TASK CLOUD INFRASTRUCTURE IN THE CENTRE OF INFORMATICS – TRICITY ACADEMIC SUPERCOMPUTER & NETWORK

PIOTR ORZECHOWSKI

Centre of Informatics – Tricity Academic Supercomputer & Network
Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdansk, Poland

(received: 25 July 2018; revised: 24 August 2018;
accepted: 21 September 2018; published online: 5 October 2018)

Abstract: The cloud solution called the TASK cloud is described. Its main components and the way of their implementation are described. Examples of deployed environments based on the cloud are also presented. Moreover, the idea of combining the cloud and big data platforms is suggested and discussed.

Keywords: cloud computing, IaaS cloud, resources, big data, Internet of Things, IoT
DOI: https://doi.org/10.17466/tq2018/22.4/d

1. Introduction

Nowadays data is everywhere and there are petabytes of data produced by different kinds of devices. Some unique and common devices can be distinguished, depending on popularity. Some good examples are the Large Hadron Collider which is one in the world and Internet of Things sensors of which there are billions. At the first look they are very different but there is one significant feature that have to be marked: they produce data, data which is very important in science.

Presently the base element for every research is data. It is needed for example to verify the rules describing models but also to organize the learning process in an artificial intelligence. The scope in which data is used is very wide and user cases are changing all the time. Depending on the kind of research to be conducted scientists have different requirements about the resources to be provided. Another case is resource sharing between different groups of scientists. Different groups need different software which could not be provided as a whole cluster, *e.g.* because of proprietary reasons. Moreover, many projects require

resources for 24 hours a day, for example, services to gather data or provide services for end users, therefore, resources available in HPC queues are not sufficient. Such cases of application were one of main issues in a project of the Centre of Competence for Novel Infrastructure of Workable Applications (NIWA) [1].

In order to meet the expectations of users, CI TASK decided to support an Infrastructure as a Service model as a resource providing method. This solution lets administrators provide a highly reliable and efficient infrastructure which can be easily manageable between projects. The CI TASK team carried out research to determine which software would be the best choice for clients assuming that it would not be a proprietary solution to avoid vendor lock-in. After a case study and market analysis the decision was to deploy OpenStack [2] software providing the IaaS model [3] of cloud computing and to use Ceph [4] as the storage provider. At the present time, they are one of most popular open source software to provide cloud services and reliable storage.

This article presents the assumed TASK cloud infrastructure. Section two demonstrates the Ceph storage and the OpenStack software deployment is presented in section three. In the fourth section cases of application in production are shown to illustrate the cloud utilities. The last section describes future works related to the development of the existing cloud infrastructure.

2. Ceph storage

The Ceph software is a solution to provide high performance, reliable and scalable distributed storage [4]. At the bottom Ceph uses commodity hardware (disk servers) instead of expensive matrices what makes it cheaper in maintenance. The data stored in Ceph is safe because it is protected by either of the two mechanisms: replication or erasure coding. A proper mechanism should be chosen depending on the application cases and the storage types used. Notwithstanding the foregoing, this Ceph has fault tolerance and it is resistant to disk failures or even a whole server failure.

The Ceph storage has been deployed as the common storage for different kinds of environments. It can be used as storage for hosts, virtual machines but also directly by applications. A general conceptual view of deployment is presented in Figure 1.

As has been mentioned before there are three types of data storage provided by Ceph. The main one is object storage which is at the bottom of others. Everything in Ceph is an object. The object storage concept is a proposition which solves problems with traditional filesystems, such as weak scalability and limits on capacity. Object storage has no limits on size and can be easily scaled up by adding new servers to the Ceph installation. Ceph provides implementation for two APIs: S3 (Amazon standard) and Swift (OpenStack standard). These APIs work as REST services over HTTP. There are many ready-to-use libraries in different programming languages and many ready-to-use software systems to store data

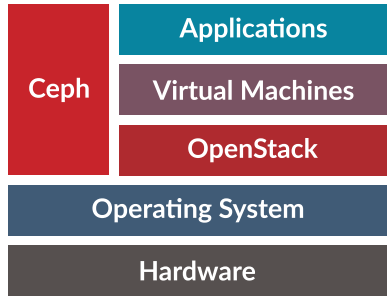

Figure 1. Conceptual view of TASK cloud deployment

in the object storage. An example of application cases for object storage could be storing data for big data analysis or archive/backup data using third party software supporting the object backend.

