Journal of large-scale research facilities

Journal of large-scale research facilities, 5, A137 (2019) http://dx.doi.org/10.17815/jlsrf-5-173

Published: 25.10.2019

# HDF Cloud – Helmholtz Data Federation Cloud Resources at the Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre *

Instrument Scientists:
- Federated Systems and Data, Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, phone: +49(0)2461 61 2828, email: ds-support@fz-juelich.de
- Supercomputing Support, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 2828, email: sc@fz-juelich.de

**Abstract:** The HDF Cloud is an OpenStack based infrastructure-as-a-service (IaaS) environment operated by Jülich Supercomputing Centre (JSC) at Forschungszentrum Jülich. It has been installed predominantly to support challenging data use cases within the Helmholtz Association's strategic initiative Helmholtz Data Federation (HDF). To this end, it has been connected to one of the central storage resources of JSC, the DATA file system that is also available on the high-performance computing systems.

## 1 Introduction

The HDF cloud infrastructure is a virtual machine (VM) hosting infrastructure based on OpenStack (OpenStack Consortium, 2019c). It allows provisioning and management of user-controlled VMs with the Linux operating system. The terms and conditions for services provided by VMs on the cloud infrastructure are regulated by an acceptable usage policy.
The main services provided by the infrastructure comprise VMs, block storage, networking, and orchestration. Details on their availability will be given in Section 5.
Use cases defined by the Helmholtz Data Federation (Helmholtz Association, 2019c) initiative of the Helmholtz Association of German Research Centres (Helmholtz Association, 2019b) as well as selected other use cases are eligible to use the resource.

* **Cite article as:** Jülich Supercomputing Centre. (2019). HDF Cloud – Helmholtz Data Federation Cloud Resources at the Jülich Supercomputing Centre. *Journal of large-scale research facilities, 5*, A137. http://dx.doi.org/10.17815/jlsrf-5-173

## 2 Hardware

The general base for all nodes of the installation are Fujitsu PRIMERGY RX2530M4 servers. Three of them are dedicated as management servers and equipped with two Intel Xeon Silver 4114 10-core processors and 196 GB of main memory (RAM) each.
Another dedicated server has the role of the network node and is equipped with two Intel Xeon Gold 5118 12-core processors and 196 GB of RAM.
There are 16 compute nodes for hosting VMs, each equipped with 384 GB of RAM and two Intel Xeon Gold 6126 12-core processors totalling about 6 TB RAM and 384 cores.
Each of the nodes is equipped with several network interfaces, connecting to the surrounding infrastructure in various ways. First, there is a two-port 40 Gb/s interface. Only one of the two ports of this interface is connected. Secondly, there is a four-port 10 Gb/s interface. Last, there are three on-board 1 Gb/s devices, that are connected depending on the node type. The details of how these interfaces are embedded in the infrastructure will follow in Section 3.

