Scientific Research in the Cloud

A Partly-Cloudy Perspective on Current Challenges

Contents

Overview

This paper captures insights developed during Partly-Cloudy 2019 and attempts to answer the following questions:

1. What is slowing down our adoption of the Cloud?
2. What changes to Cloud services would increase our adoption?

Partly-Cloudy consists of intermittent conversations amongst IT Infrastructure staff who support scientific research activities, punctuated by a one-day meeting in October, hosted at Fred Hutch in Seattle (60 attendees). Historically, we attract attendees from the university and independent biomed research communities, with a smattering from pharma, plus vendors who provide services to this community, including the Cloud providers themselves.

Summary

The bulk of this paper addresses Cost Control; this issue dominates the attention of outfits operating under a fixed income business model and tends to be of less concern to .com consumers of Cloud services (e.g. pharma).

DevOps in the Cloud has broken traditional procurement. Procurement has outsourced their job to engineers. Engineers now spend company money at will and make financial decisions on cloud providers like AWS, GCP and Azure at rapid speed.[1]

What is slowing down our use of the Cloud?

Cost Control

The current mechanisms for placing a hard-lid on expenditure do not meet our needs.

Cost Monitoring & Charge Back

Figuring out how to watch costs, to forecast costs, and to respond to unexpected cost bumps is complex and involves staff-intensive, manual process. Hiring the people, building the processes, and developing the tools are taking time.

Staff Retraining

Researchers tend to lack formal training and/or experience in formal software development practices; they possess niche expertise in Unix/shell scripting or Python, but have little incentive or resources to develop a more robust skillset. Migrating to a more professional approach to development, using Cloud-native paradigms, requires strategic and cohesive communication of

[1] https://www.finops.org/blog/why-finops-foundation/

best practices that are actually workable for common research use case. See Community Development (a community of Cloud-native learners supporting each other) and Workflow Managers[2] as contributors to the solution space.

Cybersecurity
Given our mission to make our research output freely available, for the benefit of the national economy, security has historically been challenging for us. Employing the Cloud adds yet another dimension, to which we are struggling to adapt.

What changes to Cloud Services would increase our adoption?

We propose that Cloud Providers sell Cloud Credits directly to institutions.

> We envision a process in which institutions would leverage their existing purchasing processes to refresh credits tied to budgeting objects within the Cloud providers hierarchy of account / project definitions. When a given budgeting object exhausts its credits, the Cloud provider would shut down all ensuing cost-incurring activity, until and unless the institution added more Credits.

Notes
The existing reseller approach to providing Cloud Credits places onerous limitations on us Currently, the reseller approach -- we hit the defined threshold, they call us asking either for more money or agreement to disable the account -- remains manual and thus not scalable. Further, the reseller approach does not integrate with the occasional flow of Cloud Credits directly from the Cloud Provider (provided as part of recompense for a bug, a high-level deal struck with leadership, or various promotional arrangements). Finally, the addition of a 3rd party requires that we tune our purchasing system to yet another likely transient partner -- we would rather tune our purchasing system to work with a longer-lasting provider (AWS, Azure, GCP …)

Insights

Fixed income activities are a poor fit for the Cloud's elastic costs

> The Cloud's business model -- the more you use, the more you pay -- fits well with business models for which increased utilization translates into increased revenue. The holiday season approaches, and your Cloud bill skyrockets … but your revenue skyrockets as well. The day after the holiday, your income plunges … but so does your Cloud bill. Similarly with SaaS -- the more customers who sign up for your solution, the

[2] Arguably, Workflow Managers are an all-around good thing, regardless of on-prem vs Cloud considerations.

higher your Cloud bill goes … but then again, the larger your accompanying revenue stream.

The Cloud's elastic costs pose a significant threat to the viability of fixed income tasks

However, for fixed income activities, Cloud utilization bears no relationship to income. When the size of your budget is fixed ahead of time, no change in the volume of Cloud utilization will change that budget. For example, the scientific insights you generate from your work, and your use of Cloud resources, may perhaps produce a Nature or Science paper … but your grant size remains fixed, no matter how impactful your discovery.

