EPJ Web of Conferences **245**, 07028 (2020) https://doi.org/10.1051/epjconf/202024507028
CHEP 2019

Pre-Commercial Procurement: R&D as a Service for the European Open Science Cloud

Marion Devouassoux[1], *Bob* Jones[2] , and *João* Fernandes[3]

[1,2,3]CERN, Information Technology Department, Geneva, Switzerland

Abstract. The economical and contractual aspects linked to the production use of commercial cloud services are often overlooked in research environments in Europe, preventing researchers from reaping the full benefits of these services. Since 2016, CERN, in collaboration with leading European research institutes, has launched several projects to address this problem. The preparation and execution phases of these projects have revealed key lessons in procuring commercial cloud resources for research environments in Europe. These lessons can help shape the European Open Science Cloud (EOSC) [1].

1 Introduction

CERN recently started the High-Luminosity Large Hadron Collider (HL-LHC) project with the aim to increase the number of collisions by a factor of 10 beyond the LHC's design value [2]. Such an upgrade generates a considerable surge in the volume of data produced. Estimates show that the storage needs for raw data will increase by a factor of 12 between 2016 and the launch of the HL-HLC in 2027 [3]. This factor reaches 60 when it comes to CPU needs (kHS06) [3]. Assuming constant cost, these estimates are approximately 10 times superior to what can realistically be expected from technology [3]. This challenge is not specific to the High Energy Physics domain. As an example, the European Bioinformatics Institute (EMBL-EBI) is also facing a data deluge as their current archive of 20PB is doubling every two years [4]. One solution to adapt the computing capacity to the growing needs of research institutes is to complement on-premise resources with commercial cloud services. A successful example of the combination of both types of resources for scientific purposes is the recent re-performance on top of Kubernetes of the data analysis that led to the discovery of the Higgs boson [5]. Modern data science techniques combine for example container technology and other new functionalities available in the market with scientific research statistical algorithms. This mix of resources and technologies can become a key enabler to achieve full research reproducibility. Yet, the economical and contractual aspects linked to the production use of commercial clouds are often overlooked, preventing researchers from reaping their full benefits. Since 2016, CERN, in collaboration with leading European research institutes, has launched several projects to bridge this gap. After providing an overview of the projects where CERN has a leading role in procuring commercial cloud services, this paper highlights the main results. The third section summarises fundamental

lessons learned from these projects and outlines elements to consider when involving commercial cloud providers in the EOSC.

2 Experience in Procuring Commercial Cloud Services

The Helix Nebula Science Cloud (HNSciCloud) project [6] used a Pre-Commercial Procurement (PCP) [7] approach to competitively procure R&D services from the industry and develop innovative solutions for public sector needs. With a budget of €5.3 million, the project developed a European hybrid cloud platform, linking commercial cloud service providers and publicly funded research organisations' in-house IT resources via the GEANT network. The platform was developed to support the deployment of high-performance computing and big-data capabilities for scientific use cases of 10 European leading public research organisations. The project started in January 2016 and was successfully completed in December 2018. CERN won the Procura+ Award 2019 for "Outstanding Innovation in ICT Procurement" [8] in the HNSciCloud project.

Leveraging some of the experience gained from HNSciCloud, CERN is participating in the Open Clouds for Research Environments (OCRE) project [9]. OCRE aims to accelerate cloud adoption in the European research community, by bringing together commercial cloud providers and the research and education community. The mechanism for this purpose is a pan-European tender and framework agreements with cloud service providers that meet the specific requirements of the research community. OCRE will also make €9.5 million in service credits available to research institutions and Earth Observation organisations.

In parallel, CERN is leading the Archiving and Preservation for Research Environments (ARCHIVER) PCP project [10]. ARCHIVER will introduce significant improvements in the area of archiving and digital preservation services, supporting the IT requirements of European scientists. With a procurement budget of €3.4 million, ARCHIVER will develop cost-effective solutions for data archives in the petabyte range with high, sustained ingest rates.

Finally, CERN is a partner in the EOSC-hub project [11] that brings together European research infrastructures and the core e-infrastructure community to develop a common catalogue of data, services, and software for research for the EOSC. Work Package 12 of EOSC-hub contributes to the design of future business models and procurement frameworks for acquiring digital services from both publicly funded and commercial providers. It identifies current and preferred delivery models for such services as well as funding streams and procurement strategies in order to propose areas of improvement for EOSC business models.

