

Cloud Computing – getting started

Malachi Griffith

Bioinformatics Workshop
April 17th, 2023

Why go to the cloud?

- Performance & reliability
- Access to additional storage and compute as needed
- Greater control over compute environment
 - Ability to create persistent web services
- Data sharing needs
- Access to technologies not available on local WashU compute
- It's a valuable skill? It's the future?

Why not go to the cloud?

- Cost
- Complexity. Its another thing to learn
- Security?

Public Cloud & in-house compute

- Using the cloud is not the same as using the compute0/1 cluster
 - Not HPC-oriented (no LSF, à la bsub; no NFS-based file system)
 - Essence of scientific cloud computing:
 - i. Create a VM
 - ii. Download Input Data onto VM from Cloud Storage
 - iii. Download software onto VM
 - iv. Run computation(s)
 - v. Upload Output Data on VM to Cloud Storage
 - vi. Delete VM
- “research” on the cloud vs. local in-house compute
 - Exploratory compute on the cloud can get expensive very quick.
 - Try to do small-scale “research” on in-house compute, and then move to the cloud for large-scale “production” tasks
 - Use the cloud for functionality not supported by local compute
 - Use the cloud as a backup when local system is unstable

Getting started - Overview of intial steps

1. **Initial setup.** Request an AWS or GCP cloud account from WUIT
2. **POs.** Request a purchase order from your department and link to the account and encumber with appropriate sources of funding
3. **Login.** Test login and user permissions to access billing information and other necessary functionality. Request additional permissions from WUIT as needed.
4. **Users.** Request addition of users from the lab who need access
5. **Invoices.** Set up routing of DLT (AWS) or Burwood (GCP) bills to your email
6. **Alerts.** Set up billing dashboards and monthly billing alerts.
7. **Cost control.** Consider cost optimizations such as reserved instances

Initial setup with WASHU IT

- WASHU IT Cloud Computing Service
 - <https://it.wustl.edu/services/cloud-computing/>
- Benefits of using WASHU IT
 - University negotiated, policy compliant contracts with select public cloud vendors
 - Account Management
 - University Billing
 - Access to cloud consoles via WUSTL Key login
 - HIPAA BAA (select vendors)
 - Service Discounts
- Public Cloud Providers
 - <https://it.wustl.edu/services/cloud-computing/cloud-computing-services-comparison/>
 - Google Cloud Platform (GCP)
 - i. <https://it.wustl.edu/services/cloud-computing/google-cloud-platform/>
 - Amazon Web Services (AWS)
 - i. <https://it.wustl.edu/services/cloud-computing/amazon-web-services/>
 - Microsoft Azure
 - i. <https://it.wustl.edu/services/cloud-computing/microsoft-azure/>
- Contacts
 - John Bailey (jwbailey@wustl.edu) and Steven Clement (csteven@wustl.edu)
 - Please submit all initial requests for cloud accounts through the above web forms

Project initiation – request a new GCP “project” or AWS “account”

Google Cloud



Google Cloud (formerly GCP) is a newer public cloud platform geared toward research computing and storage. For information on Google Cloud, reference the following Google web pages:

- [Google Cloud Service Catalog](#)
- [Google Cloud Pricing Calculator](#)
- [Google Cloud Service Status](#)
- [Google Cloud Technical Documentation](#)

Ready to get started with Google Cloud?

1. Review the [Cloud Computing: Features & Options](#) service description to understand what is included and excluded
2. [Request a new WashU Google Cloud project](#) ←

Already have a WashU GCP project?

- [Login to the Google Cloud console using your WUSTL Key](#)
- [Contact us to request access to an existing WashU Google Cloud project](#)

Amazon Web Services



Amazon Web Services (AWS) is the industry leader in public cloud. AWS provides a range of services including compute, storage, and databases. For information on AWS, reference the following Amazon web pages:

- [AWS Service Catalog](#)
- [AWS Product Availability by Region](#)
- [AWS Pricing Calculator](#)
- [AWS Service Status](#)
- [AWS Technical Documentation](#)
- [AWS Service Limits](#)

Ready to get started with AWS?

1. Review the [Cloud Computing: Features & Options](#) to understand what is included and excluded
2. [Request a new WashU AWS account](#) ←

Already have a WashU AWS account?