The second type of provided storage is block storage. It is more similar to traditional volumes shared by LUNs or iSCSI targets. It can be used as storage for hosts or virtual machines.

Last but not least is CephFS – the storage type provided by Ceph. It is a distributed filesystem which can be provided as a shared volume for different clients. It is also built on top of the object storage.

The actual capacity of Ceph deployed in the TASK cloud is about 0.5 PB and it is attached as based storage for OpenStack.

3. OpenStack infrastructure

The TASK cloud has been build using the OpenStack software [2] which implements the Infrastructure as a Service model of providing resources. It supports full isolation of resources such as network, storage and compute assets. Cloud clients can fully manage all the assigned resources. They can create a virtual network infrastructure, configure machines with administrative privileges and attach storage to them. Clients can also use object storage and organize objects using containers. A general concept of the provided resources is presented in Figure 2.

The TASK cloud infrastructure is a full implementation of the cloud concept. Environments deployed in clouds should be designed for tolerance instance faults, the ability to be scaled-up and down and assume high availability of solutions [3]. More specifically, the cloud idea assumes that it is only data stored in persistent storage that is protected. Instance failures are not maintained by the cloud as is the case with virtualization platforms such as VMWare where, for example, virtual machines are migrated when the host fails. Cloud platforms assume that the deployed environments are monitored and in case of failures the automated deployment method is used for the provision of failed instances.

The TASK cloud infrastructure provides over 4300 virtual CPUs, 11 PB of RAM and 1 PB of ephemeral disk storage (SSD and HDD). There are three public operating system images available to every client: Ubuntu Bionic, CoreOS and

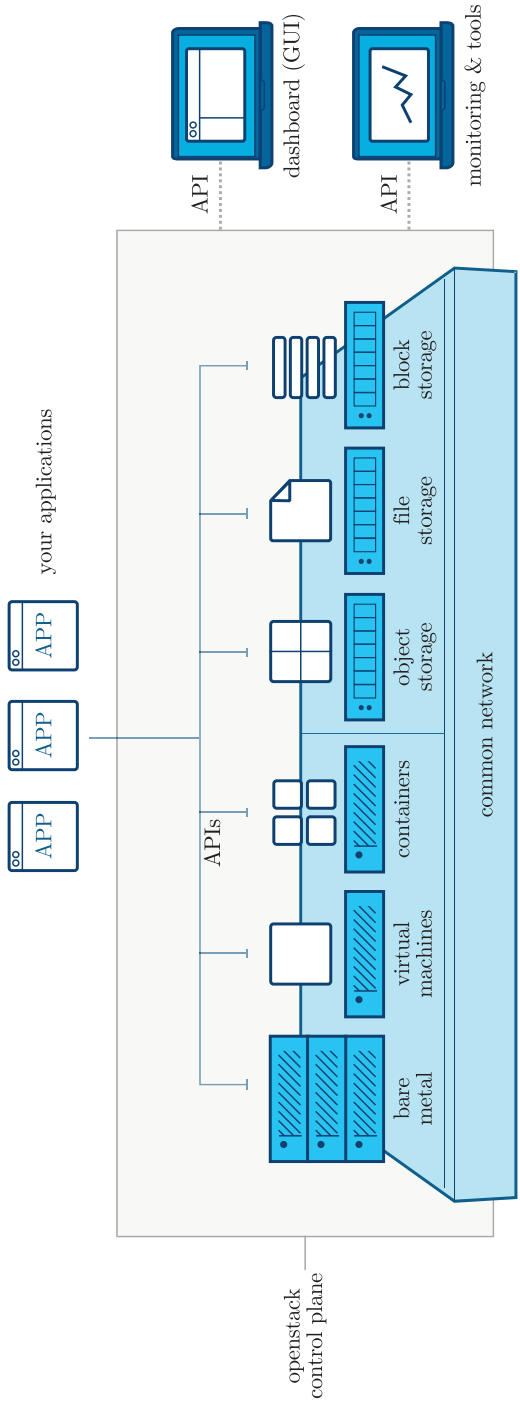

Figure 2. Resources provided by OpenStack [5]

CentOS 7. All images are maintained automatically and they are upgraded once a week with all the current security updates. Users are also able to provide their custom images such as Windows or other operating systems. It is only required to have the appropriate license, if needed.

