Portuguese **Distributed Computing** Infrastructure

First implementation of base cloud services (Mi5.1)

Execution Report

State:	**FINAL**
Dissemination:	**PUBLIC**
Authors:	**Jorge Gomes**
URL:	**http://www.incd.pt**

Date	Author	Comments
02-10-2017	Jorge Gomes	Initial version
20-10-2017	Jorge Gomes	Expand services architecture description
15-11-2017	Jorge Gomes	Latest enhancements for 1st implementation

Cofinanciado por:

Portuguese **Distributed Computing** Infrastructure

Executive Summary

INCD aims to support a wide range of scientific domains and projects with very different requirements. Supporting them across all stages of simulation, processing, analysis and data sharing requires a flexible infrastructure on top of which new added value services and specific virtual research environments can be implemented and delivered. These will include services specific to user communities and generic services delivered aimed to the whole scientific community.

Scientific applications increasingly need tailored environments often requiring heterogeneous setups and using multiple services that are difficult to support in conventional computing environments such as batch systems. In addition research is becoming more open and collaborative requiring information sharing and networked services such as scientific portals, databases, and web services among others that need to be directly accessible over the Internet. User requirements are continuously evolving and flexible approaches to computing and data storage provisioning are needed.

This milestone records the first implementation of the INCD cloud services architecture aiming at providing a flexible environment to support complex user applications, virtual research environments and added value services. The INCD cloud service is based on the Openstack cloud management framework.

Openstack

Openstack is an open source cloud management framework, mainly oriented at Infrastructure-as-a-service (IaaS) clouds. It has a modular architecture and consists of a set of components whose combination allows to provide a wide range of functionalities. The development of Openstack began in 2010 and is currently coordinated by the Openstack Foundation, a non-profit entity. Openstack has a significant adoption by commercial, academic and scientific organisations. Openstack enables flexible, scalable IaaS clouds composed of computing, storage and data communication networks. It has in its design the objective of supporting multiple tenants (users and their computing environments) with isolation and security. To better understand the concept of tenant, Openstack can have multiple tenants each one providing an isolated environment with its network environment, storage, machines and users. In this sense a tenant behaves as a virtual datacenter that can be composed with network segments, routers, servers, and storage. Each tenant can have multiple users, and each user can manage the tenant's resources. In Openstack the terms tenant and project are frequently used with the same meaning.

The INCD partners have extensive experience in cloud technologies. The choice of Openstack was therefore natural giving the accumulated experience and availability of community support namely via other infrastructures with which INCD has partnerships, namely IBERGRID at the Iberian level and EGI and the pan-European level. The architecture described in this document supported the installation of the first INCD cloud service.

Openstack architecture

The Openstack components are independent and can be combined in many ways to implement solutions with different functionalities. This modular architecture allows easier development, greater flexibility in implementation, and replacement of components.

Many components can be used both by the users and also by other components. For this purpose, each component offers a REST API. There are client side bindings for various languages that encapsulate the details of communication to facilitate access to the APIs. For the end-user, access can be provided via the command line interface clients or via the Horizon web dashboard. The command line clients offer complete access to all Openstack features, and are very convenient for the creation of scripts for management and automation. The Horizon dashboard is less complete but enables easy and intuitive access to the most important features of Openstack. The dashboard is especially suited for less experienced users or in situations where exposing the APIs to the users is not desirable.

Openstack components selection for the INCD cloud service

Openstack has a modular architecture supported by an ecosystem of projects that develops interoperable components. The selection and combination of these components facilitates the creation of tailored IaaS clouds with a varied range of functionalities. Some components of Openstack are considered nuclear because they provide fundamental functionality. There are also optional components that provide additional functionalities.

The main components that have been selected for the INCD cloud service architecture are:

Keystone

Keystone is the Openstack identity service that implements the tenant and user concepts. It is also through the keystone that the catalogue of services installed on the Openstack infrastructure and their endpoints are published.

Keystone supports several authentication methods, including username and password, federated authentication (SAML and OIDC), and authentication with digital certificates. Keystone is used by all Openstack components for authentication of users and services. INCD tested the Openstack federation capabilities both using SAML and OIDC, the decision for the first implementation was to use OIDC for identity federation purposes.