Predicting the impact of research on Cloud resources is hard

By their nature, the exploratory aspect of many scientific research activities make predicting their impact on Cloud utilization difficult. Historically, impactful discoveries were accidental byproducts of some other line of inquiry -- swerving to follow anomalies is standard operating procedure during this activity we call *scientific research*. Plenty of research employs IT resources in a long series of one-off efforts -- writing code to analyze a data set, gradually expanding / changing / morphing that code with each run It is this long series of one-off runs of custom code which make predicting the impact of research on IT resource consumption difficult

Cost Control: Institutions deploy substantial effort to put a hard ceiling on expenditures

Institutional research administration has a well-developed tool kit for keeping a lid on spending, i.e. to prevent a researcher from overrunning a grant. For example, if a researcher has 100K to spend on a particular grant, the hosting institution has an array of processes, procedures, and staff who spend considerable effort making sure that those dollars are spent according to Federal accounting guidelines … and … stopping all spending when that 100K ceiling is reached. Consider when a researcher wants to purchase a server: that purchase request travels an approval process which checks to see if the grant is (a) authorized to purchase a server (per specfics agreed upon with the funding agency when the grant was awarded), and (b) still contains enough funds to pay the resulting invoice. Only once those checks have completed does the institution ship a PO to the supplier.

In addition, historically, researchers adapt their science (consciously or not) to fit within the limitations of on-prem IT resources -- to fit within the limitations of the heavy Compute / Network / Storage which their institution provides. When they swamp the HPC cluster with jobs, consume all available Network throughput, and fill up Storage, their local IT staff may walk down the hall, tap them on the shoulder, and whine … but the institution, and their grant, takes no additional hit in terms of cost.

Pay After Use

By contrast, current use of the Cloud mostly involves consuming the resources (Compute / Network / Storage / Lambdas), and then getting billed after the fact for said consumption. This is akin to allowing a researcher to buy a server and only then getting around to figuring out whether or not they have the money left in their grant to afford it. The average non-profit research institution cannot afford to make such mistakes -- they have few *discretionary funds* available to cover up errors like this -- ergo the intense effort they invest up front to prevent such events from occurring.

The Past / The Future

In the beginning, when research began leveraging IT, the real-time cost accounting capabilities of the mainframe provided hard lids to expenditures.

With the advent of supercomputing, National Science Foundation-funded centers added a manual *compute-ready* review process, in which compute-savvy researchers reviewed algorithms and even code, proposed by grantees requesting time on these large machines. The review committees will reject (typically with suggestions) approaches which make inefficient use of machine resources.

With the advent of the microcomputer, the researcher's costs were bounded by the PCs purchased on their grant -- they could swamp their labs' computers without incurring additional cost. Similarly with the shared farms of on-prem Compute / Network / Storage which dominate the non-profit research sector today.

Looking at the Cloud, we no longer have a hard-lid on costs -- this gap drives the sophisticated cost control infrastructure which we see being developed at institutions which are leading the way in Cloud adoption. However, the leaders in this space are uniformly institutions whose leadership is investing scarce non-direct dollars into their Cloud bills and Cloud support. This is not an option available to most of us. And in any case, not sustainable long-term for any of us. As a result, Cost Control slows us down, in terms of Cloud adoption.

Looking to the future, we predict that effective use of the Cloud by fixed income research, particularly exploratory research, will be accompanied by the creation of cost-effective real-time or near-real-time cost control, likely using some combination of past techniques:

- Research institutions creating a virtual data center or mainframe, through which jobs from that institution flow, to be executed on-prem or in the Cloud, but in either case, implementing hard lids on costs

- The Cloud providers selling Cloud Credits
- Institutional review committees for particularly large or risky jobs, assessing the likelihood of runaway charges allocated to the hosting institution

Errata

This is hard

We recognize that placing ceilings on Cloud costs is hard -- the Cloud is a distributed system, distributed systems tend to exhibit entrancing scaling & resiliency properties … but distributed systems also tend to be hard to manage (and hard to trouble-shoot). The fact that the Cloud offers elastic resource consumption is an example of its scalability; the fact that Cloud providers do not currently offer hard lids on costs is an example of its manageability challenge.

Subtle Factors

Cost

The way the research community deploys on-prem resources offers some cost advantages over Cloud use.

- We tend to adopt noticeable risk in terms of Data Loss / Data Unavailable (aka losing data and service disruption)
    - Limited data protection
    - Limited physical security
    - Limited cybersecurity
    - Few staff (single points of failure, in terms of expertise)
    - Aging hardware / software
- By contrast, the Cloud providers invest a lot of effort (and cost) into reducing the chance that they will lose our data, disrupt service, and so on
- These factors allow us to operate on-prem systems with lower costs than one might otherwise predict

Flexibility

Furthermore, we tend to run systems which host ancient operating systems / libraries / in-house code. By contrast, Cloud providers bound you on just how old you get -- they retire support for services sooner than on-prem IT shops do.

