CERN

Integrating Oracle VM into an Enterprise-Grade OpenStack Cloud: CERN Case Study

Ignacio Coterillo, Giacomo Tenaglia
icoteril@cern.ch, gtenagli@cern.ch

October 1st, 2014

Table of Contents

Introduction

- About CERN
- CERN openlab

CERN Agile Infrastructure

- Overview
- Monitoring in the CERN AI

IT-DB Infrastructure

- About the IT-DB Group
  - The Oracle service
  - The On Demand Services
    - Overview
    - IT-DB Infrastructure

Oracle OVM Integration in OpenStack

- Early steps
- How is it now
- Nova computing
- AI Monitoring integration

Next Steps

Acknowledgements

Table of Contents

Introduction

- About CERN
- CERN openlab

CERN Agile Infrastructure

- Overview
- Monitoring in the CERN AI

IT-DB Infrastructure

- About the IT-DB Group
  - The Oracle service
  - The On Demand Services
- Overview
- IT-DB Infrastructure

Oracle OVM Integration in OpenStack

- Early steps
- How is it now
- Nova computing
- AI Monitoring integration

Next Steps

Acknowledgements

About CERN

- Founded in 1954
- Research: **Seeking and finding answers to questions about the Universe**
- Twenty one member states
- Seven observer states and organizations: India, Japan, the European Comission, the Russian Federation, Turkey, UNESCO, and the USA
- Cooperation and scientific agreements with over 55 additional countries

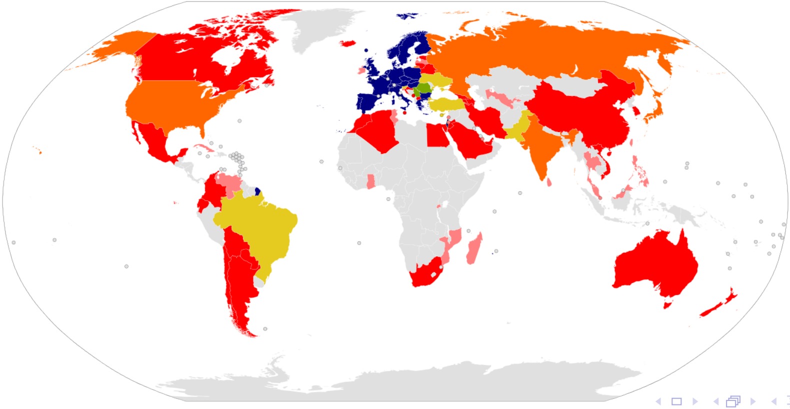
YEARS / ANS CERN

1954 2014

About CERN

People

∼ 2400 Staff, ∼ 10000 Users from 113 countries, ∼ 2000 contractors

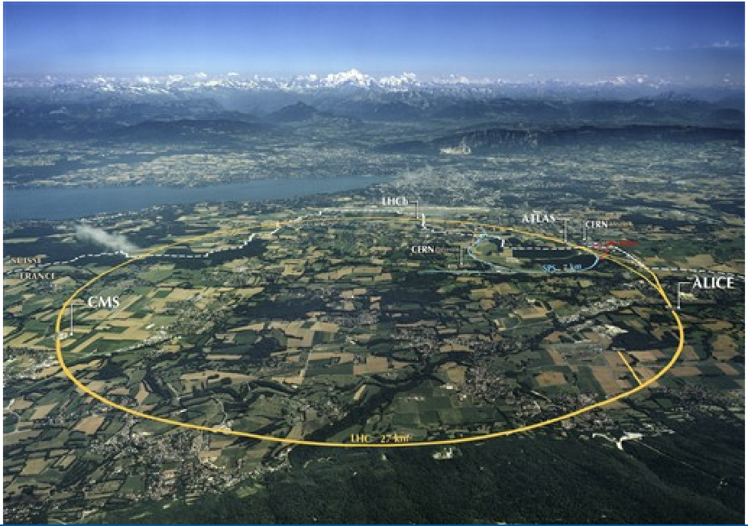


LHC, Experiments, Physics

Large Hadron Collider (LHC)

- World's largest and most powerful particle accelerator
- 27km ring of superconducting magnets
- Current undergoing upgrades, will restart in 2015
- The products of particle collisions are captured by complex detectors and analyzed by software in teh experiments dedicated to the LHC

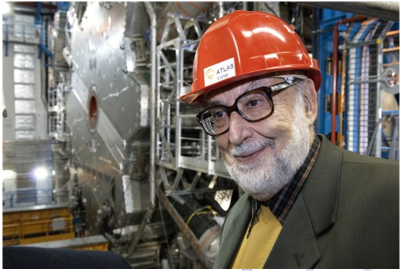
LHC, Experiments, Physics



LHC, Experiments, Physics

The Higgs Boson

The Nobel prize in Physics 2013 was awarded jointly to Francois Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that congributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*