Nova

Nova manages the lifecycle of the virtual machines instantiated by the users, namely their creation, pause, stop, start, and removal. In addition it manages the virtual machine images in association with Glance, network interfaces in association with Neutron, and storage in association with Cinder. Nova implements the virtual hardware profiles (flavours) that support the execution of the instances. Through these profiles is possible to define a set of virtual hardware configurations with different capacities of RAM, disk, swap and virtual CPUs. Nova supports multiple hypervisors. The INCD cloud uses the KVM hypervisor due to its open source nature, performance, robustness and good support. In addition INCD implemented experimental support for LXD a Linux containers based hypervisor. The INCD setup allows virtual machines to be stored either in the Nova compute nodes local disks or in remote block storage provided via Cinder.

Neutron

Neutron is the virtual network management component that ensures internal communication between virtual machines and external communication between tenant networks and the Internet. Neutron empowers the users to setup their own tenant network environment including virtual network segments, virtual routers and firewall rules. The network virtualization is achieved via virtual network switches implemented by openvswitch instances running in the Nova compute nodes and which are automatically managed by Neutron.

The INCD network setup includes:

- Support for virtual networks based on GRE and VXLAN (can be created by the users)
- Datacenter VLAN mapping to provide access to internal datacenter services (these networks can only be created by the INCD cloud administrators)
- High performance access to VLAN based networks by mapping of network interface cards into the virtual machines using SR-IOV. This approach enables Neutron to bypass the virtual network switches (openvswitches) and achieve lower latency and higher bandwidth.

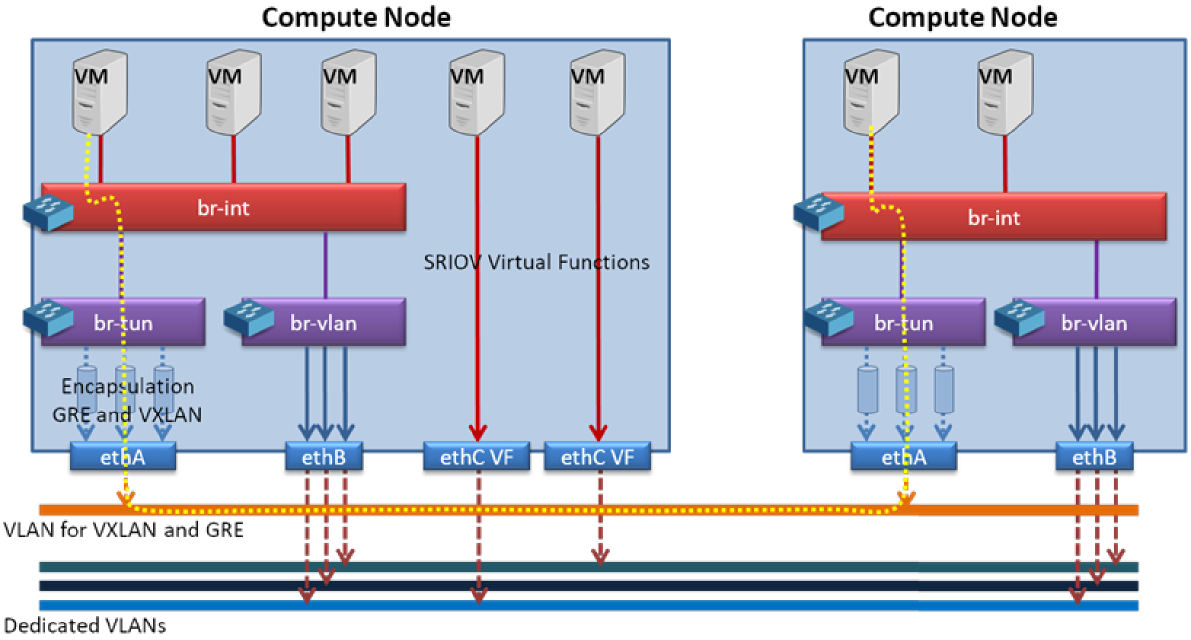


Figure 1 Openstack Neutron network setup for the INCD compute nodes

Cinder

Cinder provides access to remote block devices that can be mapped into virtual machines to provide storage. Nova can use Cinder to directly support the execution of virtual machine images on remote block storage or to add additional block storage to an existing virtual machine. The block storage provided by Cinder acts as virtual disks directly attached to the virtual machines. The Cinder storage component interfaces with the actual storage systems via drivers enabling a wide range of commercial and open source storage solutions. The INCD storage system backend used with Cinder is the Ceph storage system.

Glance

Implements the catalogue of operating system images and makes them available to Nova. Glance is used as the source of virtual machine images. Several virtual machine formats are supported. The images can be stored in several types of storage backends. The INCD storage backend for images is Ceph. The INCD setup allows the users to upload their own customized images.