## 3 Network

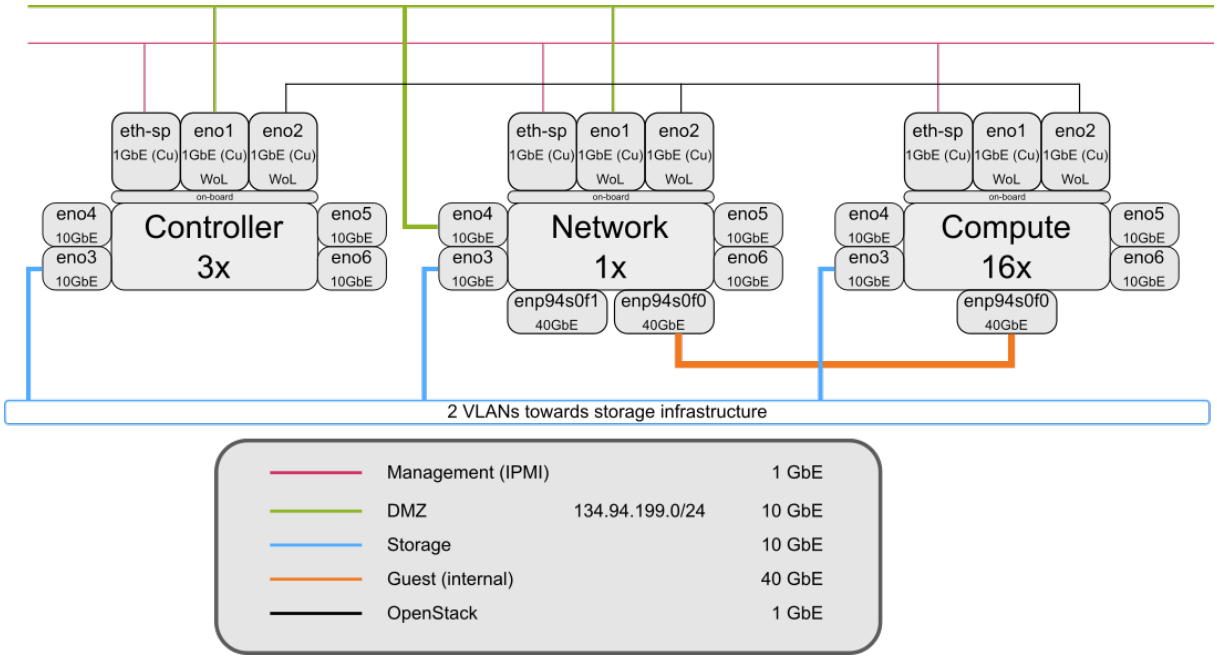


Figure 1: Network connections of the HDF Cloud system at JSC.

The 40 Gb/s Ethernet links are used for communication among virtual guests. Guest networks on these interfaces are separated using VXLAN encapsulation. The storage that supports the virtual block devices is connected with 10 Gb/s Ethernet. As can be seen in Figure 1, storage access makes use of two VLANs. Therefore, all connections to the storage systems, either for the underlying shared file system or to the XCST storage system that provides the DATA file system (cf. Section 4), share the same physical link. This allows for a bandwidth of 10 Gb/s per node with an aggregated bandwidth of 80 Gb/s configured in the infrastructure. Communication of virtual guests with the outside world is routed through the network node, which has a 10 Gb/s interface reserved for this purpose. Floating IPs for the virtual guests are within the range 134.94.199.111–134.94.199.253.