LHC, Experiments, Physics

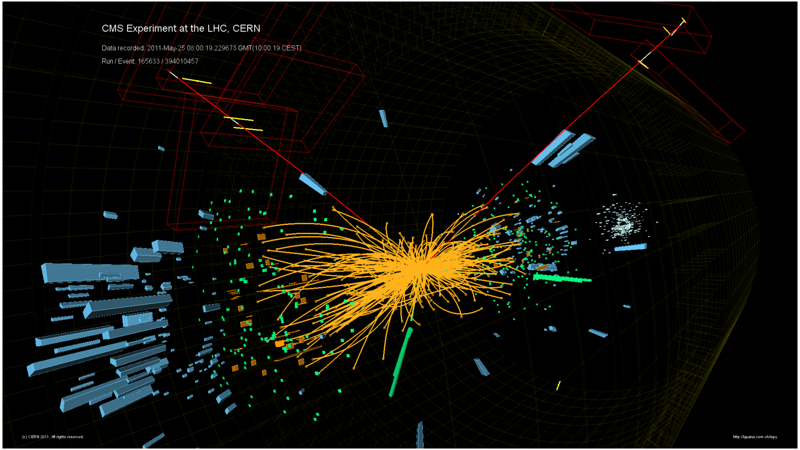

CMS Experiment at the LHC, CERN
Data recorded: 2011-May-25 08:00:19.229673 GMT(10:00:19 CEST)
Run / Event: 165633 / 394010457
(c) CERN 2011. All rights reserved.
http://iguana.cern.ch/ispy

Figure: Higgs boson decaying to ZZ candidate event

LHC Computing and storage needs

Data volume

- More than 100 Petabytes of data stored and analyzed
- Increasing $\sim$ 25 PB per year
- Over 160 computer centres in 35 countries
  - $\sim$ 260 000 CPU cores
  - $\sim$ 269 PB disk capacity
  - $\sim$ 210 PB tape capacity

CERN openlab

- Public-private partnership between CERN and leading ICT companies
- Currently in its fourth phase. It started in 2003
- Its mission is to accelerate the development of cutting-edge solutions to be used by the worldwide LHC community
- Innovative ideas aligned between CERN and the partners.

HUAWEI

intel

ORACLE

SIEMENS

rackspace

the #1 managed cloud company

Yandex

CERN openlab

Oracle and the CERN openlab

Research collaboration on several areas:

- Database replication
- Data Analytics
- Database Monitoring
- Physics analysis on the database
- Virtualization
- J2EE

Table of Contents

Introduction

- About CERN
- CERN openlab

CERN Agile Infrastructure

- Overview
- Monitoring in the CERN AI

IT-DB Infrastructure

- About the IT-DB Group
  - The Oracle service
  - The On Demand Services
- Overview
- IT-DB Infrastructure

Oracle OVM Integration in OpenStack

- Early steps
- How is it now
- Nova computing
- AI Monitoring integration

Next Steps

Acknowledgements

Motivation for CERN AI

What is CERN AI?

A new way of looking at how to manage the CERN Computer Centre, involving new strategies, tools and philoshopy.

Rationale

- Need to manage increasing (doubling) number of servers with no increasing staff
- Old tools are difficult to maintain and will not scale

Approach

- CERN is no longer a special case for compute
- Adopt an open source tool chain model
- If we have special requirements, challenge them
- If useful, contribute back

CERN AI Main components

Server Virtualization

- Trying to maximize the number of virtualized hosts
- Offer computer resources as a service
- Cloud "Operating system": **OpenStack**

Configuration Management

- **Puppet** as configuration management system
- **Foreman** as machine inventory tool

CERN AI Main components

OpenStack

"A cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives admninistrators control while empowering their users to provision resources through a web interface"

Multi hypervisor

- OpenStack Compute (Nova) has an abstraction layer for compute drivers, what allows you to choose which hypervisor(s) to use.
  - Not all of them are equally supported
- CERN current production deployment uses KVM and Hyper-V
  - Different hypervisors for different workloads
  - Hence the interest for integrating Oracle VM...

