NSDF-Cloud: Enabling Ad-Hoc Compute Clusters Across Academic and Commercial Clouds

Jakob Luettgau
Paula Olaya
Naweiluo Zhou
University of Tennessee
Knoxville, Tennessee, USA

Giorgio Scorzelli
Valerio Pascucci
University of Utah
Salt Lake City, Utah, USA

Michela Taufer
University of Tennessee
Knoxville, Tennessee, USA

ABSTRACT

Computational resources are increasingly provisioned to users through cloud-like interfaces. Both academic and commercial cloud offerings exist, but no single standardized interface for common actions such as configuration, launching, and termination of virtual resources exists. This imposes huge technical burden on domain scientist that attempt to take advantage of these resources; even expert users spend considerable time to port their applications from one cloud platform to another.

With this work, we make available to the community a unified API toolkit as well as five in-depth reports on challenges we encountered working with different academic and commercial cloud providers. Our toolkit implements automations for common tasks such as simultaneous launching and termination of large numbers of virtual machines (VM) across the cloud. We demonstrate that our toolkit brings down the time users need to spend launching and terminating these resources to mere minutes, thus enabling ad-hoc multi-cloud clusters.

KEYWORDS

VIRTUAL MACHINE, CYBERINFRASTRUCTURE, DATA FABRIC

ACM Reference Format:
Jakob Luettgau, Paula Olaya, Naweiluo Zhou, Giorgio Scorzelli, Valerio Pascucci, and Michela Taufer. 2022. NSDF-Cloud: Enabling Ad-Hoc Compute Clusters Across Academic and Commercial Clouds. In *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing (HPDC '22), June 27–July 1, 2022, Minneapolis, MN, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3502181.3533710

1 INTRODUCTION

As compute capabilities are increasingly provisioned through cloud-like interfaces by both academic and commercial providers, the diversity of APIs to launch, control, and terminate resources becomes a challenge. As part of the National Science Data Fabric (NSDF) pilot initiative, we are developing tools and methodologies to bridge the heterogeneous landscape of cloud providers while also embracing best practices to improve reproducibility and accessibility to research conducted on national cyberinfrastructure (CI) as well as on commercial clouds [**?**].

HPDC '22, June 27–July 1, 2022, Minneapolis, MN, USA.

ACM ISBN 978-1-4503-9199-3/22/06.
https://doi.org/10.1145/3502181.3533710

We identify the familiarity with the different vendor application programming interfaces (APIs), the different vocabulary, and the sometimes dramatically different sequences of steps necessary to launch virtual machines (VMs) as major sources of friction in the adoption for cloud resources for domain scientists. Even expert users have to invest significant effort to port and a cloud-based cluster from one provider to another.

In this work, we make available an extendable toolkit to spawn resources across both academic and commercial cloud providers. Furthermore, we report our experiences and challenges as part of five case studies for each of the vendors we currently integrate into our toolkit. As such, our contributions are:

- An in-depth analysis and documentation to overcome common challenges for five different academic and commercial cloud providers available on GitHub;
- A unified API that automates three common tasks such as the simultaneous creation, state listing, and tear-down of multiple VMs as well as the collection of public IPs and access credentials required by higher-level orchestrators such as Ansible and Dask [2]; and
- A quantitative analysis of create and delete latencies when requesting varying numbers of VMs for the five providers.

2 CHALLENGES ACROSS PROVIDERS

In this section we report our experience and highlight challenges that we encountered when using five different academic and commercial cloud providers. Table 1 presents the tested cloud providers listing five key characteristics. When testing the five cloud platforms we gathered the following observations: (Ob.1) Accessibility and ease-of-use are central concerns which all the providers address through different means. All offer newcomers friendly web dashboards both to monitor and control resources. (Ob.2) Most of the vendors introduce their own vocabulary for different services, as such a common challenge is to identify an equivalent services when moving from one cloud to another. (Ob.3) Most provider offer command-line interface (CLI) tools or API access in one form or another **but no standard for Identity and Access Management (IAM) has emerged yet**, requiring special procedures for each provider. (Ob.4) **Providers have their own unique sequence of steps** such as setting up security groups or networks **when VMs are launched**. (Ob.5) Gathering the **public IPs** and injecting **SSH access credentials require significant customization for each provider**. (Ob.6) Academic clouds develop high-level abstractions and often offer pathways to leverage the underlying infrastructure when asking directly; on the other hand, commercial providers tend to develop their own stack and hide details of the underlying

Table 1: Characteristics of the different cloud service providers, both academic and commercial.