The cloud infrastructure has been deployed on nearly 160 HP ProLiant and Huwei RH servers interconnected by 10 Gigabit Ethernet and Infiniband FDR networks. The deployment procedure was fully automated using special software to install and configure servers (implemented by CI TASK developers) and OpenStack Kolla to deploy OpenStack. All the software was installed based on the Ansible automation solution.

4. Deployed cloud environments

The TASK cloud was deployed over 2 years ago and currently it has several production environments deployed. In this section some of these environments will be introduced to illustrate the cloud utilities.

One of deployed environments is the Pomeranian Digital Library [6]. The digital resources gathered there are the result of cooperation between 19 associations in the Pomeranian District in Poland. At the present time there are over 66 thousand publications in the library. The average number of visitors is about 366 thousand, 3000 thousand of views and 900 new publications per month.

Another application deployed on the TASK cloud infrastructure is the SowiDocs system. It is a Plagiarism Detection Platform used by the Gdansk University of Technology. The platform uses different kinds of algorithms to detect plagiarism in student master theses or conference proceedings [7]. The deployment diagram is shown in Figure 3.

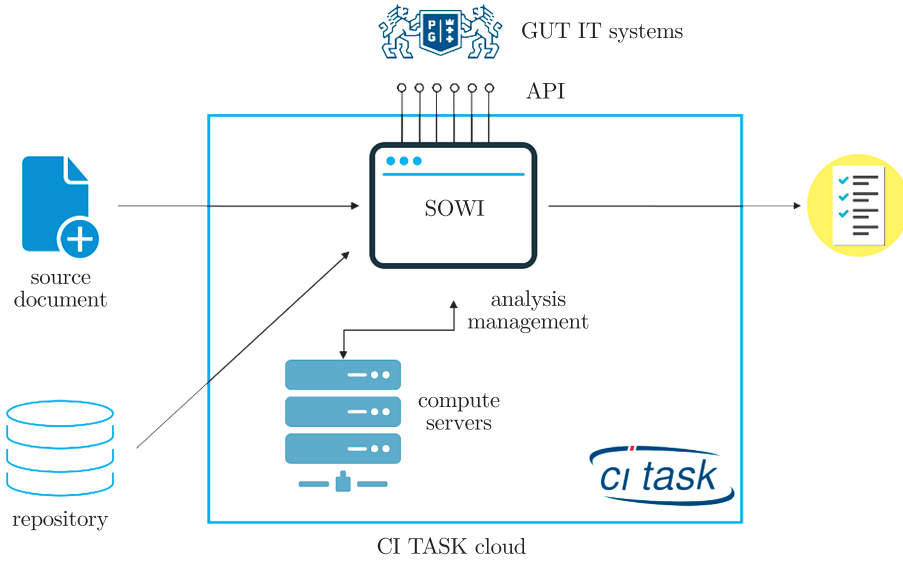

Figure 3. SowiDocs deployment model

Another example is the Big Data analysis platform [8]. It consists of a service to gather data from the Sentinel-2 ESA mission and an Apache Spark platform to process this data with custom algorithms (see Figure 4). Currently, there is about 120 TB of data gathered in the TASK cloud object store which can be used in Big Data analysis.

