EPJ Web of Conferences **245**, 07013 (2020) https://doi.org/10.1051/epjconf/202024507013
CHEP 2019

Evolution of the CloudVeneto.it private cloud to support research and innovation

Paolo Andreetto[1], *Fulvia* Costa[1], *Alberto* Crescente[1], *Sergio* Fantinel[2], *Federica* Fanzago[1], *Paolo Emilio* Mazzon[3], *Matteo* Menguzzato[4], *Gianpietro* Sella[5], *Massimo* Sgaravatto[1], *Sergio* Traldi[1], *Marco* Verlato[1,*], *Marco* Zanetti[4], and *Lisa* Zangrando[1]

[1]INFN, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy
[2]INFN, Laboratori Nazionali di Legnaro, Viale dell'Università 2, 35020 Legnaro (Padova), Italy
[3]Padova Neuroscience Center, Università di Padova, Via Orus 2/B, 35131 Padova, Italy
[4]Dipartimento di Fisica e Astronomia 'Galileo Galilei", Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[5]Dipartimento di Scienze Chimiche, Università di Padova, Via Marzolo 1, 35131 Padova, Italy

Abstract. CloudVeneto.it was initially funded and deployed by INFN in 2014 for serving the computational and storage demands of INFN research projects mainly related to HEP and Nuclear Physics. It is an OpenStack-based scientific cloud with resources spread across two different sites connected with a high speed optical link: INFN Padova Unit and the INFN Legnaro National Laboratories. The infrastructure has grown throughout the years with additional funds from ten University of Padova departments, and nowadays supports a broader range of scientific and engineering disciplines. Its hardware resources provide around 2500 computational cores and 360 TB of storage to about 250 users working for more than 70 projects. In the last months we enhanced the cloud platform in two ways: 1) by integrating a number of heterogeneous GPU cards to address the special needs of user communities whose computations involve machine learning training; 2) by enabling the users to simply deploy on-demand Kubernetes clusters for Big Data Analytics applications taking advantage of the operator framework. In particular, the Kubernetes operators for Apache Kafka and Spark platforms were integrated to address real-time data ingestion and streaming processing on the cloud. This article describes the technical details of these two solutions and their integration with the cloud infrastructure.

1 Introduction

The origin and the details of the CloudVeneto.it infrastructure have been described in a previous article [1]. In sections 2 and 3 we'll therefore only give an updated summary of its layout and capacity, while in sections 4 and 5 we'll focus on two use-cases, from Astro-particle physics and HEP, that could profit of the most recent enhancements brought to the infrastructure, namely the integration of a number of heterogeneous GPU cards and the development of a mechanism to deploy on-demand Big Data Analytics clusters based on the

* Corresponding author: Marco.Verlato@pd.infn.it

Kubernetes container orchestration framework. Section 6 will track the conclusions and future perspectives.

2 The CloudVeneto.it infrastructure

CloudVeneto.it is an OpenStack based IaaS that has been funded by INFN and ten departments of the University of Padova, and is serving the scientific user communities affiliated to them. Nowadays more than 250 users and 70 research projects are supported by this infrastructure. Its hypervisors and storage nodes are geographically spread across two sites 10 km apart (INFN Padova data center and INFN Legnaro National Laboratories) which have a dedicated network connection at 10 Gbps. The main OpenStack services (Horizon, Keystone, Neutron, Glance, Nova, Cinder, Heat, EC2 API) are hosted in two controller nodes implementing high availability in an active-active configuration. The overall fault-tolerance and high availability is achieved using three instance redundancy for a Percona XtraDB cluster, a RabbitMQ cluster and a HAProxy/Keepalived cluster. The disk devices of the hypervisors provide the ephemeral storage of the virtual machines, while two iSCSI storage systems and a Ceph cluster provide the backends for Cinder (block storage) and Glance (images) services. Ceph, through its radosgw service [2], is also used as an object storage provider. A complex network configuration (described in detail in [1]) with four virtual routers and one or more class-C virtual networks for each OpenStack project implements the different access policies defined for INFN, University and non academic (e.g. related to collaboration projects with Public Administration or industry) users, also based on the ownership of the resources. User enrollment happens through a customised Horizon dashboard where the user authenticates with his/her own institutional Single Sign On system (INFN AAI or University of Padova SSO). Username/Password authentication is also available. Ganglia, Nagios and Cacti instances continuously monitor the whole infrastructure, while a Foreman/Puppet server is used for provisioning and configuration.

Table 1. Cloudveneto.it computing resources by owner.