CERN AI Monitoring

Motivation

- Uniformity: Several independent monitoring activities in IT with similar approach and limitations, but different tool-chains
- Interdependency: Combination of data from different groups necessary, but difficult
- Performance monitoring becoming more relevant, requiring combined data and complex analysis.
- Migration to a virtualized dynamic infrastructure involves new requirements on monitoring

CERN AI Monitoring

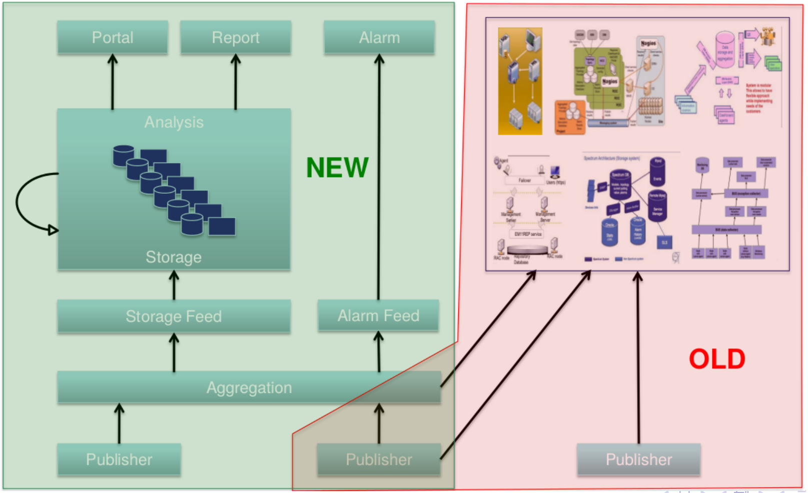


Table of Contents

Introduction

- About CERN
- CERN openlab

CERN Agile Infrastructure

- Overview
- Monitoring in the CERN AI

IT-DB Infrastructure

- **About the IT-DB Group**
  - The Oracle service
  - The On Demand Services
- **Overview**
- **IT-DB Infrastructure**

Oracle OVM Integration in OpenStack

- Early steps
- How is it now
- Nova computing
- AI Monitoring integration

Next Steps

Acknowledgements

The Oracle service

CERN Databases

- $\sim$ 100 Oracle databases, most of them RAC
  - Mostly NAS storage plus some SAN with ASM
  - $\sim$ 500 Terabytes of data file for production databases in total
- Example of critical production databases:
  - LHC logging database, currently at $\sim$ 170 TB, with an expected growth of 70 TB per year
  - 13 experiment databases between 10 and 20 TB each
  - Read only copies (Active Data Guard)

The On Demand Services

The Database on Demand platform

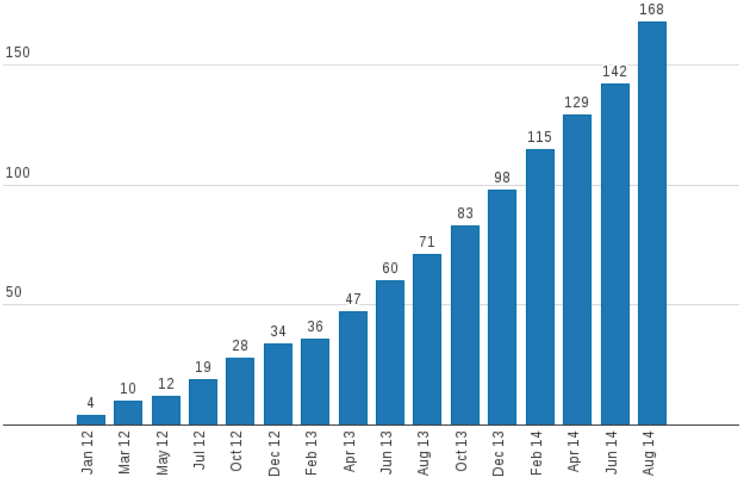
- Covers a demand from CERN community not addressed by the Oracle service
  - Users have *full* DBA privileges
  - Different RDBMS: MySQL, PostgreSQL and Oracle
- Provides automatized DBA operations: configuration, shutdown and startup, upgrades, backup and recovery operations and monitoring.
- Currently hosting $\sim$ 170 databases

The Middleware on Demand platform

- Similar concept targeting application servers
- Just launched to production

The Database on Demand Service

Evolution of the amount of MySQL, Oracle, and PostgreSQL instances in the DBOD service



IT-DB Infrastructure Overview

The big picture

- Closer placement to the IT-DB storage systems
- Specific configuration requirements (networking)
- Software licenses management

Migration process

- Started on Q2 2013
- Expected to be finished by the end of Q4 2014

IT-DB Infrastructure

Legacy infrastructure

- ∼ 500 servers
- ∼ 700 services (databases, application servers,...)
- 35 Oracle VM 2 hypervisors
  - 270 CPU cores, 1.5 TBi RAM Memory
  - ∼ 125 Virtual machines
- Storage: Netapp 3240 in 7-mode
  - 20 filers
  - ∼ 300 TBi

What we are migrating to

- 14 OpenStack Hypervisors
  - 450 CPU cores, 1.5 TBi RAM Memory
  - ∼ 120 Virtual machines
- 16 OpenStack Hypervisors being installed
  - 500 CPU cores, 2.0 TBi RAM Memory
- Storage: Netapp 6220 and 8060 in C-mode
  - 5.48PBi, 1.46 PBi Used

Some partial conclusions

About IT-DB

- Fairly heterogenous ecosystem
  - Services
  - Infrastructure
- On Demand projects are specially suited for virtualization

A great opportunity

Using Oracle VM as an OpenStack hypervisor gives up the chance of having an homogenous infrastructure across all the CERN IT ecosystem.