## 4 Surrounding Infrastructure

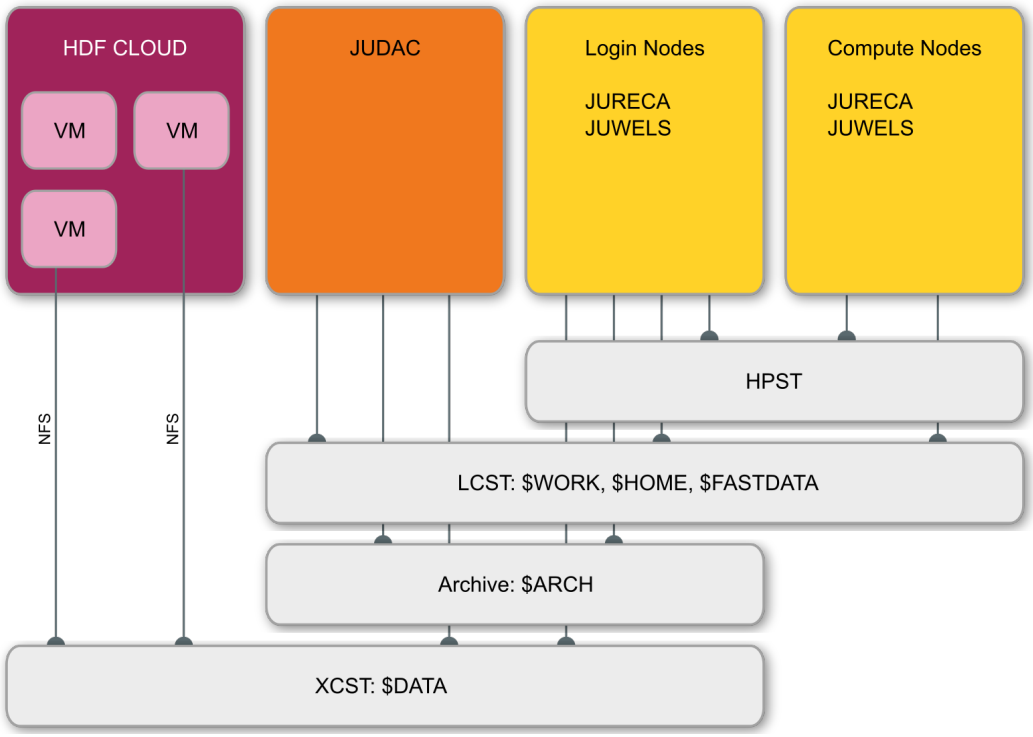


Figure 2: Access to storage tiers at Jülich Supercomputing Centre

The DATA storage service facilitates data sharing and exchange across compute projects within the JSC supercomputing facility. In order to be able to participate in this data sharing, an active user account is required. In order to enable data sharing with external users, e.g. via a web-service, additional infrastructure is required, which is provided by HDF Cloud.
Portions on the DATA file system can be exported via NFS into VMs hosted on the cloud infrastructure. This enables the creation of community-specific, data-oriented services such as large, web- accessible databases. Since the service-providing VMs are community-managed they are isolated from the user management of the supercomputing facility. For this reason, all NFS exports are protected with a "UID mapping" that alters the visible data ownership to a single user for read and write access. In particular, fine grained access control capabilities through the file system layer itself are limited and need to be implemented on the service layer if required.

### 4.1 Access to the DATA file system

The possibility to access portions of the DATA file system (Jülich Supercomputing Centre, 2019b) from the VMs within HDF Cloud facilitates the user's ability to use the same data in both the HPC and Cloud environments. This is the first time at JSC that additional services can be offered on e. g. simulation results.
Access is granted and setup on a case by case basis and involves administrator action on the storage and Cloud side, as well as the VM administrator. Technically, access is realized through an NFS export of the DATA file system through the Cluster Export Services (CES) of IBM Spectrum Scale. For security reasons, all access is mapped to a single user and group ID on the server side as negotiated with the project. This measure is important, because the user management within VMs and on the HPC systems is typically disparate.

The CES nodes share a network with the OpenStack installation (cf. Figure 1). On the OpenStack side, this network is only available to administrators, who create virtual ports and assign them to the user projects individually. In turn, these ports can be attached to any VM within the user project, allowing the consumption of NFS shares as shown in Figure 2.
A prerequisite for the access to the DATA file system is the existence of a data project that has sufficient quota in that file system. Users must apply for data projects separately.

## 5 Software and Services

At the time of this writing, the OpenStack Queens release (OpenStack Consortium, 2019b) is installed on the resources. The deployment method is "kolla-ansible" (OpenStack Consortium, 2019a), allowing for a highly automatized deployment and good customizability. All services are deployed in docker containers. Being able to customize container images is important for the integration of the underlying shared file system, IBM Spectrum Scale (IBM, 2019), and exploiting its enhanced features in the Cinder volume service.
The software stack of the OpenStack installation comprises the following services.

### 5.1 Identity Service – Keystone

Keystone is the identity and catalogue service of OpenStack. It provides adaptors for integrating a variety of backend authentication systems. Within the HDF Cloud environment, users can authenticate using

- JSC's LDAP accounts,
- HDF AAI through OIDC,
- EUDAT AAI through EUDAT B2ACCESS.

By default, users do not have any resources. For that purpose, authentication must be complemented by authorization. Currently, a manual scheme is employed in which administrators manually add users to their respective projects. An automated scheme will be taken into consideration in the future, particularly for the integration with HDF AAI.