- [Login to the AWS Portal using your WUSTL Key](#)
- [Contact us to request access to an existing WashU Azure subscription](#)

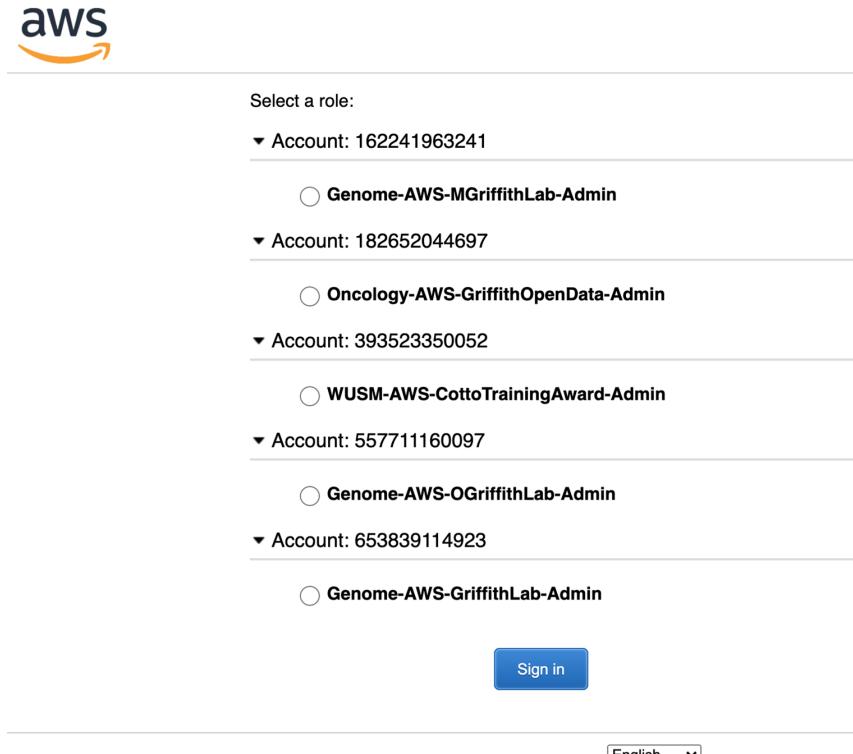
Billing, POs, encumbering ...

- Billing contacts within each department open a standing/blanket Purchase Order (POs) with each provider or reseller
 - Billing contacts vary by Department:
- You may establish multiple WASHU POs and link these to particular cloud accounts or you can just use one for Google/Burwood or AWS/DLT
- You will be asked periodically to encumber the PO for a certain amount based on how much you project to spend over the next 6-12 months
 - At this point you decide how much will be sourced to which fund numbers

Relationship with WASHU Resellers

- For billing and legal purposes WASHU has established business agreements with resellers
 - DLT (AWS)
 - Burwood Group (GCP)
- Interaction with these resellers can be quite limited
 - Does add some complexity to billing because what you see in the cloud billing console is not necessarily what you will be charged by DLT or Burwood (may be less or more)

Logging in to AWS (<http://connect.wustl.edu/awsconsole>)



The image shows the AWS login interface. At the top left is the AWS logo. Below it, a section titled "Select a role:" lists four accounts, each with a dropdown menu and a list of roles:

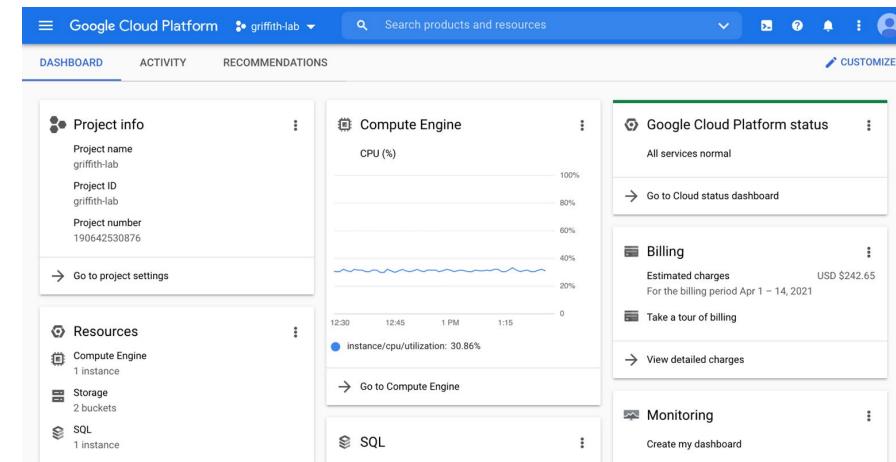
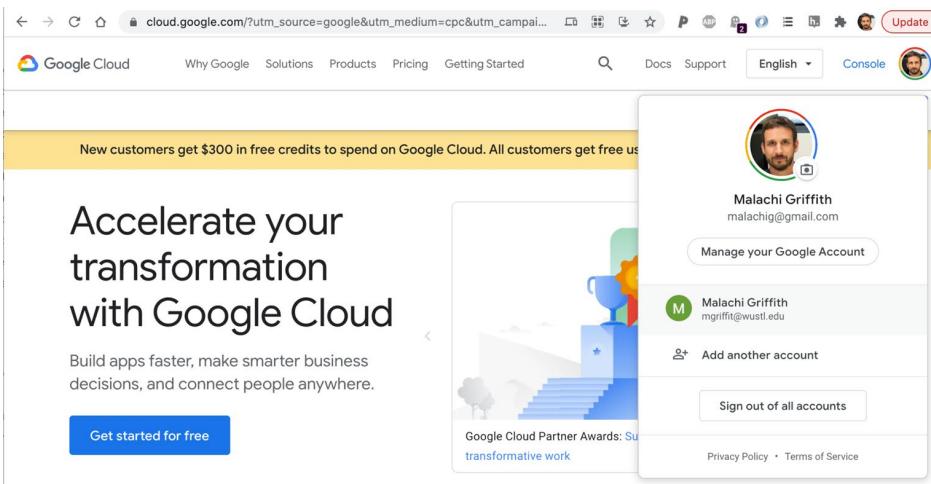
- Account: 162241963241
 - Genome-AWS-MGriffithLab-Admin
- Account: 182652044697
 - Oncology-AWS-GriffithOpenData-Admin
- Account: 393523350052
 - WUSM-AWS-CottoTrainingAward-Admin
- Account: 557711160097
 - Genome-AWS-OGriffithLab-Admin
- Account: 653839114923
 - Genome-AWS-GriffithLab-Admin

At the bottom center is a blue "Sign in" button. Below the button is a language selection dropdown set to "English". At the very bottom, there are links for "Terms of Use" and "Privacy Policy".

- Authentication uses your WUSTL Key
- You can have multiple accounts (optional)
 - Per project
 - Per billing source
- Or you can have one account and manage billing by paying one bill from multiple sources
- Multiple accounts can also be used to compartmentalize access to members of the lab working on different projects
- Each account will generate distinct billing, billing alerts, etc.

Logging into GCP (<https://cloud.google.com/>)

- Request a WASHU Google account
 - A very limited functionality Google account associated with your WASHU login
 - You now have two Google accounts. If you use the Google ecosystem a lot, this will be annoying.
 - Better functionality would happen with [Google Workspace \(formerly known as “G Suite”\)](#) integration; however, WashU IT is very invested with Microsoft and isn’t motivated yet to do this.



Adding users from a lab to a cloud account

- For AWS, the ability to add users from the lab to access AWS accounts is managed through WUIT by email
 - No way for PI to see what users are currently authorized without asking WUIT
- For GCP, the ability to add users is also limited to WUIT but at least it is possible to view the current configuration
 - You can get permission to add users yourself

The screenshot shows the Google Cloud Platform IAM & Admin interface for the project 'griffith-lab'. The left sidebar lists various IAM-related services: IAM, Identity & Organization, Policy Troubleshooter, Policy Analyzer, Organization Policies, Service Accounts, Labels, Tags, Settings, Privacy & Security, Identity-Aware Proxy, Roles, Audit Logs, Essential Contacts, Groups, Early Access Center, and Quotas. The main pane displays the 'PERMISSIONS' tab for the project. It shows a summary: '3 users with highly privileged roles Owner / Editor have excess permissions. Improve security by applying recommendations to these users.' Below this, there are tabs for 'MEMBERS' and 'ROLES'. The 'MEMBERS' tab lists users and their roles, along with their analyzed permissions and inheritance status. The table includes columns for Type, Member, Name, Role, Analyzed permissions (excess/total), and Inheritance. The data is as follows:

Type	Member	Name	Role	Analyzed permissions (excess/total)	Inheritance
	190642530876-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	3593/3598	/
	190642530876@cloudbound-services.gserviceaccount.com	Google APIs Service Agent	Editor	3555/3598	/
	acoffman@wustl.edu		Editor	3594/3598	/
	cromwell-server@griffith-lab.iam.gserviceaccount.com	cromwell-server	Genomics Admin Genomics Pipelines Runner Service Account User Cloud Life Sciences Workflows Runner	11/11 4/4 3/5 2/4	/
	john.maruska@wustl.edu		Owner	3577/3919	/
	mgrifflit@wustl.edu	Malachi Griffith	Editor	3529/3598	/
	obigriffith@wustl.edu		Owner	3912/3919	/
	susanna.kiwala@wustl.edu		Editor	3594/3598	/
	terraform@griffith-lab.iam.gserviceaccount.com	terraform	Editor Security Admin Owner Project IAM Admin Secret Manager Admin Secret Manager Secret Accessor	3536/3598 949/959 3854/3919 1/3 10/18 2/3	/
	WashUUTPublicCloud@wustl.edu		Owner	3830/3919	/

Funding Opportunities

- NIH STRIDES Partnership
 - <https://cloud.cit.nih.gov/>
- AnVIL Cloud Credits (AC2) Program
 - <https://anvilproject.org/news/2021/04/12/announcing-anvil-cloud-cost-credits-program>
- AWS Programs for Research and Education
 - <https://aws.amazon.com/grants/>
- GCP Education Credits
 - <https://cloud.google.com/billing/docs/how-to/edu-grants>

AWS/DLT Invoice example



2411 Dulles Corner Park
Suite 800
Herndon, VA 20171

Invoice Questions: 888-358-9346
General Information: 703-709-7172
Fax: 866-352-5855

Invoice No.: SI511473
Order: 4900439
Customer: WUN10
Contract #: N/A

Tax ID No: 54-1599882
CA Reseller: 101643630
DB No: 78-6468199

GST No: 82690 0003 RT0001
MB PST No: 826900003MT0001
SK PST No: 2476547
QST No: 1217287088

Bill To: Washington University in St. Louis
Accounts Payable
Campus Box 1056
700 Rosedale Ave
SAINT LOUIS, MO 63112-1408

Ship To: Washington University in St. Louis
Accounts Payable
ONCOLOGY / SCB
425 SOUTH EUCLID
SOUTHWEST TOWER
SAINT LOUIS, MO 63108

Invoices will be emailed to you monthly for approval.
Work with DLT to ensure the right people get emailed.

- Performance Period
- WASHU PO
- AWS Account #

Date	Period of Performance	ACT # / PDN #		Terms					
02/26/21	01/01/21 - 01/31/21			Net 30 Days					
Purchase Order Number	Order Date	Salesperson	Our Order Number						
2940074K	02/26/21	Jennifer Triplett	4900439						
Quantity	CLIN No.	Item Number	Tax	Unit Price	Amount				
Req.	Ship	B.O.	Description						
1	1	6538-3911-4923	413500	N	1,491.27	1,491.27			