Horizon

Horizon is the Openstack web dashboard through which users can intuitively perform many of the most frequent operations such as managing the life cycle of virtual machines, creating networks, associating IP addresses, allocating storage capacity, etc. The dashboard also provides a convenient interface to perform many infrastructure management operations such as creating users and tenants, defining resource usage quotas, and creating provider networks among many others.

The portal is made available through a Web server such as Apache httpd. The portal must be configured according to the installed components and services provided by the Openstack infrastructure.

Heat

The Openstack orchestrator called Heat allows the coordinated instantiation of a set of resources (networks, virtual machines, IP addresses, volumes, etc). Heat has its own template description language named HOT for describing the elements to be instantiated, but it also supports much of the functionality of Amazon's CloudFormation format. Heat also provides advanced features such as fault recovery and software configuration. INCD offers Heat as the mechanism to automate the deployment of complex services and applications. Translation of TOSCA specifications into the native HOT format is also supported for interoperability with other orchestration solutions.

Data storage

The INCD storage backend for Openstack is implemented through a Ceph storage system. Ceph is a state-of-the-art open source storage system that can provide block storage, object storage and file storage (file system). INCD is exploiting the block storage and object storage functionalities.

The Ceph setup includes a redundant management cluster and several storage nodes. The management cluster contains the information and rules necessary to directly access the data in the storage servers. The system is designed for high availability and resiliency with each data block always replicated across three different storage disks in different storage servers. A front-end network exposes the management cluster and the storage nodes to the Openstack nodes while a backend network is used to replicate data blocks across storage servers.

Image catalogue

The Glance nodes manage the virtual images stored on Ceph and can make the images available via its API. The catalogue of VM images to support the instantiation of VMs is made available through a combination of the Openstack Glance service together with Ceph block storage. This enables the VM images to be made available with high reliability and good performance. More than 100 VMs can be instantiated and contextualized in less than 10 minutes by a single user. Through Glance end-users can upload their own VM images.

Block Storage

The Ceph block storage is used by Cinder to provide block storage to virtual machines and for Glance to store virtual machine images. The Ceph front-end network is made directly accessible by the Nova compute nodes. When the virtual machines are stored on the Ceph backend, the INCD cloud allows live migration of virtual machines with sub-second switching of the machines between compute nodes. The INCD cloud architecture enables VMs to be stored either in the local storage space of the compute nodes or in Ceph backed storage, the latter is the recommended approach due to the excellent reliability exhibit by the Ceph block storage.

Object Storage

The object storage allows storage of unstructured data accessible via a REST API. This is a system analogous to Amazon's S3. It only has to levels of hierarchy, object container and object within the container. For greater flexibility each object can have metadata associated with it. The INCD object storage architecture is based on Ceph which provides SWIFT and S3 APIs via the rados gateway services. An initial setup of the object storage service based on CEPH was deployed and tested with good results. However given that the total storage capacity within CEPH has limited the delivery of this solution to end-users and was postponed until higher capacity storage hardware can be deployed, and actual demand for this type of service is observed.

Overall architecture

The Figure 2 provides a simplified diagram of the INCD cloud architecture including the Ceph storage system on top and the Openstack IaaS cloud below. The figure includes load balancing and high availability in the form of clusters both for Ceph monitor nodes and for Openstack controller nodes, HA proxies for the database access and APIs, multiple storage API nodes (separated from the Openstack controllers for performance) and redundant Openstack Neutron network routers. The https termination for the Openstack APIs is performed at the HA proxies with the exception of the storage APIs where it is performed in the API nodes themselves.