Owner	Storage (TB)	# Compute Nodes	# Cores (in HT)	RAM (GB)	# GPUs
INFN	270	40	1680	5824	6
University	90	20	816	3552	6
Total	**360**	**60**	**2496**	**9376**	**12**

In 2019 the accounting system CAOS [3] based on Ceilometer/Gnocchi (the telemetry services recommended within OpenStack) has been dismissed due to some scalability problems, and replaced with a homemade solution built on the time series database InfluxDB [4], the analytics platform Grafana [5] and the system statistics collection daemon Collectd [6] (in particular relying on the virt plugin). The total capacity of computing resources, summarized in Table 1, achieved 60 compute nodes for a total of about 2500 logical cores and 9 TB of RAM. Moreover, two NVIDIA TITAN Xp, one NVIDIA Quadro RTX 6000, one NVIDIA GeForce GTX TITAN, four NVIDIA Tesla T4 and four NVIDIA V100 GPUs were gradually embedded in the infrastructure making use of the OpenStack support for

KVM hypervisor with PCI passthrough virtualization [7]. Some of them were used in a scientific computation described in section 4.

Besides the elastic on-demand HTCondor batch cluster service already provided to CloudVeneto.it users, another PaaS-type service designed to deploy a Big Data Analytics platform was developed and put in production during 2019.

3 Kubernetes clusters for Big Data Analytics

It is well known that the cloud is better exploited using a cloud-native approach in building and running applications, that basically have to be container packaged, dynamically managed and micro-services oriented. Cloud-native principles and open source software are fostered e.g. by the Cloud Native Computing Foundation (CNFC) [8]. In particular, two of its first "graduated projects" are the container orchestration framework Kubernetes [9] and the systems monitoring and alerting tool Prometheus [10]. Apache Kafka [11] and Apache Spark [12] are among the most popular open source Big Data Analytics tools maintained within the Apache Software Foundation. Kafka is a distributed highly scalable and fault-tolerant streaming platform used in thousands of companies for building real-time data pipelines and streaming applications. Spark is a unified analytics engine for large-scale data processing. It supports cloud-native deployments using Kubernetes as resource manager implementation since March 2018.

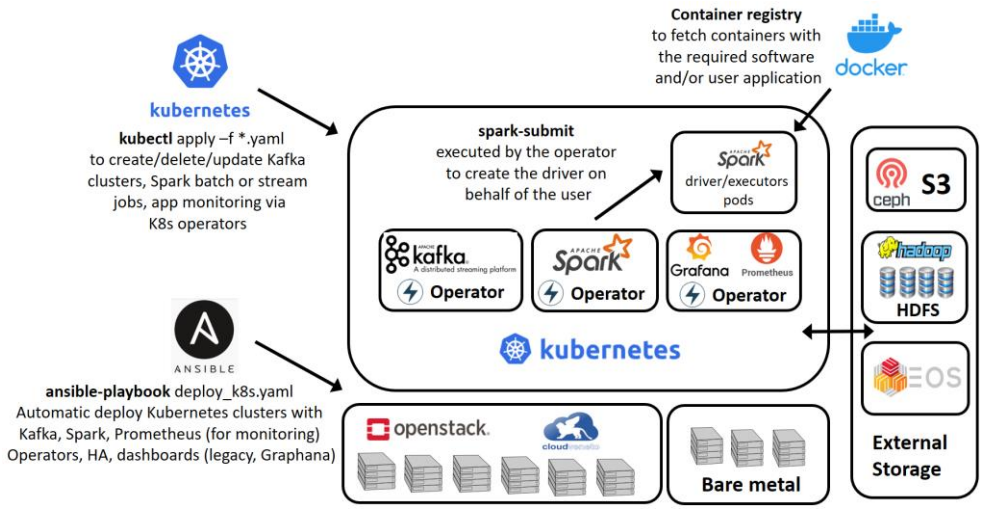

Fig. 1. Automatic deployment and execution of workloads on the Big Data Analytics platform based on Ansible, Kubernetes and Docker technologies.

We developed a set of Ansible [13] playbooks to automatically deploy Kubernetes clusters including Kafka, Spark and Prometheus operators. The deployment can optionally be in High Availability, both over cloud and bare metal resources, and includes the Kubernetes legacy dashboard and the Grafana dashboard for visualising cluster and application monitoring data. The Kubernetes operator pattern adds a further level of automation that simplifies running, monitoring and fine-grained lifecycle management of applications within a Kubernetes cluster. It enables in fact declarative application specification and management of application through Kubernetes custom resources. In the example shown in Fig. 1, a Spark application properly defined in a YAML file can be submitted via the Kubernetes client to the Kubernetes

API, and the Spark operator executes the spark-submit command on behalf of the user. The application software is typically packaged into container images uploaded into a Docker registry. While streaming data can be injected into a Kafka cluster and further processed by Spark, the access of data hosted in external storage systems like e.g. Ceph [2], HDFS [14] and CERN EOS [15] is also possible from the cloud through the appropriate protocols/connectors.