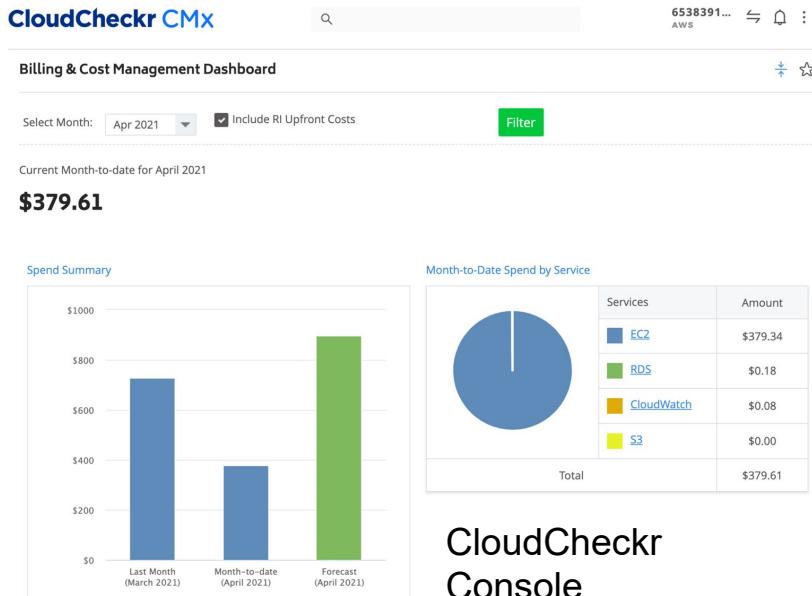
AWS/DLT Invoice example

Description	Qty	Rate	Price
4900439 - 6538-3911-4923 - McDonnell Genome Institute - Genome Griffith Lab			
CloudWatch			
Alarm metric month (standard resolution) - US West (Oregon)	2	0.09500	0.19
GB-mo of log storage - US East (Northern Virginia)	0.01	0.00000	-
GB-mo of log storage - US West (Oregon)	0	0.00000	-
CloudWatch Total	2.01	0.03167	0.19
DynamoDB			
GB - first 10 TB / month data transfer out beyond the global free tier	0	0.00000	-
GB - next 40 TB / month data transfer out	0	0.00000	-
DynamoDB Total	0	0.00000	-
EC2			
Elastic IP address not attached to a running instance per hour (prorated)	2231	0.00475	10.60
EUN1-AWS-Out-Bytes EU (Stockholm)	0	0.00000	-
GB - first 10 TB / month data transfer out beyond the global free tier	1738.81	0.08550	148.67
GB - regional data transfer - in/out/between EC2 AZs or using elastic IPs or ELB	26.48	0.00944	0.25
GB - US West (Oregon) data transfer to Asia Pacific (Mumbai)	0	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Seoul)	0.01	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Singapore)	0.23	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Sydney)	0	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Tokyo)	0.06	0.00000	-
GB - US West (Oregon) data transfer to AWS GovCloud (US)	0.02	0.00000	-
GB - US West (Oregon) data transfer to Canada (Central)	0.23	0.00000	-
GB - US West (Oregon) data transfer to EU (Germany)	0.01	0.00000	-
GB - US West (Oregon) data transfer to EU (Ireland)	1.3	0.01538	0.02
GB - US West (Oregon) data transfer to EU (London)	0	0.00000	-
GB - US West (Oregon) data transfer to EU (Paris)	0	0.00000	-
GB - US West (Oregon) data transfer to South America (Sao Paulo)	0	0.00000	-
GB - US West (Oregon) data transfer to US East (Northern Virginia)	19.4	0.01907	0.37
GB - US West (Oregon) data transfer to US East (Ohio)	0.27	0.03704	0.01
GB - US West (Oregon) data transfer to US West (Northern California)	0	0.00000	-
GB-month of General Purpose SSD (gp2) provisioned storage - US West (Oregon)	2560	0.09500	243.20
GB-Month of snapshot data stored - US West (Oregon)	1824.65	0.04750	86.67
In Asia Pacific (Hong Kong)	0.02	0.00000	-
In Middle East (Bahrain)	0	0.00000	-
On Demand Linux m5.2xlarge Instance Hour	744	0.36480	271.41
On Demand Linux m5a.2xlarge Instance Hour	744	0.32680	243.14
On Demand Linux t2.large Instance Hour	744	0.08816	65.59
On Demand Linux t2.medium Instance Hour	744	0.04409	32.80

- DLT gives a detailed breakdown of cloud services you are being charged for
- This only loosely corresponds to the billing information you see in AWS console
 - The usage should be the same, but *how* you are billed for it can be very different
 - We have seen DLT bills >2x what is shown in AWS
 - Related to reserved instances. DLT calls this “RI arbitrage”. Basically DLT is reserving instances, keeping the savings from Amazon and billing you the full agreed rate.