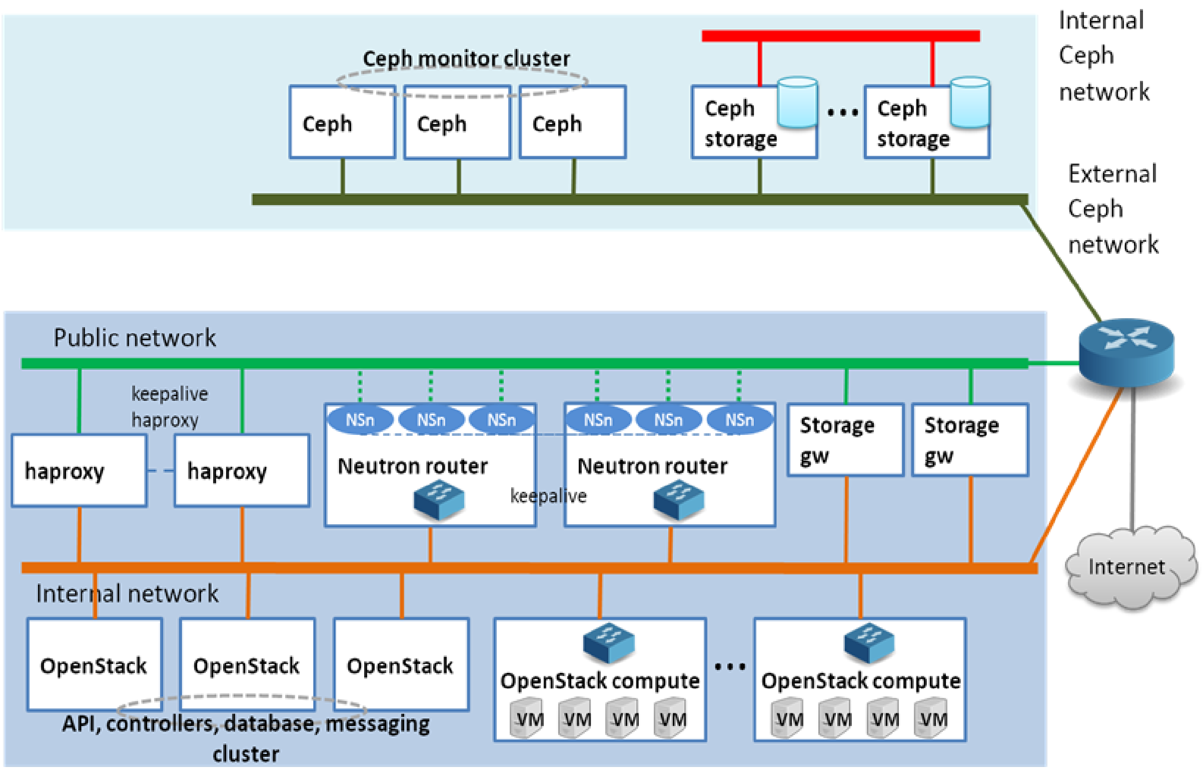


Figure 2 Overview of the 1st implementation of the INCD base cloud services

Advanced computing

The INCD cloud infrastructure is designed to support demanding computing and data access requirements. Three main types of scenarios have been considered: use of GPGPUs, bare-metal performance and network intensive applications.

GPGPUs

The architecture includes support for the allocation of GPGPUs to virtual machines using PCI pass-through. This setup allows each GPGPU PCI device to be made directly visible inside a virtual machine. Validation was performed using NVIDIA Kepler GPUs together with applications using CUDA and OpenCL libraries.

Bare-metal performance

Bare metal like performance can be achieved by combining Linux containers technology with Openstack. The INCD cloud architecture comprises the use of LXD as a bare metal hypervisor. Compared with other container technologies LXD aims to provide virtual machine like capabilities using Linux containers, LXD is thus suitable to run almost complete operating systems. For integration with Openstack the nova-LXD driver was deployed and tested. In addition the use of Docker in bare-metal was also evaluated using the nova-Docker driver however this driver is now in the process of being replaced by a new implementation. The benefit of using containers is reduced memory footprint and less overhead for compute and I/O operations.

Network intensive applications

The INCD cloud design includes measures to improve the performance of network intensive applications. These measures include:

- The support for network interface cards with SR-IOV capabilities. With SR-IOV physical network interface cards (PF) can show additional virtual instances of themselves in the PCI bus (VF) that can be mapped to virtual machines. This approach allows the virtual machines to access the network interface cards directly providing much higher performance.
- Tuning of the network packet size that has been enlarged using jumbo frames to reduce the packet processing overheads.
- Support for VLAN based virtual networks to exploit the native performance of the datacenter networks and access other datacenter services. This setup has been tested and used in multiple scenarios including high performance data transfer between virtual machines, execution of MPI parallel applications in combination with SR-IOV and extension

of the INCD High Throughput Computing farm to use virtual nodes deployed on top of the INCD cloud taking full advantage of the available capacity.

