv6
  - Ganeti hooks create files about tap devices configuration (bindings)
  - nfdhcpd listens on NFQUEUE, reads bindings and receives/sends packets on tap devices

Private Networking Modes

- Bridged networks
  - Usecase: L2VPNs from research institutions/labs
  - Every new one needs provisioning from network team (slow)
  - Network equipment does not like >XXX vlans per port for thousands of DC switch ports
  - Limited number or real vlans (how can we go above >4096 vlans?)
- MAC-filtered "private VLANs" for synnefo/~okeanos
  - Assign MAC-address prefix per user
  - One (real) VLAN carries all traffic
  - ebtables filtering on tap for user prefix
  - *Warning!* Performance penalties noticed (at least with Wheezy/Wheezy-bpo kernels)
  - Not recommended for clusters with a *lot* of VMs/traffic
- Open vSwitch for private and cross-dc networks of VMs (ganeti-ovsd)

Ganeti + Open vSwitch

Why: We need *cross-DC, cross-cluster* private networks with the least possible dependency on vendor specific solutions

Considerations

- Ganeti supports Open vSwitch link type
- OVSDB is faster that querying RAPI
- Ganeti does not provide an external event handler
  - Difficult to scale ganeti hooks for every event

Our approach

- Use topological changes seen by switch instead of using ganeti hooks
- Create a dedicated ovs bridge with single VXLAN tunnel port
- Modify kvm-vif-bridge to add special tags to OVSDB `external_ids`

ganeti-ovsd Design doc 1/3

- Add a new instance tag for every openvswitch link (tap)
  - `external_ids`: `grnet_private_lan=iface:ethX:lan_id:1234`
  - `lan_id` is VXLAN VNI
- Learning:
  - Use *Nicira Learn OpenFlow* extension to learn MAC addresses
  - Local MAC addresses: learn input port and associate with the instance's private LAN, encoded in the `tunnel_id` flow parameter
  - Remote MAC addresses: the switch also learns the tunnel endpoint (IP)

ganeti-ovsd Design doc 2/3

Pipeline

- stage 0: *Filtering*
  - Drop unwanted traffic (eg multicast source mac)
- stage 1: *Port-based LAN classification*
  - One can assign physical ports to a VNI
- stage 2: *Learning*
  - Learn per-MAC tunnel endpoints from VXLAN traffic
  - Learn about locally connected MACs
- stage 3: *Output pre-processing*
  - Always flood multicast/broadcast traffic directly
  - Try learned rules, flood otherwise
- stage 4: *Output port selection*

ganeti-ovsd Design doc 3/3

Handling Broadcast, Unknown, Multicast instance traffic

- Flood (BUM) traffic using multicast
  - VXLAN is UDP
  - easy mapping of adminstratively scoped IP multicast block (RFC2365) (239.192.0.0/16 → 65535 private networks)
    - VNI 10 → 239.192.0.10
    - VNI 20 → 239.192.0.20
  - No need for OpenFlow controller

Another approach

- Flood (BUM) traffic using unicast
  - Would lead to traffic amplification
  - Needs OpenFlow controller to keep track which node has VMs for which VNIs

ganeti-ovsd

Implementation

- kvm-vif-bridge adds tap to ovs switch and sets external_ids
- ganeti-ovsd daemon in python
  - Creates initial flow rules for ovs switch
  - Monitors OVSDB for port change events + changes in external_ids
  - Subscribes to multicast groups for each private lan ID (VNI)
- Simple and effective
  - No Ganeti modifications needed
- Currently only supports IPv4 multicast groups
- Bonus: VM tap rate limiting on ovs switch using classifier tags
- Code soon on github

Written by Apollon Oikonomopoulos

Security considerations: Cross-DC setups need to protect multicast traffic from leaking outside of the network

New DCs

- 3 new datacenters to be deployed
  - VMC (= VM Container) + SC (= Storage Container) + Traditional Storage
- ~700 New compute nodes
  - ~600 VMCs (20 cores, 192Gb RAM, 2x300Gb SAS disks)
  - ~100 enhanced VMCs (20 cores, 384Gb RAM, 2x300Gb SAS + 4x900Gb SSD disks)
- ~140 SCs (16 cores, 128Gb RAM, 2x300Gb SAS + 6x200Gb SSD + 12x4TB SATA disks)
  - probably for RADOS (userspace RBD using ExtStorage)
- 2 DCs w/ additional NetApp Storage
- 1 DC w/ only distributed storage

We want to run Ganeti there as well!

but...

will it scale ?