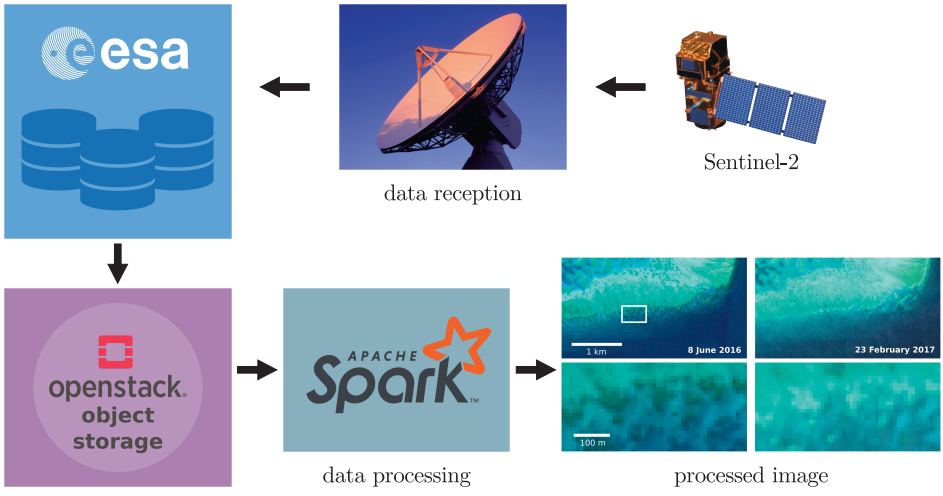

Figure 4. Sentinel data processing

A specific workload deployed in the TASK cloud is the MeteoPG platform. It integrates HPC with the cloud resources. The forecasting model is a process on the HPC cluster and all results of analysis are sent to the cloud environment (see Figure 5). Then, they are processed and new graphical layers for each parameter (wind, temperature, humidity) are created. There is also a MeteoPG web application deployed in a cloud which presents the predicted forecast. It is one of the fully operational models which shows possible cooperation between the HPC cluster and the cloud.

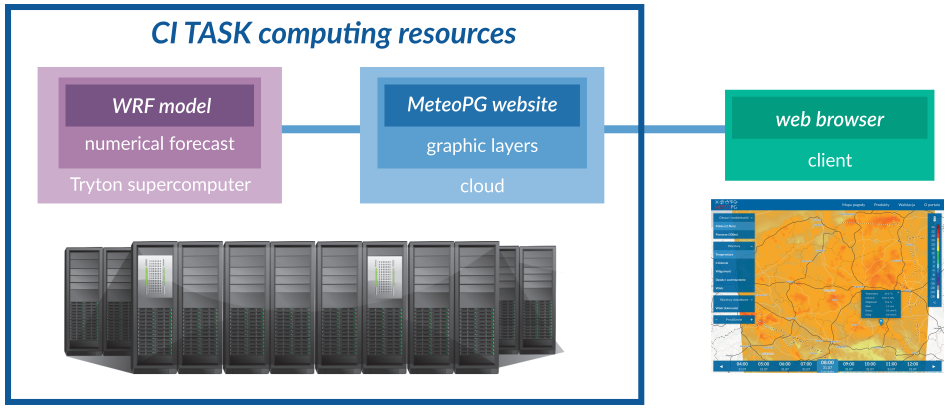

Figure 5. MeteoPG components architecture

5. Future works

The presented TASK cloud platform is a first step of a solution to provide more efficient, flexible resources for clients. As far as further works are concerned, there is a plan to extend the portfolio of OpenStack components available to our users: provisioning Bare Metal nodes and Containers, providing DNS as a Service, Kubernetes as a Service or Applications as a Service model. Another idea is to integrate the TASK cloud with LoRaWAN and deploy a dedicated infrastructure to gather information and process it [9]. A subsequent concept is deeper consolidation with the HPC cluster to let users share resources between environments. The currently described cloud environment is first of all available for the scientific purposes of our clients. However, the plan of CI TASK is to bring the cloud to the market and provide resources also for business applications. If you are interested in resources of such kind, please feel free to contact us at cloud@task.gda.pl.

References

[1] Krawczyk H 2015 *TASK Quarterly* **19** (4) 357 doi: 10.17466/tq2015/19.4/b
[2] OpenStack 2018 [online] https://www.openstack.org/ [accessed 2-December-2018]
[3] Serrano N, Gallardo G and Hernantes J 2015 *IEEE Software* **32** (2) 30 doi: 10.1109/MS.2015.43
[4] Ceph 2018 [online] https://ceph.com/ [accessed 2-December-2018]
[5] OpenStack Components 2018 [online] https://www.openstack.org/software/ [accessed 2-December-2018]
[6] Pomeranian Digital Library [online] http://pbc.gda.pl [accessed 2-December-2018]
[7] Sobecki A and Kepa M 2018 *Semantic Keyword-Based Search on Structured Data Sources* 56 doi: 10.1007/978-3-319-74497-1_6
[8] Proficz J and Drypczewski K 2017 *TASK Quarterly* **21** (4) 365 doi: https://doi.org/10.17466/tq2017/21.4/y
[9] Wiszniewski Ł and Klimowicz D 2017 *TASK Quarterly* **21** (4) 355 doi: https://doi.org/10.17466/tq2017/21.4/q