Federation

The INCD cloud has been federated and integrated with the EGI (European Grid Infrastructure) using the AAI checkin service. This proxy service can accept multiple identity providers including social networks, EDUGAIN, ELIXIR AAI, DARIAH AAI, IGTF X.509 certificates and ORCID. In terms of AAI INCD experimented federation using with both SAML and OIDC, with the final choice being OIDC due to its features and simpler integration. The EGI checkin service is thus used as a proxy that aggregates several identity providers. The INCD Openstack infrastructure sees the EGI checkin service as an OIDC identity provider.

Besides the authentication and authorization aspects, the INCD cloud was integrated with the EGI accounting service and with the EGI images catalogue. In this context the INCD cloud is ready for the EOSC (European Open Science Cloud) where it participates as an EGI federated cloud provider capable of supporting international user communities.

Resiliency and load balancing

The INCD cloud architecture includes resiliency and load balancing at several levels.

1. The SQL databases supporting the Openstack components are provided by a cluster of three machines running MariaDB.
2. The messaging system used by Openstack was setup using a redundant RabbitMQ.
3. The API and controller nodes for the several Openstack services are split across a cluster of three machines.
4. Load balancing and high availability both for MariaDB and for the Openstack API and controller nodes is implemented by a fault-tolerant *haproxy* setup that includes VRRP provided by *keepalived*.
5. Multiple Openstack compute nodes are available therefore virtual machines whose images are stored in Ceph can be live migrated between the nodes without noticeable impact. This facilitates both redistribution on VMs across compute nodes enabling better load-balancing and also fast recovery of virtual machines in case of compute node failure.
6. Neutron network gateways are setup in a fault tolerant setup enabling redundant routing.

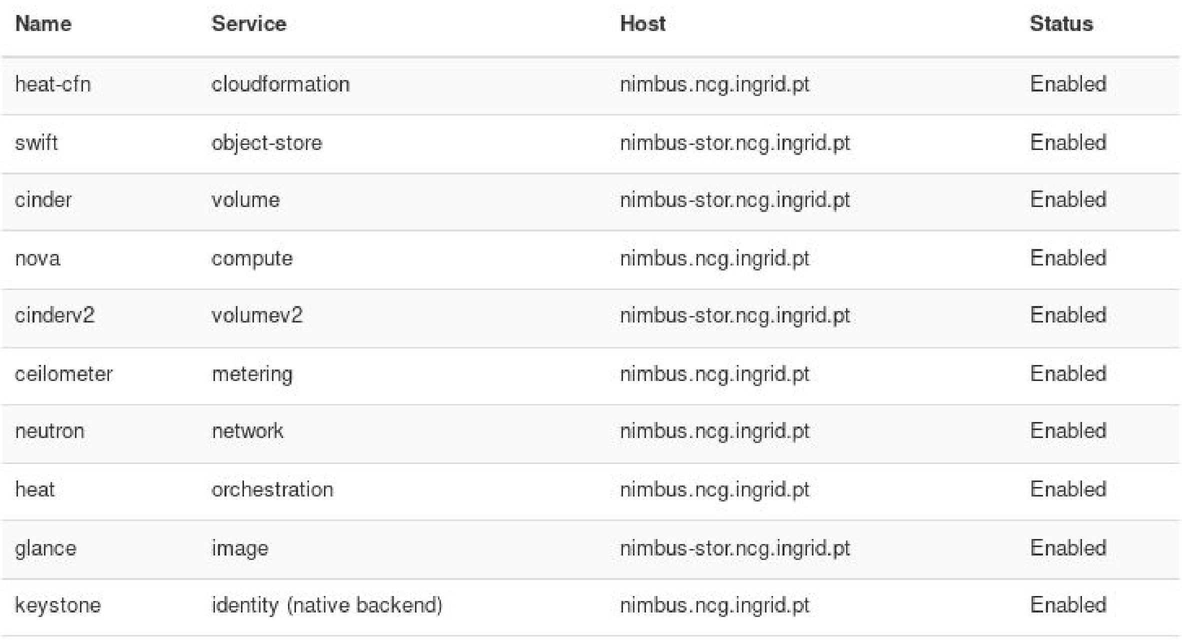
7. Storage gateways have been deployed in a scalable redundant setup enabling access to images and object storage.
8. The Ceph storage system has each data block replicated across three failure domains.
9. Resiliency and load balancing

10. The INCD cloud architect

11. The Ceph monitor nodes are setup in a Paxos cluster composed of three nodes.

INCD Openstack Services

The Figure 3 shows the INCD Openstack services of the first INCD IaaS cloud implementation.