4 GPU use-case: the MAGIC experiment

The MAGIC experiment [16] aims at detecting primary γ rays originated from galactic and extragalactic sources with Imaging Atmospheric Cherenkov Technique (IACT). A set of telescopes with a mirror diameter up to 28 m located on the Canary island of La Palma at 2200 m above the sea detects the Cherenkov light emitted in extensive air showers initiated by very high energy rays hitting the atmosphere. The Cherenkov light can be reflected and focused by the mirrors and collected by a camera composed of several photomultiplier tubes. Electronic signal conversion and digitalization gives a pixelated image of the shower induced by the primary ray. The main problem affecting the identification of the primary γ rays is the huge background coming from the cosmic rays, with a signal to noise ratio less than 1/2000. The University of Padova researchers working in the MAGIC experiment designed and implemented a novel full analysis based on a Deep Learning pipeline from the pixel-wise information. The new method, described in detail in [17], improved significantly both the shower direction reconstruction (~ 20%) and the energy reconstruction (~ 30% above 1 TeV), over the standard MAGIC analysis, which inevitably loses potentially useful information in its image parametrization step. Fig. 2 shows a simplified comparison of the two analyses.

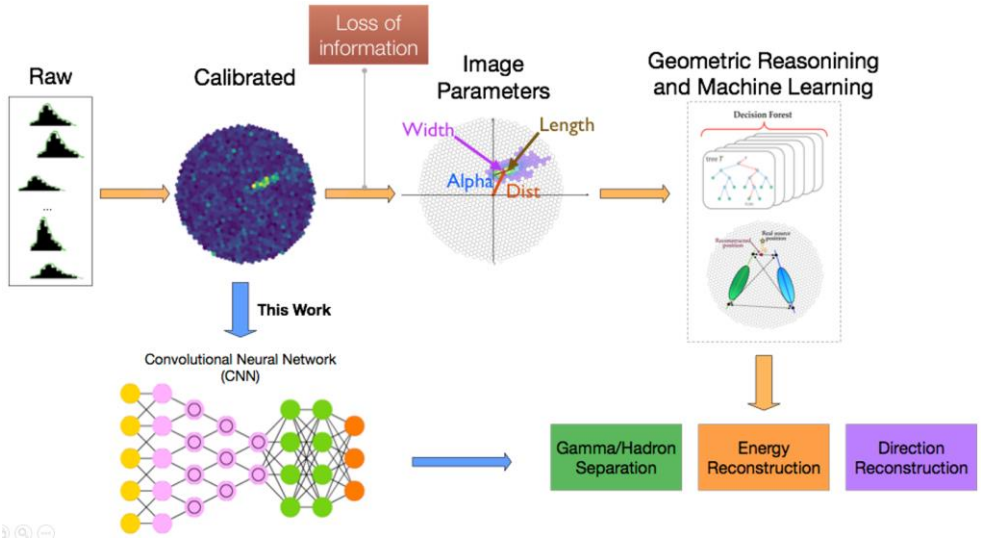

Fig. 2. The new analysis approach based on CNN working directly from the calibrated data compared with the traditional analysis pipeline.

We configured and set up customized images and specific flavors to allow users to create, in our cloud infrastructure, a virtual machine with one or more GPUs. The training and optimization of the Convolutional Neural Network (CNN) adopted in the new analysis could greatly benefit from the use of GPUs integrated in CloudVeneto.it. By using a virtual server

connected with a NVIDIA Titan GPU, a factor 10 of speedup compared with the use of a 24 CPU core server was achieved in the training time. The reading performance of the 250 GB input dataset was initially a serious bottleneck for the computation pipeline when using magnetic hard disks on the hypervisors hosting the GPU cards, so we decided to upgrade the bare metal hypervisor with solid state disk (SSD-NVMe) and customize how the instances execute I/O instructions. This improvement was crucial to achieve the full utilization of GPUs connected to the virtual servers, because the non-sequential reading performance increased by more than one order of magnitude.