In addition, research on-prem tends to re-use hardware purchased on one grant to support additional activities, whether explicitly funded or not. With its pay-as-you-go model, such re-purposing is not possible in the Cloud (well, not possible without paying additional monthly fees).

Hidden Subsidies

On-prem systems tend to benefit from hidden subsidies -- the cost of rent/mortgage, power, IT staff which may or may not be explicitly charged to a particular grant budget, the way a Cloud bill might be. Similarly, at least today, Cloud systems tend to require substantial investment in Cost Control / Chargeback systems, involving additional accounting staff and budget managers whose salaries may or may not be charged to a grant.

Future

Trajectory

Participants suggest the following trajectory for Research Computing

- Because of agility, scale, and cost, the future will see most Research Computing hosted in the Cloud. However, we are unclear how long we will spend reaching this future (5? 10? 20?)
- We anticipate a long tail of on-prem usage, driven by leveraging sunk assets (e.g. existing Data Centers), legacy issues (e.g. POSIX-compliant applications), and hidden subsidies (e.g. Facilities pays the power bill, grant recipients do not)
- That being said, we acknowledge that some situations and use cases may sustain or even expand on-prem systems for the foreseeable future. For example:
    - Access to low-cost power makes running on-prem less expensive than is typical world-wide
    - Substantial on-prem data generation: historically, Networks have lagged Compute & Storage, with the result being that transferring petabytes of data remains slow and expensive. If you generate a lot of bits on-prem, you may find it cheaper and faster to Store and Compute over those bits on-prem as well
    - Exploratory research on a fixed income: some of us propose that Cloud Credits solve this problem for most use cases; others suggest that there will remain substantial use cases where the simplicity of the hard-cost lids provided by on-prem gear will remain compelling
    - Hidden, or not-so-hidden, subsidies will continue to play a role in the Federally-funded space: Federal policy deliberately accepts some inefficiency (e.g.

Facilities paying the electric power while grant recipients do not, through the indirect / direct cost accounting methodology) in exchange for promoting researcher independence. Want more? See http://www.skendric.com/seminar/business/Business-Model-Grant-Funded-Research-Institutes.pdf

Partly-Cloudy 2020

Tentatively, Fred Hutch intends to host Partly-Cloudy 2020 as follows
2019-10-15: Hands-on training in Python & Bash
2019-10-16: Partly-Cloudy 2020 Conference

Possible Themes for Partly-Cloudy 2020

- Cost Control
- Collaboration across Clouds: You and your collaborators at other institutions use different Clouds – how do you share data & applications?
- Send your suggestion to partlycloudy {at} fredhutch {dot} org

Exploration

- For a community concerned with cybersecurity and the Cloud, consider https://www.rhedcloud.org
- For a community concerned with controlling Cloud costs, see https://www.finops.org
- Unanswered question: what are the particular financial impacts to grant-funded institutions as IT CapEx spend (buying on-prem Compute / Network / Storage) migrates to IT OpEx spend (Cloud bills)? This migration changes both the numerator and the denominator in the N_i / N_d indirect rate calculation
- Unanswered question: what kind of support or requirements are Federal funding agencies providing around Cloud costs?

Appreciation

To the Partly-Cloudy community who, through their vigorous participation in break-out groups, plenary sessions, and email exchanges, developed the insights presented here.

Allen Institute
Amazon (AWS)
Benaroya Research Institute
Cancer Research & Biostatistics

Celgene
Diamond Age
Fred Hutch
Genentech
Igneous
Invitae
Institute for Health Metrics and Evaluation
Institute for Systems Biology
Jackson Laboratory
Lifebit
New York Genome Center
Pacific Northwest National Laboratory
Pennsylvania State University
Salk Institute
Scripps Research
Seattle Genetics
University of British Columbia
University of California
University of Central Florida
University of Virginia
University of Washington
Washington State University
WestGrid

And to correspondents at research institutes who contributed insights.
CERN
Cornell University
Jet Propulsion Laboratory
NOAA

Authors: Partly-Cloudy Steering Committee
Stuart Kendrick, Dirk Petersen, Robert J. Robbins