### 5.2 Compute service – Nova

The Nova compute service is available on 16 compute nodes, the API and other management services of the Nova suite of services are available on the three management nodes.
Currently, we employ the default OpenStack overprovisioning factors of 16 for the number of cores and 1.5 for the amount of RAM. These figures will be adjusted over time as we gain experience with the typical workloads in the environment. Whereas a dedicated allocation of cores would technically be possible, we do not currently support this. Again, a final decision about this will be taken in case there is any demand.
Resources are allocated in OpenStack as flavours. A flavour comprises the

- number of VCPUs,
- amount of RAM,
- root disk size,
- ephemeral disk size,
- swap disk size, and
- RX/TX factor.

All parameters except VCPUs and RAM are the same for all flavors in our deployment.
As shown in Table 1, we have five categories–*t*iny, *s*mall, *m*edium, *l*arge, e*x*tra *l*arge–describing the ratio of VCPUs and RAM that ranges from 2:1 to 1:8. The remaining parameters have been fixed to the values shown in Table 2.
The size of the root disk is low as compared to other deployments, the settings for the additional storages *ephemeral* and *swap disk* are zero. All users are advised to use the volume service (cf. Section 5.5) for

| RAM \ VCPUs | 1 | 2 | 4 | 8 | 16 |
|---|---|---|---|---|---|
| **.5 GB** | t1 | - | - | - | - |
| **1 GB** | s1 | t2 | - | - | - |
| **2 GB** | m1 | s2 | t4 | - | - |
| **4 GB** | l1 | m2 | s4 | t8 | - |
| **8 GB** | xl1 | l2 | l4 | s8 | t16 |
| **16 GB** | - | xl2 | m4 | m8 | s16 |
| **32 GB** | - | - | xl4 | l8 | m16 |
| **64 GB** | - | - | - | xl8 | l16 |
| **128 GB** | - | - | - | - | xl16 |

Table 1: OpenStack flavours

| Parameter | Value |
|---|---|
| root disk | 10 GB |
| ephemeral disk | 0 GB |
| swap disk | 0 MB |
| RX/TX factor | 1.0 |

Table 2: Fixed parameters for defined flavours

additional storage, in particular for data underlying the service. Also, given the availability of data projects at JSC, VM administrators can request their VM to be connected to a corresponding data project in the DATA file system (cf. Section 4.1).

### 5.3 Network service – Neutron

The Neutron network service comprises a number of services that are deployed on various systems. First of all, the Neutron server that provides the API is available on all three management nodes for high availability. The L3 and DHCP agents are available on the network node. These services host the virtual routers and provide instances with IP addresses. In order to connect these latter two services to internal networks and to provide connectivity for virtual routers to the DMZ and thus the outside world, the network host also hosts an instance of the Open vSwitch (OVS) agent. Finally, the OVS agent is also deployed on all compute nodes to allow for integrating virtual networks with the VMs as well as connecting VMs with the storage network towards the Spectrum Scale Cluster Export Services (CES) as described in Section 4.1.

### 5.4 Image service – Glance

The Glance image service is available on the three management nodes. The backend is file system based, files are stored on a dedicated IBM Spectrum Scale file system, allowing for fast availability of the image at instantiation time.

### 5.5 Volume service – Cinder

The Cinder volume service is located on the three management nodes. The backend is using the GPFS driver. A total volume of several hundred Terabytes is available to this service.

### 5.6 Client software

Two mechanisms to access the cloud resources are available to end users: the OpenStack Horizon Dashboard[1] and the OpenStack APIs. When using the Horizon Dashboard, a web-based interface, users can authenticate either through their JSC LDAP accounts managed by JuDoor (Jülich Supercomputing Centre, 2019a), or through federated accounts via the EUDAT or HDF AAI federations (EUDAT CDI, 2019; Helmholtz Association, 2019a). Federated access is currently not possible when using the OpenStack APIs. The APIs can only be accessed with an account in JuDoor.