Table of Contents

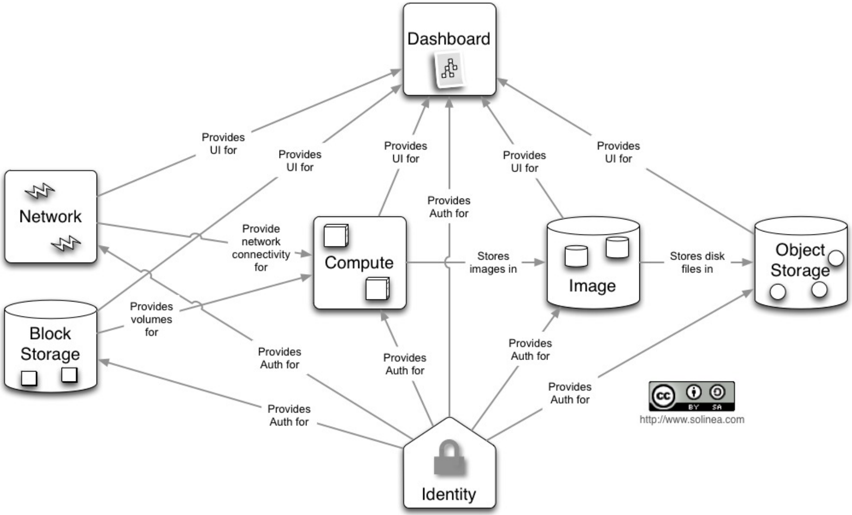
- **Introduction**
  - About CERN
  - CERN openlab
- **CERN Agile Infrastructure**
  - Overview
  - Monitoring in the CERN AI
- **IT-DB Infrastructure**
  - About the IT-DB Group
    - The Oracle service
    - The On Demand Services
  - Overview
  - IT-DB Infrastructure
- **Oracle OVM Integration in OpenStack**
  - Early steps
  - How is it now
  - Nova computing
  - AI Monitoring integration
- **Next Steps**
- **Acknowledgements**

Why are we doing this?

Continuation of previous collaboration

During the past few years, CERN and Oracle have collaborated researching and testing in the field of virtualization:

- Networking performance under Oracle VM with SR-IOV
- Testing and evaluation of Oracle VM
- Oracle VM integration with Oracle EM


Dashboard
Network
Compute
Image
Object Storage
Block Storage
Identity
Provides UI for
Provides UI for
Provides UI for
Provides UI for
Provides UI for
Provides Auth for
Provide network connectivity for
Stores images in
Stores disk files in
Provides volumes for
Provides Auth for
Provides Auth for
Provides Auth for
Provides Auth for
Provides Auth for
http://www.solinea.com

Early steps: Notes about OpenStack RDO installation

Supossedly straighforward

1. Run packstack –allinone
2. Add extra nova nodes to config file and re-run

In reality

- Problems with dependencies versions ⇒ YUM repository priorities
- Bugs:
  - Required services not being started (MongoDB/Ceilometer)
  - Some python modules not having the right imports
- Fast iteration: Configuration parameters changing names day to day

Early steps: A custom Oracle VM hypervisor

Why?

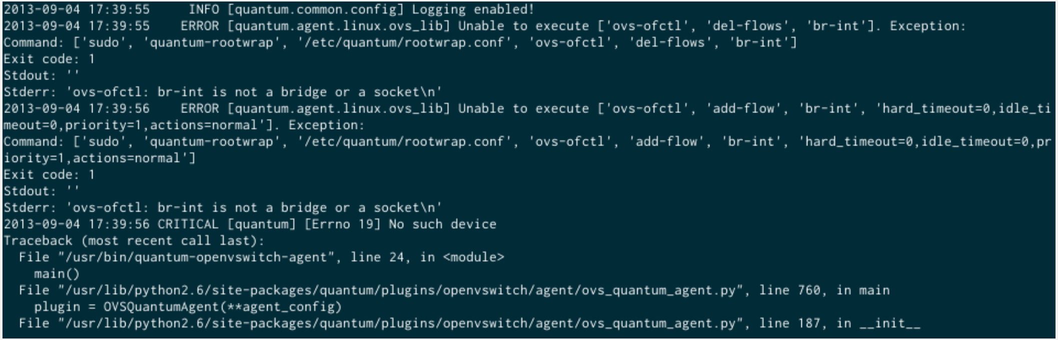
- Our work started between Oracle VM 3.2 and 3.3
- Oracle VM Hypervisor was based on OL5, and following the black-box model
- Impossible to work out OpenStack RDO dependencies
  - Grizzly release at the time we started working