Provider	Type	Credentials	Multi-Region	Stack	Custom Images
AWS	Commercial	Token+Secret	Yes (Int.)	Custom	Yes
Chameleon	Academic	Token	Yes (US)	CHI on OpenStack	Yes*
CloudLab	Academic	Certificate	Yes (US)	Custom	Yes
Vultr	Commercial	Token+IP-Whitelist	Yes (Int.)	Custom	Yes
JetStream	Academic	Token	Yes (US)	Atmosphere on OpenStack	No*

open source infrastructure. (Ob.7) Most providers feature multiple regions either globally or in the US. (Ob.8) Some academic clouds enforce a lease-based model that requires users to make reservations before launching resources; on the other hand, commercial clouds usually offer on-demand services.

3 A UNIFIED API TO CREATE AND DELETE VIRTUAL MACHINES ACROSS PROVIDERS

From the observations listed above, we identify three key challenges that are especially hard to overcome by researchers who are new and want to get starting with the use of cloud platforms for their scientific computations. The researchers often spend countless hours to execute specific deployment adaptations and technical tasks for: 1) *Setting up, collecting, and distributing credentials and security settings*; 2) *Launching VMs and gathering basic information such as public addresses to hand over to higher-level orchestrators such as, for example, Ansible or Dask*; and 3) *Terminating resources once computations or experiments are done.*

NSDF-Cloud is out solution to these challenges. NSDF-Cloud is an extendable Python-based toolkit that unifies the different cloud APIs to automate common cloud tasks. Speficially, NSDF-Cloud provides three easy to use commands: `create nodes <prefix> –num <N>`, `get nodes <prefix>`, and `delete nodes <prefix>`. Furthermore, users can configure and register their credential for different providers using a `vault.yaml` which is honored by the NSDF-Cloud CLI utility. Our toolkit can be used both from in the CLI and in Python-scripts. After successfully launching nodes, our toolkit automatically generates Ansible inventory files.

4 DEMONSTRATING NSDF-CLOUD POTENTIALS

We demonstrate the potentials of our toolkit by launching varying numbers of *VMs* $\{1, ..., 16\}$ with each of the different providers and measuring the latency to complete different *actionss* {create, delete}. For *create* we measure the time until logging onto the nodes using SSH succeeds. An overview of the results faceted by the amount of VMs requested is plotted in Figure 1.

The task of launching between one and 16 VMs usually completes within 10 minutes. The amount of resources requested in many cases does not have an effect on the overall launch latency in our experiments. The two commercial providers AWS and Vultr often launch resources more reliably than their academic counterparts.

The resources to perform our study where in part donated by vendors and part acquired through smaller resource allocation grants directly through the provider as well as XSEDEs XRAS process. Resources to perform experiments on AWS cloud were generously provided by CloudBank.

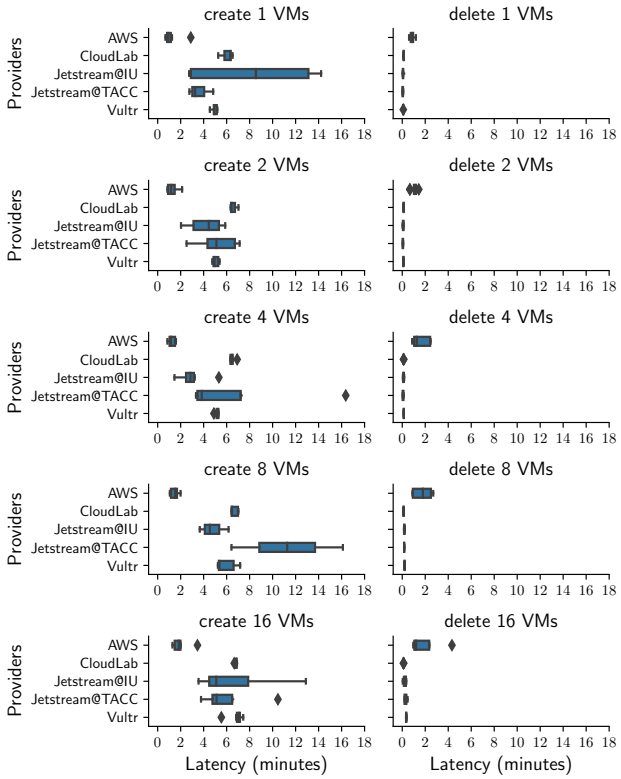

Figure 1: Latency in seconds to complete the *create* and *delete* commands using the NSDF-Cloud CLI tools.

5 CONCLUSION

In this work, we first present important observations when using five cloud providers, both academic and commercial. We leverage the lessons learned to implement the NSDF-Cloud toolkit which enables spawning resources across different cloud providers thus supporting users to easily create ad-hoc compute clusters across different clouds with higher-level orchestrators.

ACKNOWLEDGMENTS

Support from the National Science Foundation under ACI #2138811 and #1548562 is acknowledged; AWS access is from Cloudlab.

REFERENCES

[1]]nsdf [n. d.]. National Science Data Fabric: A Platform Agnostic Testbed for Democratizing Data Delivery. http://nationalsciencedatafabric.org/. [Online; accessed 04-03-2022].

[2] Matthew Rocklin. 2015. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*. Citeseer.