Leveraging Distributed Networked Cloud Testbeds for Domain Science Research and Experimentation

Anirban Mandal[1a]*, Eric Lyons*[b]*, George Papadimitriou*[c]*, Cong Wang*[a]*, Komal Thareja*[a]*,*
Paul Ruth[a]*, Ilya Baldin*[a]*, Michael Zink*[b]*, Ewa Deelman*[c]
[a]Renaissance Computing Institute (RENCI), UNC - Chapel Hill
[b]University of Massachusetts, Amherst
[c]USC Information Sciences Institute

Computational science today depends on complex, data-intensive applications operating on datasets from a variety of scientific instruments. A major challenge is the integration of data into the scientist's workflow. Recent advances in dynamic, networked cloud resources provide the building blocks to construct reconfigurable, end-to-end infrastructure that can increase scientific productivity. However, applications have not adequately taken advantage of these advanced capabilities. In the context of the DyNamo [4] project funded under the NSF Campus CyberInfrastructure program, we have developed a novel network-centric platform, Mobius [7], which enables high-performance, adaptive data flows and coordinated access to distributed multi-cloud resources (cloud research testbeds like ExoGENI [1], Chameleon [2], XSEDE JetStream [3], etc.), and data repositories for atmospheric scientists.

We have demonstrated the effectiveness of our approach by evaluating time-critical, adaptive weather sensing workflows, which utilize advanced networked infrastructure to ingest live weather data from radars and compute data products on dynamically provisioned resources on hybrid, multi-cloud platforms, which are used for timely response to weather events. The workflows are orchestrated by the Pegasus workflow management system [6] and were chosen because of their diverse resource requirements. We have shown that our approach results in timely processing of CASA [5] weather workflows under different infrastructure configurations and network conditions. We have also shown how workflow task clustering choices affect throughput of an ensemble of workflows with improved turnaround times. Our findings show that using our network-centric platform powered by advanced layer2 networking techniques results in faster, more reliable data throughput, makes multi-cloud resources easier to provision, and the workflows easier to configure for operational use and automation.

We are currently extending our work such that domain science data flows can be effectively adapted and optimized by leveraging Software-Defined Exchanges (SDX), and the Quality of Service of the end-to-end provisioned infrastructure can be transparently maintained by active monitoring and control. Our current plans also include supporting a wider federation of cloud infrastructure (public clouds like Amazon EC2 and other research clouds like Massachusetts Open Cloud) using Mobius. Using the connected, distributed, multi-cloud federation enabled by DyNamo, we are continuing to support (a) a wider range of adaptive weather sensing workflows performing wind computations and hail formation, and (b) ingest of streaming data and on-demand computations for workflows employing data from the Ocean Observatory Initiative (OOI) NSF Large Facility.

Another novel use of distributed cloud testbeds is for studying the integrity and reproducibility of domain science applications when faced with scientific data integrity issues resulting from both intentional and unintentional sources. Testbeds are ideal to simulate different kinds of attack vectors that might be exhibited in a production cyberinfrastructure. In our recent work on the SWIP project [8], we developed a suite of software called the “Chaos Jungle” [9], which can be deployed on the ExoGENI testbed to inject different

[1] Corresponding author, Email: anirban@renci.org

kinds of infrastructure anomalies, including those related to integrity errors. One example use was when Chaos Jungle intentionally corrupted the scientific data from a bio-science workflow by mangling the network packets without affecting the checksum, the integrity check features introduced in the workflow system orchestrating the application were able to suitably catch those integrity errors. We could test these integrity checking capabilities by leveraging distributed cloud testbed infrastructures.

Hence, our position is that global, distributed, networked cloud testbeds are rapidly becoming virtual labs for the development of research infrastructures for domain science evaluation and validation, and for experimenting with novel algorithms, models and data management approaches for science applications.

References

[1] I. Baldin, J. Chase, Y. Xin, A. Mandal, P. Ruth, C. Castillo, V. Orlikowski, C. Heermann, J. Mills. “ExoGENI: A Multi-Domain Infrastructure-as-a-Service Testbed.” The GENI Book, pp. 279--315, 2016.

[2] NSF Chameleon Cloud. https://chameleoncloud.org/

[3] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. Scott, N. Wilkins-Diehr, Xsede: Accelerating scientific discovery, Computing in Science & Engineering 16 (05) (2014) 62–74. doi:10.1109/MCSE. 2014.80.

[4] E. Lyons, G. Papadimitriou, C. Wang, K. Thareja, P. Ruth, J. J. Villalobos, I. Rodero, E. Deelman, M. Zink and A. Mandal, Toward a Dynamic Network-centric Distributed Cloud Platform for Scientific Workflows: A Case Study for Adaptive Weather Sensing, IEEE eScience 2019, San Diego, CA, September 2019.

[5] B. Philips, D. Pepyne, D. Westbrook, E. Bass, J. Brotzge, W. Diaz, K. Kloesel, J. Kurose, D. McLaughlin, H. Rodriguez, and M. Zink. “Integrating End User Needs into System Design and Operation: The Center for Collaborative Adaptive Sensing of the Atmosphere (CASA).” In *Proceedings of Applied Climatol., American Meteorological Society Annual Meeting*, San Antonio, TX, USA, 2007.

[6] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, “Pegasus: a workflow management system for science automation.” *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[7] Mobius. https://github.com/RENCI-NRIG/Mobius

[8] Mats Rynge, Karan Vahi, Ewa Deelman, Anirban Mandal, Ilya Baldin, Omkar Bhide, Randy Heiland, Von Welch, Raquel Hill, William L. Poehlman, and F. Alex Feltus, “Integrity Protection for Scientific Workflow Data: Motivation and Initial Experiences.” In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (PEARC '19)*. ACM, New York, NY, USA, Article 17, 8 pages. 2019. DOI: https://doi.org/10.1145/3332186.3332222

[9] Chaos Jungle, https://github.com/RENCI-NRIG/chaos-jungle