What we did

- Starting from Oracle Linux 6
  - Xen 4.1.6-rc1 compiled from source
  - Libvirt 0.10.2 re-compiled from source to enable Xen support
  - Add node as a nova compute node on an OpenStack RDO installation

Early steps: Issues

Network problems!

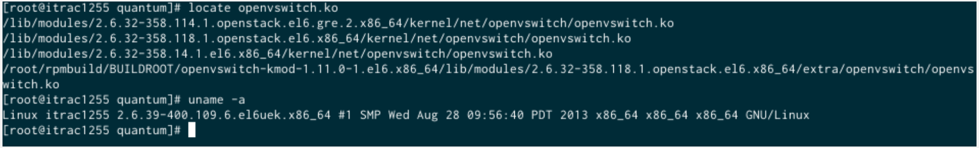


```
2013-09-04 17:39:55     INFO [quantum.common.config] Logging enabled!
2013-09-04 17:39:55    ERROR [quantum.agent.linux.ovs_lib] Unable to execute ['ovs-ofctl', 'del-flows', 'br-int']. Exception:
Command: ['sudo', 'quantum-rootwrap', '/etc/quantum/rootwrap.conf', 'ovs-ofctl', 'del-flows', 'br-int']
Exit code: 1
Stdout: ''
Stderr: 'ovs-ofctl: br-int is not a bridge or a socket\n'
2013-09-04 17:39:56    ERROR [quantum.agent.linux.ovs_lib] Unable to execute ['ovs-ofctl', 'add-flow', 'br-int', 'hard_timeout=0,idle_ti
meout=0,priority=1,actions=normal']. Exception:
Command: ['sudo', 'quantum-rootwrap', '/etc/quantum/rootwrap.conf', 'ovs-ofctl', 'add-flow', 'br-int', 'hard_timeout=0,idle_timeout=0,pr
iority=1,actions=normal']
Exit code: 1
Stdout: ''
Stderr: 'ovs-ofctl: br-int is not a bridge or a socket\n'
2013-09-04 17:39:56 CRITICAL [quantum] [Errno 19] No such device
Traceback (most recent call last):
  File "/usr/bin/quantum-openvswitch-agent", line 24, in <module>
    main()
  File "/usr/lib/python2.6/site-packages/quantum/plugins/openvswitch/agent/ovs_quantum_agent.py", line 760, in main
    plugin = OVSQuantumAgent(**agent_config)
  File "/usr/lib/python2.6/site-packages/quantum/plugins/openvswitch/agent/ovs_quantum_agent.py", line 187, in __init__
```

Early steps: Issues

Network problems!

- Quantum (OpenStack networking module) requires **openvswitch** and its kernel module
- There was no openvswitch kernel module for the Oracle UEK



```
[root@itrac1255 quantum]# locate openvswitch.ko
/lib/modules/2.6.32-358.114.1.openstack.el6.gre.2.x86_64/kernel/net/openvswitch/openvswitch.ko
/lib/modules/2.6.32-358.118.1.openstack.el6.x86_64/kernel/net/openvswitch/openvswitch.ko
/lib/modules/2.6.32-358.14.1.el6.x86_64/kernel/net/openvswitch/openvswitch.ko
/root/rpmbuild/BUILDROOT/openvswitch-kmod-1.11.0-1.el6.x86_64/lib/modules/2.6.32-358.118.1.openstack.el6.x86_64/extra/openvswitch/openvs
witch.ko
[root@itrac1255 quantum]# uname -a
Linux itrac1255 2.6.39-400.109.6.el6uek.x86_64 #1 SMP Wed Aug 28 09:56:40 PDT 2013 x86_64 x86_64 x86_64 GNU/Linux
[root@itrac1255 quantum]# 
```

- Tried different things but nothing worked...