Can we reach 30-40.000 **manageable** VMs?

2000-5000 VMs per cluster feasible ?

Need to explore options

- Queue concurrency
- Lots of clusters vs fewer clusters and more node-groups?
- Smarter allocator/interactions with external APIs for CPU load/IOPS weight
  - We could use cgoups in tags but is there any planned cgroup support by Ganeti ?
- Distributed storage handling (RBD or what ?)

Problems with Ganeti

Documentation

- Lack of good documentation
  - HOWTO guides
  - Whitepapers for specific setups
- Status of design doc implementation is not clear
  - at least without looking at the code
- Object UUIDs frequently exposed to errors, cli instead of friendly names

Automation

- Hard to manage cluster settings in an automated way
  - Anyone has cluster settings in puppet/chef/salt ?

Upgrades

(Gnu)TLS issues when upgrading from 2.12 → 2.15 (Yeap, it's Debian specific but that's Ganeti's most used platform)

Anyone who has upgraded knows what I'm talking about.

- Most painful upgrade so far
- More testing definitely needed, can we help somehow?

Default cluster init values

Not good enough for modern (10GbE) networking

- Sloooow migrations / DRBD sync / replace-disks
  - Can be amazingly improved by adding/changing 3 lines in the config
- New qemu-kvm migration algorithms are available
  - Why not automatically switch to them when possible?
- *[Feature req]* Ability to override cluster migration_* settings per node group and fallback to cluster values when migrating VMs from one node group to the other
- *[Feature req]* Networking Profiles for hardware nodes

Locking/Scheduler concurrency

- Has definitely improved but...
- Long running jobs delay tens of minor ones from starting
  - Predictive scheduler looks very promising!
- `Detected death of job` issue

DRBD timeouts

- DRBD sometimes fails to release devices

-
  ```
  Error 28: Operation timed out after 900433 milliseconds with 0 out of
  -1 bytes received
  ```
- Further investigation needed

Mixed logging format

- HTTP like logs + RunCmd + other info + multi-line exceptions in same file
  - really painful parsing
- Huge json `Validating hv kvm` lines in logs
  - node-daemon.log is chaotic
- Sometimes INFO is too talkative w/o being informative for the operators
  - notable example is wconf-daemon.log
- JobFile with subjobs parsing
  - A job with subjobs writes the json of the completed subops constantly on a new event. Anyone who monitors the file gets "duplicate" entries.
  - *[Feature req]* Separate the jsons of each subjob or split to different file

node-group awareness

A node needs to know its node-group so to expose it to puppet and get from Hiera the proper storage backend

- Our implementation: cron in cluster master writes a file to each node via ssh (meh)
    - Time to use ConfD maybe

Clearing OS parameters

- Cannot re-install VM using different OS provider because of different image properties
  - *[Feature req]* Remove image properties
  - *[Feature req]* Optional reset of image properties on reinstall

Our TODO list

- RBD ExtStorage Driver
  - Bypass Ganeti's pipeline that needs an existing block device
  - Flexibility
- Unfork/rename snf-* Ganeti-related packages
  - Make them easier to be used by vanilla Ganeti setups
  - Make them more open → get more contributors
- Public image directory to be used by snf-image installations
- cgroups
  - We need resource pools, at least for I/O and CPU
- Accounting
  - Better/More precise resource usage statistics
  - Take advantage of monitoring-daemon

Most needed features

- Many clients ask for small-fast OS disk and huge-slower data disk
  - If we give them two slow disks they are not happy
  - If we give them two fast disks we are not happy
- Instance with multiple disks from different pools of the same storage backend
  - Multiple DRBD on different VGs works (metavg though...)
  - Multiple NFS pools with ExtStorage should work
- Instance with multiple disks from different storage backends
  - much much needed
  - DRBD for OS + NFS/RBD for data ? :)
- We're still waiting for gnt-disk and macvtap support to be finalized/reviewed/merged...

Thank you!

Questions?