3 Key Results

The following section lists key results of the above-mentioned projects, starting with two business models that were successfully tested: the Buyers Group model and cloud vouchers.

A Buyers Group can be defined by a group of organisations committing funds to a shared procurement budget overseen by a Lead Buyer. These organisations are legally represented by the Lead Buyer organisation and sign a Joint Procurement Agreement (JPA) that includes the financial commitments and management aspects of the tender process. Buyers Group

members must be in a position of handling the procurement process, internal accounting, invoicing and payments. Since 2016, CERN is the Lead Buyer of two Buyers Groups:

- The HNSciCloud Buyers Group composed of ten research organisations (CNRS, CERN, EMBL, ESRF, DESY, IFAE, INFN, KIT, SURFsara, and STFC) from across Europe, purchasing R&D of cloud services to support their scientific programmes.
- The ARCHIVER Buyers Group composed of four research organisations (CERN, DESY, EMBL-EBI and PIC) representing a diverse range of scientific domains and use cases for archiving and digital preservation services.

Aggregating procurement funds through a Buyers Group has several advantages:

- The procurement process is faster, simpler and more cost-efficient as it is carried out by one entity (i.e. the Lead Buyer).
- The aggregation increases demand as some organisations are unlikely to engage in isolated procurement activities.
- Higher volumes stemming from the demand aggregation generate economies of scales leading to volume discounts and preferential terms.
- All members benefit from past experience and acquired expertise in selecting digital services for research, assessing services against key requirements, such as data sovereignty, data protection and security.

Details on the Buyers Group definition are available in the report "Research Infrastructures & NRENs Requirements for Commodity Cloud Services" [12].

The second business model explored the use of pre-purchased vouchers for accessing commodity cloud services in research environments. Cloud vouchers were initially used in the HNSciCloud project pilot phase to encourage the uptake of commercial cloud services by end-users. Indeed, while large-scale services meet the needs of the project's Buyers Group, a flexible scheme was needed to lower the entry barriers to cloud services for new users. Cloud vouchers have proven to be efficient for this purpose as they are easy to redeem within an organisation, have a defined face value suitable for small-scale projects, and enable researchers to explore innovative architectures (such as GPUs, FPGAs) and software libraries (such as TensorFlow) before adopting them in production. The description of the process followed in the HNSciCloud project to test vouchers is detailed in a report published by the project [13]. The use of cloud vouchers is currently being tested at a larger scale in the OCRE project, where a significant share of the project budget will be distributed to institutions and Long-Tail-of-Science researchers in the form of cloud vouchers.

Another key result springing from these projects is a market analysis of current usage and requirements for commodity cloud services in the research sector across Europe. The EOSC-hub published a report presenting the results of a demand-side market survey to understand the need and level of demand for digital services for research in the context of the EOSC [14]. The report explores the manner in which the need and demand are currently met and the challenges faced to support analysis workflows, data management, and related infrastructure and services. This survey highlighted a growing demand for digital services for research, especially for data repositories, data registries services, analytics, data management tools, collaboration services and general computational services.

In parallel, in the context of the OCRE project, CERN analysed the requirements for commodity cloud for Research Infrastructures and National Research and Education Networks (NRENs) [15] as well as the Long-Tail-of-Science [16] in view of preparing a pan-European tender. These groups of stakeholders have similar requirements:

- They are mostly interested in compute (often orchestrated through software container technology) and storage (object storage capacity and synchronization and sharing services).
- There is an increasing need for Machine Learning and Deep Learning algorithms that often require the availability of large amounts of accelerator hardware (such as GPU and FPGAs).
- Similarly, there are growing requirements for engineering software tools and services (such as MATLAB, Mathematica, Comsol, SolidWorks, Ansys, and AutoCAD).

The analysis also shows that Research Institutions need a testing environment for commercial cloud services before procuring at scale [17]. To this end, a technical validation test-suite [18] was created for both the OCRE and ARCHIVER project. This test-suite enables the Buyers Group to package and easily deploy tests derived from scientific use cases. The test-suites include tests from multiple domains providing a scalable and uniform way of deploying validation tests and therefore assessing, transparently and efficiently, commercial cloud services.

Finally, financial attractiveness of the solutions developed during a project is a key success factor of these projects. To ensure that this factor is taken into account, contractors of the HNSciCloud project were asked to perform a Total Cost of Ownership (TCO) study of their proposed solutions by the end of the pilot phase. The study was carried out for two selected scientific use cases in the fields of High Energy Physics and Biology. The study was an iterative and collaborative process, where the Buyers Group provided the high-level input requirements and necessary workflow information. The outcome of this process are two comprehensive studies of the TCO of the hybrid cloud platform developed by consortia led by T-system [19] and RHEA Group [20]. These TCO studies showed that while the PB-scale disk storage currently remains cheaper on-premise, the compute workload is competitive on commercial cloud. Such TCO studies have been included in the ARCHIVER work plan.

4 Lessons learned and Implications for the EOSC

During the preparation and implementation phases of these projects, several lessons concerning the procurement of commercial cloud resources for research environments were identified. These lessons can be valuable input for the definition of the Rules of Participation for cloud providers in the EOSC.

First, some of the identified business models have proven to be efficient and therefore can be beneficial for the EOSC. As described above, Buyers Groups generate volume discounts due to the demand aggregation, increase demand and simplify the tendering process. The HNSciCloud project demonstrated that nominating a Lead Buyer that already had long-standing relationships with the members of the Buyers Group is a successful approach. The close cooperation between the members of the Buyers Group was essential to the success of the project. Furthermore, careful planning is needed concerning the management of the procurement exercise itself in order to achieve a genuine commitment from the buying organisations. To this end, the ARCHIVER project has established a Joint Procurement Agreement that defines the roles of each Buyer Group member and establishes its governance as well as the levels of pre-commitment of funds for procurement. A binding document that fulfils this role is a necessity to ensure the smooth functioning of Buyers Groups in the EOSC.

Additionally, a cloud voucher is an efficient mechanism to encourage the uptake of commercial cloud services made available in the EOSC catalogue. This statement is in line with the publication "Prompting an EOSC in practice" [21], which recommends investigating the use of euro-denominated EOSC vouchers as a mechanism to grant access to researchers to cloud resources within the EOSC. Based on the experience gained while distributing cloud vouchers in the HNSciCloud project, best practices have been identified and documented [22]. To be attractive to end-users, voucher schemes in the EOSC should include the following elements:

- A fixed face value in monetary units
- Be free at the point of use
- Have a defined duration of validity
- No limitation on the number of vouchers that can be consumed by an end-user
- Be promoted via a defined catalogue of all services accessible with the voucher
- Procured services to possess up-to-date and detailed documentation and tutorials for end-users
- Near real-time consumption monitoring accessible by end-users
- An applicable Service Level Agreement
- A clear data repatriation policy and transparent associated costs

The Buyers Group and cloud vouchers are suitable business models for the EOSC. However, both models imply a cyclic access to cloud resources. Researchers can access cloud resources only until the voucher is exhausted while members of a Buyers Group can procure resources under a framework agreement that has a limited duration. To ensure their viability, a comprehensive data management solution must be associated with these models. This plan must offer cost-effective and easily implementable solutions to repatriate, transfer or archive data uploaded on commercial clouds.

To this end, a risk mitigation strategy for vendor lock-in by cloud service providers should be provided in the EOSC. The switching and porting Codes of Conduct [23] that was launched by the European Commission is an important step in this direction and should be taken into account for the EOSC Rules of Participation for service providers. In addition, the EOSC should promote the use of open source software, vendor independent standards and interfaces to port data and the provision of generic open APIs. This will allow the integration of the purchased services either in innovative workflows, or existing IT infrastructure. An example of the implementation of a successful exit strategy can be found in ARCHIVER report "Initial Data Management Plan" [24].

Data repatriation from commercial cloud to on-premise cloud should also be facilitated in the context of the EOSC. To this end, CERN developed a cloud data exporter [25] that repatriates data uploaded onto commercial cloud using OCRE cloud vouchers to Zenodo [26], an open-access repository developed under the European OpenAIRE program and operated by CERN. Data repatriation tools should be accessible to researchers in the EOSC. Furthermore, solutions must be available for users willing to archive their data in trustworthy scientific data repositories, which follow standard best practices and guidelines to ensure the findability, accessibility, interoperability, and reuse of the data. The innovative archiving and preservation solutions developed as part of the ARCHIVER project, offer such functionalities and will become part of the catalogue of the EOSC by December 2021. EOSC stakeholders should be able to easily move their data to these innovative and cost-effective solutions.

Finally, the compliance to the EOSC Rules of Participation must be validated before providers can offer their services in the EOSC catalogue. As proved in the HNSciCloud, ARCHIVER and OCRE projects, test-suites are powerful tools to validate that the cloud providers meet the minimum acceptance requirements and should be used in the EOSC. Tested requirements should be technical and organisational, to ensure the conformity with the current European legislation (such as the General Data Protection Regulation and the Free Flow of Data) and to safeguard the safe production, usage and sharing of scientific data across Europe.

5 Acknowledgements

The authors would like to thank the persons working on HNSciCloud, OCRE, ARCHIVER and the EOSC-hub projects. In addition, we would like to thank Emanuele Storti (Eurodoc [27]), Gareth O'Neill (Technopolis Group [28]) and Marco Masia (Marie Curie Alumni Association [29]) for their active involvement in the gathering of the Long-Tail-of-Science requirements for commodity cloud services.

References

1. European Open Science Cloud: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
2. High-Luminosity Large Hadron Collider project: https://hilumilhc.web.cern.ch/content/hl-lhc-project
3. S. Campana, *R&D session: Software and Computing Requirements and possible architecture*, ECFA High Luminosity LHC (2016)
4. EMBL on FIRE Use Case Technical Summary: https://archiver-project.eu/sites/default/files/Technical%20Summary%20%20-%20EMBL%20on%20FIRE.pdf
5. L. A. Heinrich, R. B. Da Rocha, *Reperforming a Nobel Prize discovery on Kubernetes*, CHEP (2019) (to be published)
6. Helix Nebula Science Cloud project: www.hnscicloud.eu
7. PCP process: https://ec.europa.eu/digital-single-market/en/pre-commercial-procurement
8. Procura+ Award 2019: https://procuraplus.org/awards/
9. Open Clouds for Research Environments project: www.ocre-project.eu
10. Archiving and Preservation for Research Environments project: www.archiver-project.eu
11. EOSC-hub project: www.eosc-hub.eu
12. J. Fernandes, B. Jones, M. Devouassoux, J. Tendel, *Research Infrastructures & NRENs Requirements for Commodity Cloud Services*, Zenodo, 4 (2020)
13. B. Jones, J. Fernandes, M. Devouassoux, *Voucher Schemes for Accessing Commercial Cloud Services in the Research Environment*, Zenodo (2019)
14. S. Andreozzi, H. Goodson, B. Jones, A. Bens, C. Veys, M. Williams, P. Matthews, D. Wustemberg. D. Mallmann, H. Koers, A. Giuliano, A. M. Pantea, *Procurement requirements and demand assessment*, (2019)

15. J. Fernandes, B. Jones, M. Devouassoux, J. Tendel, *Research Infrastructures & NRENs Requirements for Commodity Cloud Services*, Zenodo (2020)
16. M. Devouassoux, B. Jones, J. Fernandes, *Long-Tail-of-Science's Requirements for Commodity Cloud Services in Europe*, Zenodo (2019)
17. J. Fernandes, B. Jones, M. Devouassoux, J. Tendel, *Research Infrastructures & NRENs Requirements for Commodity Cloud Services*, Zenodo, 9 (2020)
18. I. Lozada, J. Urban, J. Fernandes, "EOSC-Testsuite" [software], available from https://github.com/cern-it-efp/EOSC-Testsuite [accessed 2020-06-25]
19. J. de la Mar, *D-PIL-3.13 TCO Study T-Systems*, Zenodo (2019)
20. RHEA Group, *D-PIL-3.13 TCO Study RHEA*, Zenodo (2019)
21. Directorate-General for Research and Innovation, *Prompting an EOSC in practice*, EU Publications, 30 (2018)
22. B. Jones, J. Fernandes, M. Devouassoux, *Voucher Schemes for Accessing Commercial Cloud Services in the Research Environment*, Zenodo, 9 (2019)
23. Switching and porting Codes of Conduct: https://swipo.eu/
24. J. Fernandes, B. Jones, M. Devouassoux, J. Shiers, D. Foster, M. Coelho dos Santos, *ARCHIVER D1.1 - Initial Data Management Plan*, Zenodo (2019)
25. I. Lozada, "cloud–exporter" [software], available from : https://github.com/cern-it-efp/cloud-exporter [accessed 2020-06-25]
26. Zenodo : https://zenodo.org/
27. Eurodoc : https://eurodoc-net.com/en
28. Technopolis Group : https://www.technopolis-group.com/
29. Marie Curie Alumni Association : https://www.mariecuriealumni.eu/