How is it now

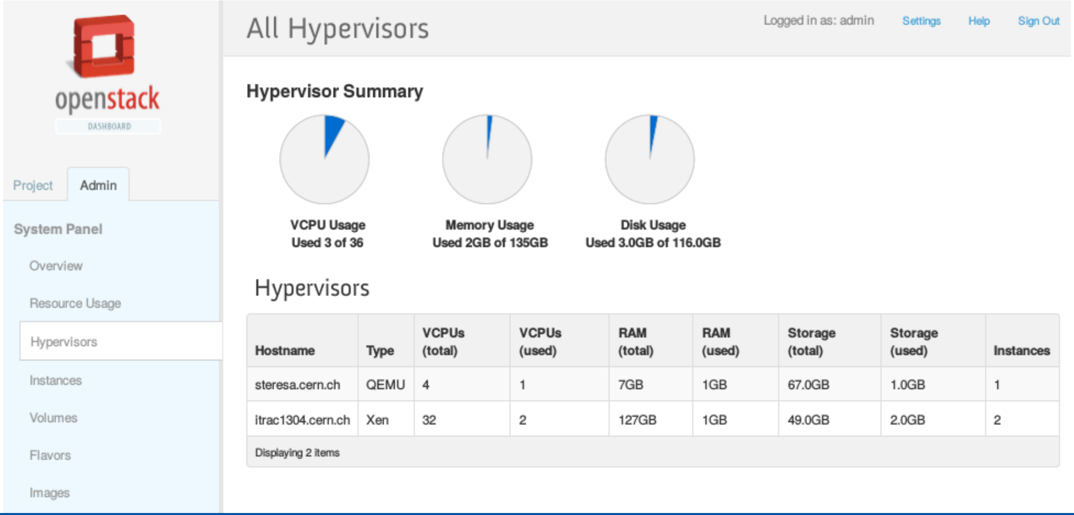
Things happened...

1. OpenStack Havana released ⇒ Quantum now is Neutron...
2. Oracle VM 3.3.1 r776 was released ⇒ No more Xen compiling
3. Oracle OpenStack Beta tech preview released ⇒ No more libvirt compiling

Current procedure

1. Install Oracle VM
2. Install libvirt from Oracle OpenStack YUM repository
3. Add node as a nova compute node on an OpenStack RDO installation
4. Change nova configuration to use Xen as hypervisor

Hypervisor details



All Hypervisors

Hypervisor Summary

VCPU Usage
Used 3 of 36

Memory Usage
Used 2GB of 135GB

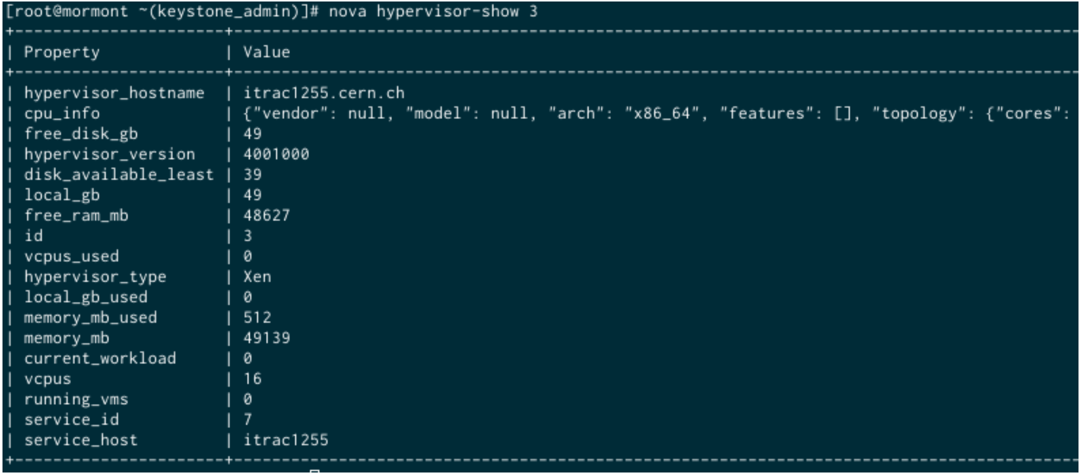
Disk Usage
Used 3.0GB of 116.0GB

Hypervisors

Hostname	Type	VCPUs (total)	VCPUs (used)	RAM (total)	RAM (used)	Storage (total)	Storage (used)	Instances
steresa.cern.ch	QEMU	4	1	7GB	1GB	67.0GB	1.0GB	1
itrac1304.cern.ch	Xen	32	2	127GB	1GB	49.0GB	2.0GB	2

