Contribution ID: **22** Type: **Short Talk**

Virtual Data Center: a platform for enabling data scientists to access integrated and scalable online environments

Wednesday, 2 October 2024 17:05 (10 minutes)

The advancement of **EOSC** promotes a research paradigm more reproducible and verifiable in response to the growing complexity and interdisciplinarity of modern research, necessitating an unprecedented level of collaboration and data sharing. In line with this, federated data infrastructures, like the **Blue-Cloud project**, have been established, integrating marine data sources across Europe to catalyze advancements in marine science. Among these initiatives, the *NEw REsearch Infrastructure Datacenter for EMSO* (**NEREIDE**) developed by INGV and located near the Western Ionian Sea facility, is designed to drive data science forward through its Virtual Data Center (VDC) platform.

The core of NEREIDE innovation is the **Virtual Data Center**, a managed environment where Data Scientists (DSs) can control cloud resources, including Virtual Machines (VMs), behind a dedicated customizable **Gateway** with a public IP address. DSs have administrative privileges over their cloud segments, enabling them to **autonomously** create VMs and manage network components including firewalls, VPNs and advanced routing configurations. Meanwhile, the overarching management of cloud infrastructure, including the physical data center, remains in charge and under the control of the Infrastructure Administrators. This **dual-structure** ensures a balance between stringent infrastructural measures and DSs operational autonomy. VDCs provide a sophisticated, plug-and-play infrastructure that sidesteps the complexities of traditional data center management, allowing DSs to focus on their **data services**, bypassing the complexities of traditional physical data center management.
DSs want to work into dynamic and customizable environments where they manage substantial computational resources to tackle complex scientific questions. The main component of VDCs is **OpenStack**, an open-source "Infrastructure as a Service"platform that enables seamless scalability and flexibility in resource management, aided by workflow automation tools such as **MaaS** and **JuJu**. This setup allows DSs to optimize computing and storage capacities according to project needs, essential for processing extensive datasets and performing complex simulations.
VDCs rely also on **Ceph**, a distributed software defined storage engine, which offers flexible and scalable storage resources, in conjunction with data security and integrity. This solution allows DSs to face heavy scientific data loads and to efficiently manage multiple storage types.
Additionally, VDCs not only provide bare computational power and raw storage space to DSs, but also enable the integration of sophisticated arrays of tools, such as **JupyterHub** for interactive data analysis, **ERDDAP** for data distribution, **ElasticSearch** for data querying, etc. These tools underpin data science activities including data analysis, visualization, and collaborative research, thereby making complex data comprehensible and accessible across various scientific domains.

Essentially, with the introduction of a custom gateway, VDCs represent an evolution of the concept of virtualization, extending it beyond individual virtual machines to include an entire ready-to-use data center infrastructure.

The potential integration of VDC platforms within federated data infrastructures, like Blue-Cloud, suggests a future where seamless data and resource sharing could significantly boost the analytical and operational capacities within different scientific domains. These advancements foster new scientific applications and innovations, accelerating the achievement of open science goals and easing the work of data scientists.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: CACCIAGUERRA, Stefano (INGV); Dr CHIAPPINI, Stefano (INGV)

Presenters: CACCIAGUERRA, Stefano (INGV); Dr CHIAPPINI, Stefano (INGV)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways