CoESRA: From virtual desktop to science gateway.

Hoang Anh Nguyen*, David Abramson, Research Computing Centre; Siddesware Guru, Yi Sun Terrestrial Ecosystem Research Network;
**Corresponding author address:* Research Computing Centre, The University of Queensland, Australia; email: uqhngu36@uq.edu.au

Abstract: *The Collaborative Environment for Ecosystem Science Research and Analysis (CoESRA) is a Web-based virtual desktop environment that integrates existing eResearch infrastructure in Australia for synthesis and analysis of scientific data for the ecological science community. Data synthesis and analysis is performed through scientific workflows. Even though this is powerful, it has a large learning curve for novice users. We have implemented a Web layer on top of an existing virtual desktop layer to hide this complexity from users. This new layer allows users to execute scientific workflows without requiring a desktop, and thus reduces the learning curve. The virtual desktop is still accessible for more advanced users.*

1. Introduction

Current scientific experiments increasingly require more distributed resources for computation and storage. In the Australian context, researchers typically use the National research platforms National eResearch Collaboration Tools and Resources (NeCTAR) [1] and the Research Data Storage (RDS) [2] to obtain compute and storage Cloud resource. However, both RDS and NeCTAR provision resources using an Infrastructure as a Service (IaaS) model. Scientists apply for resources, build and maintain a platform, and run experiments on that platform. This requires significant system administration skills, which many often domain scientists do not have.

CoESRA is a project initiated by Terrestrial Ecosystem Research Network (TERN) [3] to build an environment for synthesis and analysis of scientific data for ecosystem science community. Its main objective is to "*demonstrate the value in better connecting eResearch infrastructure by bringing ecosystem science date streams closer to a common environment for further analysis tasks, and by supporting data publishing of the results*" [4]. CoESRA, as a common environment, promotes: 1) use and reuse of existing analysis tools and techniques in the form of flexible pipelines - scientific workflows 2) a collaborative workspace for sharable ecology analysis.

CoESRA provides a virtual laboratory for ecology research that integrates existing eResearch infrastructure. Scientific workflows are used to chain analysis processes into pipelines and orchestrate the execution on integrated eResearch infrastructure. The Desktop-as-a-Service (DaaS) approach delivers scientific workflows to users. This DaaS platform allows scientists to access a ready-to-use Linux desktop environment through a Web browser, and this relieves scientists from various system administration tasks and lowers the barrier to access infrastructure for research purposes. CoESRA is available at https://portal.coesra.org.au and it currently serves a community of approximately 50 registered users.

Although serving scientific workflows on virtual desktops provides flexibility, users still need to access and navigate the desktop to access tools. We observe that a lack of familiarity in the operating environment is the main hurdle for novice users.

This paper presents the latest development in CoESRA to make the system more accessible and user friendly. Web services, and a Web layer, are developed on top of existing infrastructure that make it possible to launch workflow executions without requiring a virtual desktop. Since most scientists are familiar with Web interfaces, the learning curve is reduced. In addition, the virtual desktops are still accessible for advanced users.

2. CoESRA v.1: a virtual desktop platform

CoESRA virtual desktops are served from a SLURM [5] virtual cluster, which are built on a research cloud offered by NeCTAR and RDS. Each node of the cluster is bundled with the Kepler

Presented at Gateways 2016, San Diego Supercomputer Center, La Jolla, CA, November 2-3, 2016.
https://gateways2016.figshare.com/

scientific workflow management system [6], and other popular packages for ecological data analysis such as Rstudio [7], QGIS [8], etc. Nimrod/G [9] is used to distribute computational jobs to clusters and the Cloud.

Virtual desktop management in CoESRA is handled by Strudel-web [10]. This OAuth2 compliant framework allows users to access the virtual desktop via their Web browsers. Strudel-web is responsible for three main tasks: 1) authenticate users using Australian Access Federation [11] and Shibboleth 2) authorize users to launch a VNCServer as a SLURM job and 3) create a VNC tunnel from user's browsers to the VNCServer via Guacamole [12].

3. CoESRA v.2: a science gateway

In order to support the execution of a scientific workflow from the portal, the following aspects of a workflow execution need to be supported:

- *Data management*. This involves the management of workflow input and output data. NeCTAR's object store is used for all input/output data in CoESRA.
- *Resource management*. This involves the management of distributed computational resources.
- *Workflow management*. This involves the management of workflow contents, for instance creating, editing or sharing workflows.
- *Execution management*. This involves the management of workflow execution, i.e. preparation of the execution environment.

Fig.1 shows the architecture of CoESRA v.2. New Web services are developed on top of current Strudel-web services to utilize existing authentication and authorization. The Web components are developed as portlets conforming to JSR 286 and they are deployed in a Liferay portal.

CoESRA also captures essential information about each execution, including which workflow is being executed, with what distributed resource, parameters, and datasets and at which node of the virtual cluster. The whole execution process is illustrated in Fig. 2. Once the *Deployment Manager* receives an execution request from the

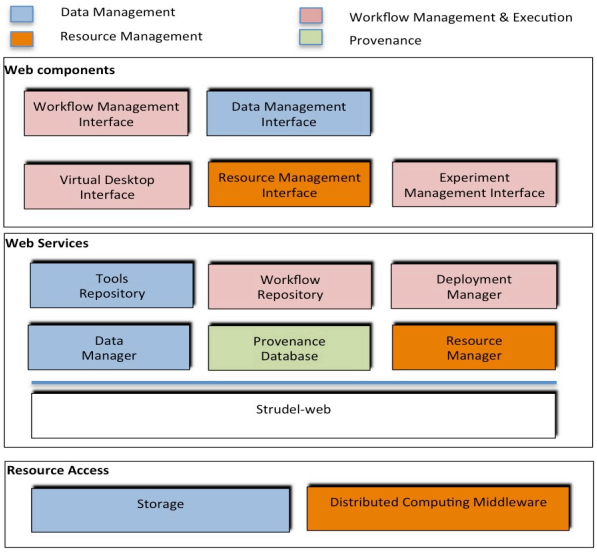

Fig. 1. CoESRA v.2 Architecture.

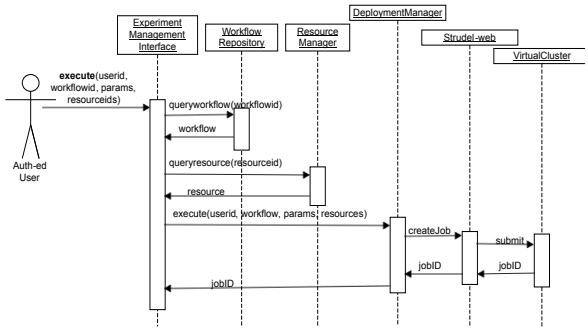

Fig. 2. Execution Process in CoESRA v.2

Experiment Management Interface, it then creates a SLURM job with those settings and submits to the cluster via Strudel-web. Virtual desktop requests are also handled via the *Deployment Manager*.

4. Conclusions

We have briefly presented the latest development in CoESRA, a platform for synthesis and analysis of scientific data for the ecological science community. The newly developed Web interface reduces the learning curve for novice users.

5. Acknowledgments

We thank Chris Hines and Jason Rigby for their contributions to the project. CoESRA is funded by UQ RCC, TERN and RDS. TERN and RDS are supported through the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS).

6. References

[1] "NeCTAR.," 2016. [Online]. Available: http://www.nectar.org.au. [Accessed: 01-Sep-2016].

[2] "Research Data Services.," 2016. [Online]. Available: https://www.rds.edu.au. [Accessed: 01-Sep-2016].

[3] S. Guru, X. Shen, C. Love, A. Treloar, S. Phinn, R. Wilkinson, C. Brady, and T. Clancy, "Sharing Australia's Nationally Significant Terrestrial Ecosystem Data: A Collaboration between TERN and ANDS," in *2013 IEEE 9th International Conference on eScience (eScience)*, 2013, p. pp.53,60.

[4] S. Guru, I. C. Hanigan, H. A. Nguyen, E. Burns, J. Stein, W. Blanchard, D. Lindenmayer, and T. Clancy, "Development of a cloud-based platform for reproducible science: A case study of an IUCN red list of ecosystems assessment," *Ecological Informatics*, 2016.

[5] A. B. Yoo, M. A. Jette, and M. Grondona, "Simple linux utility for resource management," in *Workshop on Job Scheduling Strategies for Parallel Processing*, 2003, pp. 44–60.

[6] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, and S. Mock, "Kepler : An Extensible System for Design and Execution of Scientific Workflows," in *16th International Conference on Scientific and Statistical Database Management*, 2004, pp. 423–424.

[7] J. Verzani, *Getting started with RStudio*. O'Reilly Media, Inc., 2011.

[8] A. Graser, *Learning QGIS 2.0*. Packt Publishing Ltd, 2013.

[9] D. Abramson, R. Sosic, J. Giddy, and B. Hall, "Nimrod : A Tool for Performing Parametised Simulations using Distributed Workstations," in *Proceedings of the Fourth IEEE International Symposium on High Performance Distributed Computing.*, 1995, no. August, pp. 112–121.

[10] "Strudel-web," 2016. [Online]. Available: https://www.massive.org.au/userguide/cluster-instructions/strudel-web. [Accessed: 01-Sep-2016].

[11] "Australian Access Federation," 2016. [Online]. Available: http://aaf.edu.au. [Accessed: 30-Aug-2016].

[12] "Guacamole," 2016. [Online]. Available: https://guacamole.incubator.apache.org/. [Accessed: 01-Sep-2016].