Displaying 2 items

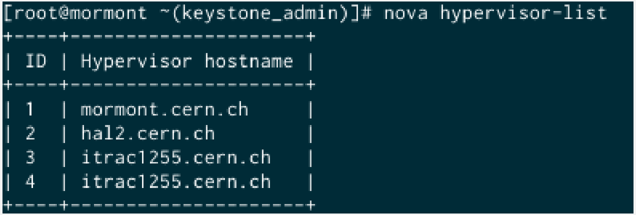
Hypervisor details



```
[root@mormont ~(keystone_admin)]# nova hypervisor-show 3
+----------------------+-------------------------------------------------------------------------------------
| Property             | Value
+----------------------+-------------------------------------------------------------------------------------
| hypervisor_hostname  | itrac1255.cern.ch
| cpu_info             | {"vendor": null, "model": null, "arch": "x86_64", "features": [], "topology": {"cores":
| free_disk_gb         | 49
| hypervisor_version   | 4001000
| disk_available_least | 39
| local_gb             | 49
| free_ram_mb          | 48627
| id                   | 3
| vcpus_used           | 0
| hypervisor_type      | Xen
| local_gb_used        | 0
| memory_mb_used       | 512
| memory_mb            | 49139
| current_workload     | 0
| vcpus                | 16
| running_vms          | 0
| service_id           | 7
| service_host         | itrac1255
+----------------------+-------------------------------------------------------------------------------------
```

Hypervisor details

OpenStack curiosities



```
[root@mormont ~(keystone_admin)]# nova hypervisor-list
+----+---------------------+
| ID | Hypervisor hostname |
+----+---------------------+
| 1  | mormont.cern.ch     |
| 2  | hal2.cern.ch        |
| 3  | itrac1255.cern.ch   |
| 4  | itrac1255.cern.ch   |
+----+---------------------+
```

- Hypervisors 1 and 2 are KVM hypervisors
- Hypervisor 3 is our Oracle VM hypervisor
- Hypervisor 4 is ...

Now...

Nova is ready

- You can try to create instances

A bit of advice

- Beware of automatic system updates
  - Can break your dependencies
  - Can break the enviroment
- Follow Oracle patch submissions to OpenStack

CERN AI Monitoring integration

What is needed?

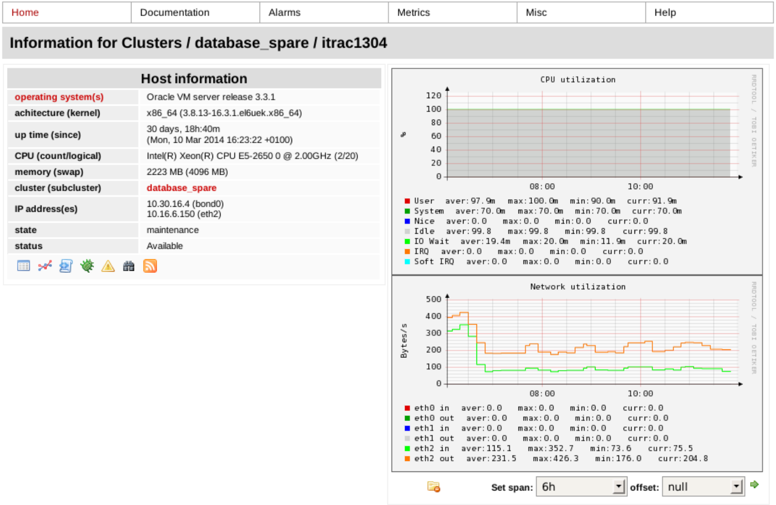
- **CERN monitoring agent**. A server/client based monitoring system, using a push/pull protocol with sensors.
- Apache **Flume** agent. Flume is a distributed service for collecting, aggregating and moving large amounts of log data.

How to?

- In a tipical installation, the agents will be setup by puppet
- In our case:
  1. Set up CERN AI Yum Repositories
  2. Peek at sister machine list of installed packages
  3. Copy configuration files
  4. Set up host certificates

