Cross-Layer Infrastructure Optimization for Data-Centric Applications

Yufeng Xin
RENCI, UNC-CH
USA
yxin@renci.org

Yin Chen
EGI Foundation
the Netherlands
yin.chen@egi.eu

Zhiming Zhao
System and networking Lab
University of Amsterdam
the Netherlands
z.zhao@uva.nl

GEFI 2019 Attendees: Zhiming Zhao, Yufeng Xin

I. Research Description

Modern data-centric applications are among the major drives for next generation Internet and network infrastructure innovation. These applications, often founded in broad societal challenges such as overpopulation and diminishing natural resources, cut across many different scientific domains and require collection, transfer, and processing capabilities on data from broad range of sources.

These applications can only be effectively enabled however in the presence of a supporting research infrastructure, which should provide the necessary tools for searching, accessing and integrating data and software for different workflows within scientists research activities. Recent paradigm shift towards data centric approaches further motivated the development of advanced network and computing technologies, e.g., SDN (software defined networking), ICN (Information-Centric Networking) and 5G, as well as the Cloud technologies in Edge Cloud and machine learning (ML). In the following, we use our recent research experience in supporting environmental research as an example to help lay out our collaborative research agenda.

The Environmental research infrastructures (RIs) have been developed to support user communities by integrating large-scale sensor and observer networks with dedicated data curation, data analysis tools and other research assets installed behind a unified service interface[1], ICOS (atmospheric science)[2], and SeaDataNet (ocean/marine science) [3].

Integrating large number of heterogeneous RIs incurs many challenges to the networking infrastructure. The first challenge is the limited FAIRness (Find-ability, Accessibility, Interoperability and Reusability) of digital assets (data, software and services) to make discovering suitable data objects from different RIs less difficult and time consuming. Several projects are recently funded to tackled such FAIRness challenges, e.g., ENVRI-FAIR[4]. Secondly, the evolution of the digital objects results in different versions, which often leads to complicated dependencies in creating new objects. Without rich contextual information of different versions, the (re)usage of data objects in a data workflow is often error prone. Thirdly, many data applications are often based on community collaboration, for example, composing or collaborative editing of online education or training material using community contributed media. As the size of digital objects are very large, e.g., video content, the performance of the object sharing is critical for high quality user experiences. Centralized data storage often creates performance bottlenecks for distributed users. Lastly, data privacy and security are always big concerns.

Advanced virtual infrastructure service has been identified to be a promising solution. With this service capability, user defined services and workflows can be efficiently and securely deployed, discovering, accessing and processing data from different RIs will be highly efficient, and performing more advanced data centric researches by scientific users becomes much easier. However, control and management of virtual infrastructures, across different RIs (senors and instruments), network (wired and wireless), and Cloud (Internet and Edge) infrastructure, remain very difficult.

A. Research Collaborations

In the past, within the context of GENI collaboration, we have developed necessary information model, notably the NDL (Network description language), and end-to-end resource provisioning mechanisms in federating distributed network and Cloud resource.

In this collaboration, we aim to further investigate the optimization models for virtual infrastructure dynamic control with the RIs and the up-layer data-centric applications explicitly in the loop.

1) Virtual Infrastructure control across the distributed Cloud-Edge-RI infrastructure system. While there are many Cloud and SDN infrastructural components available in US, EU, etc, there is no existing mechanism to link the RI testbeds to these Cloud/Network assets. We will explore (a) information model for resource co-allocation; (b)a programmable virtual networking interconnectors between the two systems. The goal is to be able to provision end-to-end virtual experiment system (data source - edge - Cloud) with data-centric constraints (latency, bandwidth, deadline, security, etc).

[1] epos-ip.org/
[2] www.icos-cp.eu/
[3] www.seadatanet.eu
[4] www.envri-fair.eu

2) Time critical virtual infrastructure management: in order to map and reconfigure applications/experiment to the infrastructure intelligently, we need real-time state information of the multi-domain infrastructure. Yet, monitoring capability is very limited to applications. We will explore a new tomography-like approach, using advanced machine learning techniques, to dynamically draw a heat map of the inter-Cloud infrastructure based on endmost measurement and traffic statistics. With this capability, it is possible to compose and experiment with different new network technologies in overlays, like ICN and network slicing, to support real-time data applications.
3) Applications on heterogeneous virtual infrastructure: experiment with virtual container clusters connecting to a set of distributed (real or virtual) data sources. Several use cases have been identified in the context of a number of EU H2020 projects, include ENVRI-FAIR, ARTICONF and BlueCloud.