[1] OpenStack Horizon Dashboard: https://hdf-cloud.fz-juelich.de/

## 6 Cloud Images

A number of default images are deployed and managed by administrators of the HDF Cloud installation. It will always be ensured that up-to-date versions of these images are available to all users of the infrastructure. Similarly, outdated images will be deactivated and ultimately deleted after a reasonable grace period. This allows for using a recently withdrawn image in case of any problems with the most up-to-date one. An image is only deleted, if no active VM depends on it. The default operating systems currently available are:

- Ubuntu Xenial 16.04 LTS
- Ubuntu Bionic 18.04 LTS
- Debian Stretch
- Debian Buster
- Centos 7

The images are downloaded from their respective official repositories (Canonical Ltd., 2019; Debian Cloud Team, 2019; The CentOS Project, 2019) and provided unmodified. Users can use their own images. In the future, we plan to provide newer releases of the above mentioned distributions as appropriate. The metadata associated with the images is optimized for running in the given cloud environment.

## 7 Access

End users can gain access to the system and apply for resources by sending an email to ds-support@fz-juelich.de. The email should contain a problem statement, a justification of the requested resources, and, if applicable, a sketch of the envisioned architecture. The latter part is particularly important, if more than just the HDF Cloud resources are involved in a certain scenario.
The primary target audience for the service are the use cases defined in the Helmholtz Data Federation strategic initiative (Helmholtz Association, 2019c). Following this are other projects or communities funded by the Helmholtz Association (Helmholtz Association, 2019b). In the long run, it is envisioned that a resource allocation mechanism will be established that is comparable to the applications for computing time at JSC.

## References

Canonical Ltd. (2019). *Ubuntu Cloud Images.* https://cloud-images.ubuntu.com/. (Last accessed: 2019-09-03)

Debian Cloud Team. (2019). *Debian Cloud Images.* https://cdimage.debian.org/cdimage/openstack/. (Last accessed: 2019-09-03)

EUDAT CDI. (2019). *B2ACCESS.* https://b2access.eudat.eu/. (Last accessed: 2019-09-03)

Helmholtz Association. (2019a). *HDF AAI.* https://login.helmholtz-data-federation.de/. (Last accessed: 2019-09-03)

Helmholtz Association. (2019b). *Helmholtz Association of German Research Centres.* https://www.helmholtz.de/en/. (Last accessed: 2019-09-03)

Helmholtz Association. (2019c). *Helmholtz Data Federation.* https://www.helmholtz.de/en/research/information_data_science/helmholtz_data_federation/. (Last accessed: 2019-09-03)

IBM. (2019). *IBM Spectrum Scale product webpage.* https://www.ibm.com/marketplace/scale-out-file-and-object-storage. (Last accessed: 2019-09-03)

Jülich Supercomputing Centre. (2019a). *JuDoor – Portal for managing accounts, projects and resources at JSC.* https://judoor.fz-juelich.de. (Last accessed: 2019-09-03)

Jülich Supercomputing Centre. (2019b). JUST: Large-Scale Multi-Tier Storage Infrastructure at the Jülich Supercomputing Centre. *Journal of large-scale research facilities JLSRF*, *5*(A136). http://dx.doi.org/10.17815/jlsrf-5-172

OpenStack Consortium. (2019a). *Kolla-Ansible.* https://docs.openstack.org/kolla-ansible/latest/. (Last accessed: 2019-09-03)

OpenStack Consortium. (2019b). *OpenStack Releases.* https://releases.openstack.org/. (Last accessed: 2019-09-03)

OpenStack Consortium. (2019c). *OpenStack web site – Build the future of Open Infrastructure.* https://www.openstack.org/. (Last accessed: 2019-09-03)

The CentOS Project. (2019). *CentOS Cloud Images.* https://cloud.centos.org/centos/7/images/. (Last accessed: 2019-09-03)