Results: Lemon Metrics

Lemon Monitoring



Information for Clusters / database_spare / itrac1304

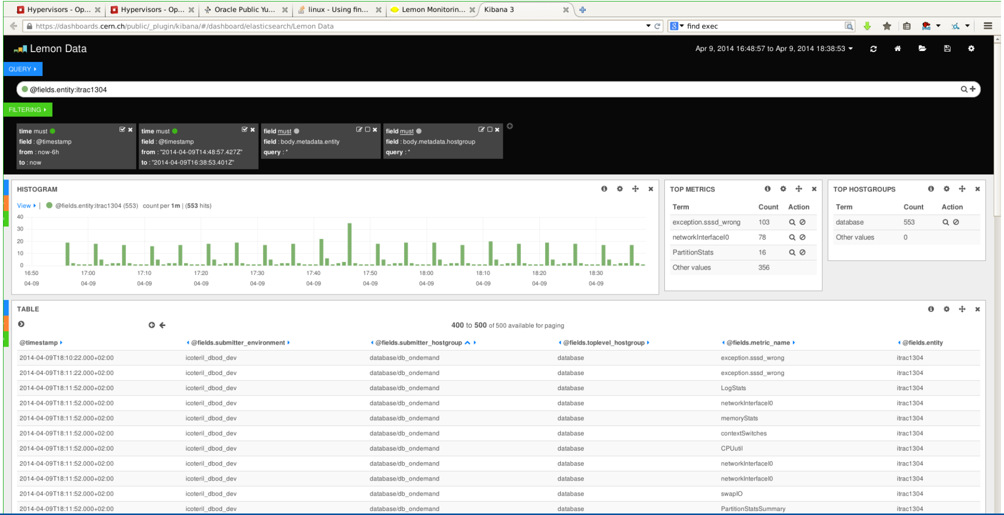
Host information	
operating system(s)	Oracle VM server release 3.3.1
achitecture (kernel)	x86_64 (3.8.13-16.3.1.el6uek.x86_64)
up time (since)	30 days, 18h:40m (Mon, 10 Mar 2014 16:23:22 +0100)
CPU (count/logical)	Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz (2/20)
memory (swap)	2223 MB (4096 MB)
cluster (subcluster)	database_spare
IP address(es)	10.30.16.4 (bond0) 10.16.6.150 (eth2)
state	maintenance
status	Available

```
■ User     aver:97.9m   max:100.0m   min:90.0m   curr:91.9m
■ System   aver:70.0m   max:70.0m    min:70.0m   curr:70.0m
■ Nice     aver:0.0     max:0.0      min:0.0     curr:0.0
■ Idle     aver:99.8    max:99.8     min:99.8    curr:99.8
■ IO Wait  aver:19.4m   max:20.0m    min:11.9m   curr:20.0m
■ IRQ      aver:0.0     max:0.0      min:0.0     curr:0.0
■ Soft IRQ   aver:0.0   max:0.0      min:0.0     curr:0.0
```

```
■ eth0 in    aver:0.0     max:0.0     min:0.0     curr:0.0
■ eth0 out   aver:0.0     max:0.0     min:0.0     curr:0.0
■ eth1 in    aver:0.0     max:0.0     min:0.0     curr:0.0
■ eth1 out   aver:0.0     max:0.0     min:0.0     curr:0.0
■ eth2 in    aver:115.1   max:352.7   min:73.6    curr:75.5
■ eth2 out   aver:231.5   max:426.3   min:176.0   curr:204.8
```

Set span: 6h offset: null

Results: Flume acquistion



@fields.entity:itrac1304

time must ●
field : @timestamp
from : now-6h
to : now

time must ●
field : @timestamp
from : "2014-04-09T14:48:57.427Z"
to : "2014-04-09T16:38:53.401Z"

field must ●
field : body.metadata.entity
query : *

field must ●
field : body.metadata.hostgroup
query : *

HISTOGRAM

TOP METRICS

Term	Count	Action
exception.sssd_wrong	103	
networkInterface0	78	
PartitionStats	16	
Other values	356	

TOP HOSTGROUPS

Term	Count	Action
database	553	
Other values	0	

TABLE

400 to 500 of 500 available for paging

@timestamp	@fields.submitter_environment	@fields.submitter_hostgroup	@fields.toplevel_hostgroup	@fields.metric_name	@fields.entity
2014-04-09T18:10:22.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	exception.sssd_wrong	itrac1304
2014-04-09T18:11:22.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	exception.sssd_wrong	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	LogStats	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	networkInterface0	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	memoryStats	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	contextSwitches	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	CPUutil	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	networkInterface0	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	networkInterface0	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	swapIO	itrac1304
2014-04-09T18:11:52.000+02:00	icoteril_dbod_dev	database/db_ondemand	database	PartitionStatsSummary	itrac1304

Table of Contents

Introduction
- About CERN
- CERN openlab

CERN Agile Infrastructure
- Overview
- Monitoring in the CERN AI

IT-DB Infrastructure
- About the IT-DB Group
  - The Oracle service
  - The On Demand Services
- Overview
- IT-DB Infrastructure

Oracle OVM Integration in OpenStack
- Early steps
- How is it now
- Nova computing
- AI Monitoring integration

Next Steps

Acknowledgements

Next Steps

Starting next week

- Trial production deployment
  - Several Oracle VM hypervisors added to our OpenStack production pools
  - CERN OpenStack Nova upgraded to IceHouse
- Database workload testing and evaluation

Challenges

- Automate Oracle VM Installation (Kickstart based)
- CERN puppet integration

Table of Contents

Acknowledgements

- Ronen Kofman, Monica Marinucci, Greg Doherty

- David Collados, Ruben Gaspar Aparicio, Miroslav Potocki, Lisa Azzurra, Jan van Eldik, Belmiro Moreira, Nacho Barrientos, Pedro Andrade

CERN

www.cern.ch