5 The CMS experiment use-case

The University of Padova researchers working in the CMS experiment exploited the Big Data Analytics platform of CloudVeneto.it described in section 3 to perform online data processing of the CMS muon detector composed of Drift Tube (DT) chambers [18]. They built a prototype at INFN-Legnaro Laboratories with two DT chambers for testing the online event reconstruction with data streams coming from the front-end electronics at the same rate as the LHC clock (40 Mhz), without archiving data on disk. Muon tracks crossing the DTs generate hits converted into electronic signals which are digitized and acquired by an FPGA board hosted on a server running a Kafka producer. This broadcasts the hit data to the Kafka cluster hosted in CloudVeneto.it as "messages" of 1kB minimum size (128 hits). Kafka brokers provide then the data to the Spark cluster. Spark executors process data from Kafka through the Spark Streaming API that performs event reconstruction and produces useful data for monitoring the detector status. Spark outputs are injected back to Kafka that makes them available for monitoring and runtime visualisation with Prometheus, Grafana and Bokeh [19]. Fig. 3 shows the entire setup and data processing workflow.

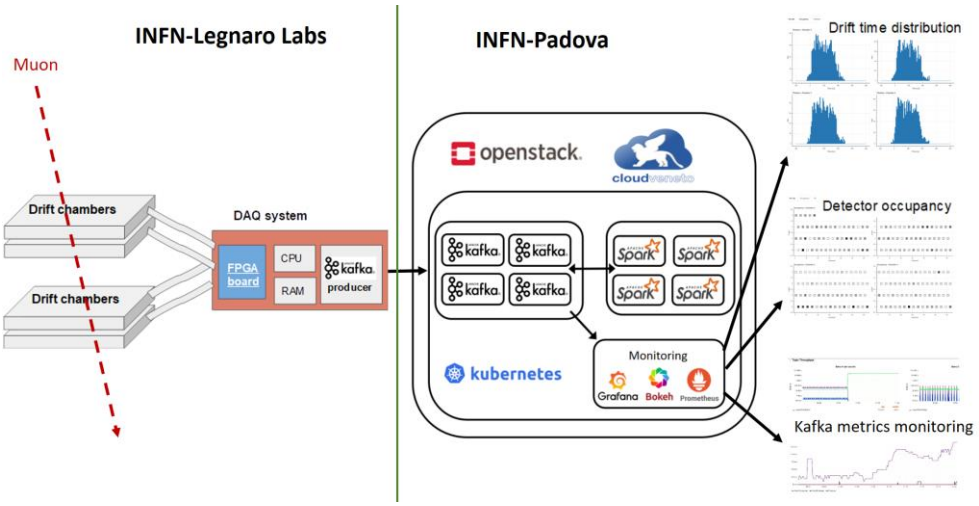

Fig. 3. Muon DT chambers Online data acquisition setup at INFN-Legnaro Labs and remote streaming processing through the Big Data Analytics platform of CloudVeneto.it.

CloudVeneto.it implementation through Kubernetes enabled the maximum flexibility in changing both Kafka and Spark clusters configuration. This allowed us to perform extensive

tests which demonstrated the scalability of the system up to the level of throughput expected at HL-LHC from the DT chambers.

6 Conclusions

As discussed in this article, CloudVeneto.it infrastructure has evolved to face not only its continuously growing user base, but also the increasingly high demand of Big Data and Machine Learning workloads required by scientific applications that need accelerated hardware and high level services on top of the lowest IaaS cloud level.

As a next challenge, we are exploring the possibility to confederate CloudVeneto.it into a larger INFN nationwide cloud infrastructure geographically distributed across a few big INFN data centers, which is expected to enter gradually in production during 2020.

References

1. P. Andreetto et al., EPJ Web of Conferences **214**, 07010 (2019)
2. Ceph home page, https://ceph.io/
3. P. Andreetto et al., EPJ Web of Conferences **214,** 07006 (2019)
4. InfluxDB home page, https://www.influxdata.com
5. Grafana home page, https://grafana.com
6. Collectd home page, https://collectd.org
7. P. Andreetto et al., PoS (ISGC2017) 020, https://doi.org/10.22323/1.293.0020
8. Cloud Native Computing Foundation home page, https://www.cncf.io
9. Kubernetes home page, https://kubernetes.io
10. Prometheus home page, https://prometheus.io
11. Kafka home page, http://kafka.apache.org
12. Spark home page, https://spark.apache.org
13. Ansible home page, https://www.ansible.com
14. Hadoop home page, https://hadoop.apache.org/
15. A.J. Peters et al., Journal of Physics: Conference Series **664,** 042042 (2015)
16. MAGIC home page, https://magic.mpp.mpg.de
17. E. Mariotti, *Deep Learning on MAGIC: a Performance Evaluation for Very High Energy Gamma-Ray Astrophysics*, Thesis submitted for the degree of Laurea Magistrale In Telecommunication Engineering, University of Padova, April 2019.
18. M. Migliorini et al., *Big Data solutions for the online processing of trigger-less detectors data*, Proceedings of CHEP2019.
19. Bokeh home page, https://bokeh.org