Blueprint for an Open Source On-Premise Cloud Infrastructure to Serve as a Research Data Infrastructure for Universities

Raimund Vogl[1], Jürgen Hölters[1], Martin Ketteler-Eising[1], Dominik Rudolph[1], Markus Blank-Burian[1], Holger Angenent[1], Christian Schild[1], Stefan Ost[1]

[1]Zentrum für Informationsverarbeitung, Westfälische Wilhelms-Universität, Röntgenstraße 7-13, 48149 Münster, Germany, {rvogl|holters|mke|d.rudolph|blankburian|holger.angenent|schild|ost}@uni-muenster.de

Keywords
Research data management, European open science cloud, European Data Infrastructure, IT infrastructure, ceph, openstack, repository, open science, reproducible research.

1. Summary

In response of the significant increase in the amount and variety of research data, the European Data Infrastructure (EDI) and the European Open Science Cloud (EOSC) foster persistent, highly available and compatible data infrastructures where data from various disciplines can be stored and accessed. These infrastructures should not only provide storage but also tools for processing and analysis. To prepare the implementation of such an extensive research data infrastructure for a group of five Universities, Münster University (lead of the consortium) has invested substantial manpower in developing a versatile, scalable and performance optimized hyperconverged deployment of OpenStack (cloud stack for virtual machines and Storage as IaaS) and Ceph (as underlying Software Defined Storage) using kubernetes as container orchestration engine on industry standard hardware. This is the first instant that advanced leading edge cloud technology like kubernetes has been put to use in any of the participating university IT centers and we see this as pivotal for out future approach to system architecture. The Open Source approach was adopted for cost reduction and sustainability. Remarkable is the approach to build on community versions of the Open Source software only, without vendor support. A scaled down pilot system has been operational for well over a year now, and demand for such an infrastructure is mounting from numerous research groups from a wide range of disciplines. Implementation of the full scale cloud system is planned for mid 2019. This is an update on the very preliminary report on the project given at EUNIS 2018.

2. Extended Abstract

The IT service facilities of the five German universities Bielefeld, Bonn, Münster, Paderborn and Siegen have formed a consortium to jointly develop a synergetic, sustainable and cost-efficient research data infrastructure using standardized hardware components and free open-source software.

A joint operating group with experts from all five universities was established early on to guarantee an efficient operations management and to support the self-reliant community only Open Source software approach without vendor support, sharing the work of testing and packaging new software versions. Open Stack and Ceph are deployed as kubernetes packages for flexible scalability and resilience.

With regard to functionality, first the storage of research data (Ceph), and, second, fast access to extensive data sets for processing and analysis via an integrated platform for virtual machines and containers (OpenStack), are of central importance according to potential users. The OpenStack platform also serves as an execution environment for research data containing executable code – "executable" publications to allow access and reuse of open data through publication of data together with integrated execution for analyses have been pursued at WWU with the O2R (Open Reproducible Research) project, and sustainable development and availability of research software is the key focus of the pyMOR (python Model Order Reduction) project. Apart from a comprehensive feature setup to insure data integrity build into Ceph, the architecture of this platform will also allow to store data

redundantly in multiple data-centers (even geo-redundant at the sites of the cooperating universities) for disaster protection for important data sets. Fault tolerance and resilience in the application layer is achieved by deploying microservice containers under the container orchestration engine kubernetes.

Detailed surveys of user demands at all 5 participating universities have resulted in a design of a Ceph/OpenStack infrastructure with a total of 120 hyper-converged storage and virtualization hosts with 100 GE networking and a total of 33 Petabyte of disk capacity, to be installed as autonomous OpenStack clouds at each of the 5 universities.

A special interface for this infrastructure will be established through a set of research data services (RDS), which will help researchers to automate data management workflows: from maintaining data management plans, annotating metadata and persistent identifiers, creating and submitting archival packages, to making data available according to rulesets. Easy access to these RDS will be provided through the well-established cloud storage platform “sciebo”, which is already a key collaboration space for researchers in the German state of North Rhine-Westphalia, hosting approximately 110,000 users and 2,100 projects. This approach to a joint multi-university research data infrastructure is inspired by the ideas of the European Open Science Cloud (EOSC) and the European Data Infrastructure (EDI), and also represents a first step towards the National Research Data Infrastructure planned for Germany.