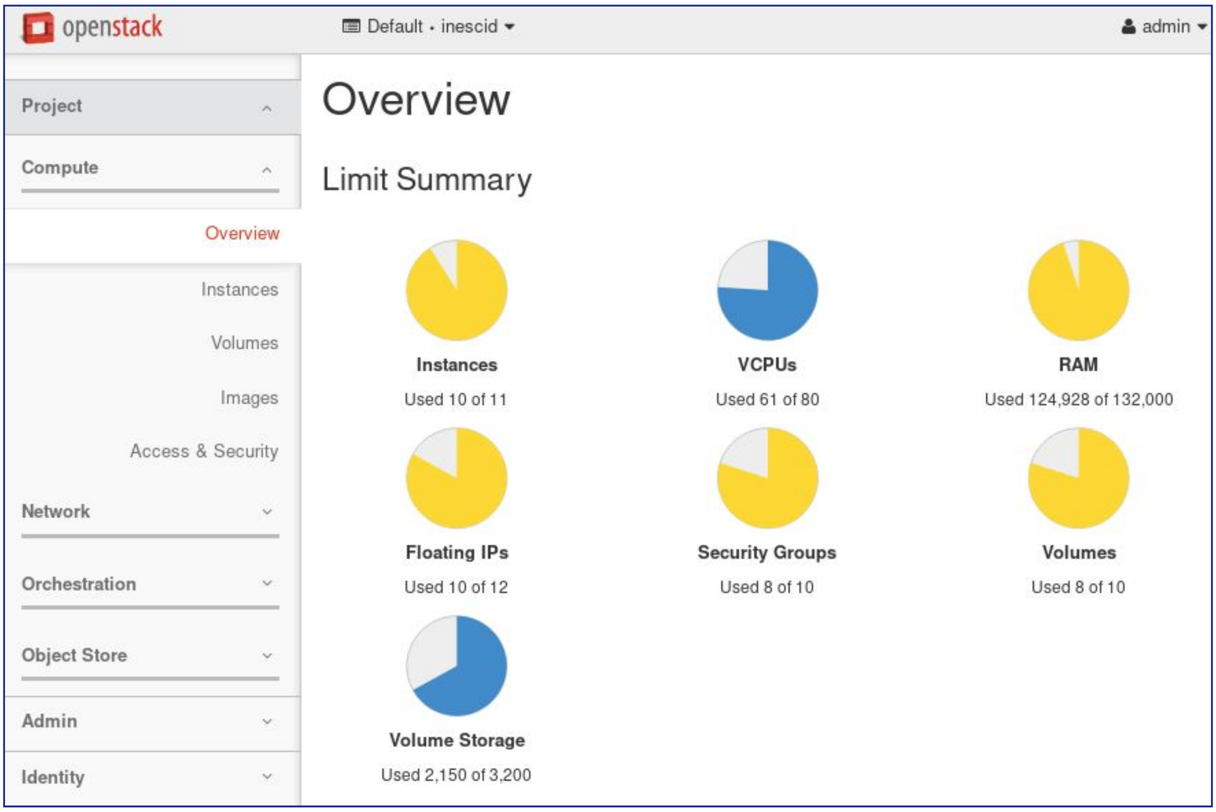


Name	Service	Host	Status
heat-cfn	cloudformation	nimbus.ncg.ingrid.pt	Enabled
swift	object-store	nimbus-stor.ncg.ingrid.pt	Enabled
cinder	volume	nimbus-stor.ncg.ingrid.pt	Enabled
nova	compute	nimbus.ncg.ingrid.pt	Enabled
cinderv2	volumev2	nimbus-stor.ncg.ingrid.pt	Enabled
ceilometer	metering	nimbus.ncg.ingrid.pt	Enabled
neutron	network	nimbus.ncg.ingrid.pt	Enabled
heat	orchestration	nimbus.ncg.ingrid.pt	Enabled
glance	image	nimbus-stor.ncg.ingrid.pt	Enabled
keystone	identity (native backend)	nimbus.ncg.ingrid.pt	Enabled

Figure 3 INCD Openstack services

INCD Openstack Dashboard

The Figure 4 shows a panel of the Openstack Horizon web dashboard, available at: https://nimbus.ncg.ingrid.pt



Overview

Limit Summary

Instances
Used 10 of 11

VCPUs
Used 61 of 80

RAM
Used 124,928 of 132,000

Floating IPs
Used 10 of 12

Security Groups
Used 8 of 10

Volumes
Used 8 of 10

Volume Storage
Used 2,150 of 3,200

Figure 4 Compute overview panel of an Openstack tenant