II. International Collaboration

Our collaborative team consists of researchers from the infrastructure, middleware, and data centric applications. We have long history of developing, operating, and conducting research on different kinds of experimental infrastructures that are complimentary to this collaboration.

RENCI. Researchers from RENCI and UvA have long history of collaboration, from the semantic web based resource description model (NDL), to the federated Cloud infrastructure, to the experiment based research covering networking, Cloud, and data applications. Dr. Xin was the lead developer for ExoGENI control framework. He is also the lead at RENCI in the just announced NSF PAWR Award led by NC State University. He visited UvA a few times for the research and infrastructure collaboration.

University of Amsterdam. The University van Amsterdam (UvA) embeds a concentration of e-Science development and application groups, benefiting from its central position in the Science Park Amsterdam. The park holds the main European Internet Exchange (AMS-IX), one of the SURFnet core locations including the optical network exchange NetherLight, the national supercomputer and grid center (Surf-SARA) and a number of Dutch Science Organisation institutes. Dr. Zhiming Zhao leads Quality critical distributed computing team within the Multi-Scale Networking group, System and Networking Lab. He leads the development effort in the EU H2020 ENVRIPLUS and ENVRI-FAIR projects. During the past years, he focuses on a novel DevOps framework, time critical cloud computing, and the infrastructure optimization for the big data management. His team is an active user of both GENI and EGI infrastructures.

EGI Foundation. The Stitching European Grid Initiative (also referred to as EGI.eu) is a not-for-profit foundation established under Dutch law to coordinate and manage the European Grid Infrastructure (EGI) federation on behalf of its members: National Grid Initiatives (NGIs) and European International Research Organisations (EIROs). Building on over a decade of investment by national governments and the European Commission, EGI supports more than 22,000 researchers across many scientific fields with a wide range of technical services such as high-throughput data analysis, federated cloud, federated operations and community building.

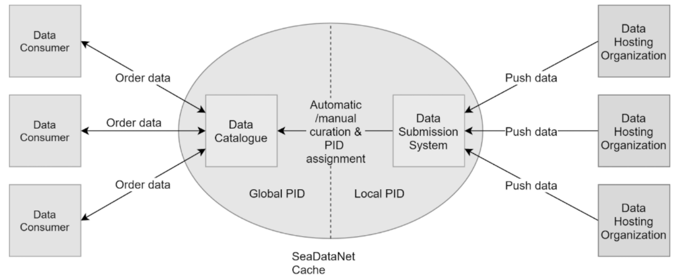

Fig. 1. SeaDataNet's current infrastructure.

In near term, the team plans to conduct collaboration in two major areas: (1) infrastructure federation; (2) experimental data-centric application research. We use the architecture of our ongoing SeaDataNet project as a concrete example, shown in Fig.1. We are developing software solutions for two representative scenario:

1) Data management scenario. SeaDataNet originally only gathers metadata from remote repositories, and a user has to access the individual remote repositories. [5] It is possible to leverage the edge Cloud system to move this process closer to the repositories.
2) On demand workflow scenario. A user can initialize scientific workflow based on the data, models and other software tools provided by SeaDataNet or his/her own repository. Based on the geo-locations, the envisioned cross-layer optimizer provisions customized virtual infrastructure and executes the workflow for the user in a cross-layer control loop.

Potential Testbed federation. We envision a federated virtual infrastructure among several partners: NSF GENI, Cloud and PAWR Testbeds. They are built as distributed infrastructure system with wide geographical footprint including a federation site at UvA. We are further extending the control and management of the testebd toward an edge Cloud architecture, which could be federated with our environmental RIs to support our data centric application research agendas.

EGI.eu offers a variety of services to the wider EGI community such as overseeing infrastructure operations, coordinating user community support, working with technology providers, representing EGI in collaborative projects. EGI.eu and its members provide EGI, a pan-European infrastructure of publicly funded computing, storage and data resources to support excellent research and innovation in Europe.

[5] We distinguish repository and registry based on the Digital Object Architecture: registry is for storing metadata while the repository is for storing the data objects. They can often be combined to obtain data.