REFERENCES

(1) Lopez A, Vogl R, Roller S (2017) Research Data Infrastructures - A Perspective for the State of North Rhine-Westphalia in Germany. In: EUNIS 2017, Münster , Book of Proceedings p 105ff (http://www.eunis.org/download/2017/EUNIS2017_Book-of_Proceedings_1.pdf).

(2) The Rectorate of the WWU Münster has adopted on 11.5.2017 the "Principles for the handling of research data" in which it is committed to the condition for the fulfillment of these principles to create (https://www.uni-muenster.de/Forschungsdaten/)

(3) European Commission (2016) The European Open Science Cloud (https://ec.europa.eu/research/openscience/pdf/realizing_the_european_open_science_cloud_2016.pdf)

(4) Open Commons Consortium (http://occ-data.org/)

(5) The FAIR Principles for Research Data - Findable, A Accessible, Interoperable, Reusable (http://www.forschungsdaten.org/index.php/FAIR_data_principles)

(6) A Data Biosphere for Biomedical Research (https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d)

(7) Genomic Data Commons Data Pool (https://portal.gdc.cancer.gov/)

(8) The OCC Environmental Data Commons (https://portal.gdc.cancer.gov/)

(9) The Jetstream Project (http://jetstream-cloud.org/)

(10) Vogl, R., Rudolph, D., Thoring u. a. (2016) How to Build a Cloud Storage Service for Half a Million Users in Higher Education: Challenges Met and Solutions Found. In: Proceedings of the 49th Annual Hawaii International Conference on System Sciences (HICSS 2016), p. 4272-4281. doi: 10.1109/HICSS.2016.658.

3. AUTHORS' BIOGRAPHIES

R. Vogl is the CIO of the University of Münster (Germany) and is the director of the University IT center since 2007. He holds a PhD in elementary particle physics from University of Innsbruck (Austria). After completing his PhD studies in 1995, he joined Innsbruck University Hospital as IT manager for medical image data solutions and moved on to be deputy head of IT. He is board member and president of EUNIS (European University Information Systems Organisation) and active as a member of the IT strategy board of the Universities in North Rhine-Westfalia. He is a member of GMDS, EuSoMII and AIS and is representing Münster University in EUNIS, DFN, ZKI, DINI and ARNW. His current research interest in the field of Information Systems and Information Management focuses on the management of complex information infrastructures. More Info: http://www.uni-muenster.de/forschungaz/person/10774

J. Hölters is the deputy head of the department for systems at the IT center (Zentrum für Informationsverarbeitung, ZIV) of the University of Münster (Germany).

More info: https://www.uni-muenster.de/forschungaz/person/8172

M. Ketteler-Eising is a research assistant at the Zentrum für Informationsverarbeitung (the IT center) of the University of Münster (Germany).

More info: https://www.uni-muenster.de/forschungaz/person/20615

D. Rudolph is managing director of the IT center (Zentrum für Informationsverarbeitung, ZIV) of the University of Münster (Germany). He received his PhD from the University of Münster, where he also studied communication sciences, economics and modern history. His graduate thesis has been appraised as one of the best dissertations in 2014 (German Thesis Award). His research focuses on the diffusion of innovations, the management of research data, and digitalization processes in higher education. More info: https://www.uni-muenster.de/forschungaz/person/7445

M. Blank-Burian is a research assistant at the Zentrum für Informationsverarbeitung (the IT center) of the University of Münster (Germany).

More info: https://www.uni-muenster.de/forschungaz/person/17330

H. Angenent has studied physics at the Technical University of Ilmenau and the University of Münster (Germany). He worked four years as a research assistant at the institute of Theoretical Physics in Münster. Since 2010 he is a research assistant at the Zentrum für Informationsverarbeitung (the IT center) of the University of Münster and responsible for high performance computing systems and cloud services. More info: https://www.uni-muenster.de/forschungaz/person/8041

C. Schild is a research assistant at the Zentrum für Informationsverarbeitung (the IT center) of the University of Münster (Germany) and responsible for authentication systems. He studied physics at University of Münster. He was a leading member of the JOIN project at the university of Münster which participated in several european cooperative projects (some ipv6 task forces, 6WIN, 6NET) to develop and propagate the IPv6 protocol.
More info: https://www.uni-muenster.de/forschungaz/person/11099

S. Ost is the deputy director of the Zentrum für Informationsverarbeitung (the IT center) of the University of Münster (Germany) and head of the department for systems. More info: https://www.uni-muenster.de/forschungaz/person/10355