CloudCheckr CMx (<https://app-us.cloudcheckr.com/>)

- More detailed information on what DLT will actually charge you for AWS usage
- Must request that DLT set this up for you



CloudCheckr
Console



AWS Billing
Console

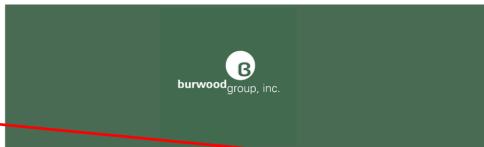
Google / Burwood Invoice example

INVOICE - 02_23_WASUNI-Oncology -
Griffith

Billing Interval: 02-01-23 - 02-28-23

Invoice Date: 02-28-23

Purchase Order: P000021529



CUSTOMER

Washington University in St. Louis
Group: Oncology - Griffith
7425 Forsyth Blvd
Clayton, 63105

CONTACT

Burwood Group Inc
8582 Solutions Center
Chicago, IL 60677, United States

CLOUD PROJECT SUMMARY

Griffith-Lab (Purchase Order: P000021529): \$365.30
Terra-5f2d205 (Purchase Order: P000021529): \$0.16
Terra-D3606022 (Purchase Order: P000021529): \$0.16

Total: \$365.94

DETAILS

Griffith-Lab
(Purchase
Order:
P000021529
)

SERVICE	LIST PRICE	I2 DISCOUNT	PROMOTIONS	ADJUSTMENTS	TOTAL
Cloud Storage	\$166.47	\$-7.93	\$-34.34	\$0	\$124.20
Compute Engine	\$254.56	\$-12.12	\$-1.35	\$0	\$241.10
TOTAL	\$421.04	\$-20.05	\$-35.69	\$0	\$365.30

Invoices will be emailed to you monthly for approval. Work with Burwood to ensure the right people get emailed.

- Performance Period
- WASHU PO
- Google Project

The google cloud console - <https://console.cloud.google.com/>

← → C https://console.cloud.google.com/home/dashboard?organizationId=270103800160&project=griffith-lab Search (/) for resources, docs, products, and more

Google Cloud griffith-lab Search

Cloud overview Products & solutions

PINNED

- Cloud Storage
- Compute Engine
- IAM & Admin
- Billing

VPC network APIs & Services Marketplace Kubernetes Engine BigQuery Cloud Run SQL Logging Security MORE PRODUCTS ▾

RECOMMENDATIONS

Info

Compute Engine

CPU (%)

3 PM 3:15 3:30 3:45

instance/cpu/utilization: 71.78%

Go to Compute Engine

Google Cloud Platform status

Google Compute Engine

Global: High Memory Issues on GCE Windows VMs using Google Cloud Ops Agent >= 2.27.0

Began at 2023-04-11 (15:28:29)

All times are US/Pacific

Data provided by status.cloud.google.com

Go to Cloud status dashboard

Billing

Estimated charges USD \$80.23

For the billing period Apr 1 – 13, 2023

Take a tour of billing

View detailed charges

Monitoring

Create my dashboard

Set up alerting policies

Create uptime checks

View all dashboards

API APIs Requests (requests/sec)

3 PM 3:15 3:30 3:45

Google Billing Dashboard

Billing Reports PRINT SHARE SAVE VIEW LEARN

Billing account WUSTL - Oncology - Griffith - Burw

Overview

Cost management

Reports Cost table Cost breakdown Budgets & alerts Billing export

Cost optimization

Committed use discounts (CUD analysis)

Pricing

Cost estimation

Billing management

Account management

Release Notes

Saved views

March 2023 (total cost) ?
\$743.45 ↑ 101.83%
includes \$102.74 in credits \$375.09 over February 2023

Daily ▲ CSV

Last month

Group by SKU

Folders & Organizations All folders/organizations (11)

Projects All projects (7)

Services All services (11)

SKUs All SKUs (169)

Locations Filter by location data like region and zone.

Labels ? Select the key and values of the labels you want to filter.

Credits

Discounts ?

Sustained use discounts ?

Spending based discounts (contractual) ?

Negotiated savings ?

Invoice level charges ?

Tax

Cost table

Cost & usage report

Cost breakdown

Cost optimization

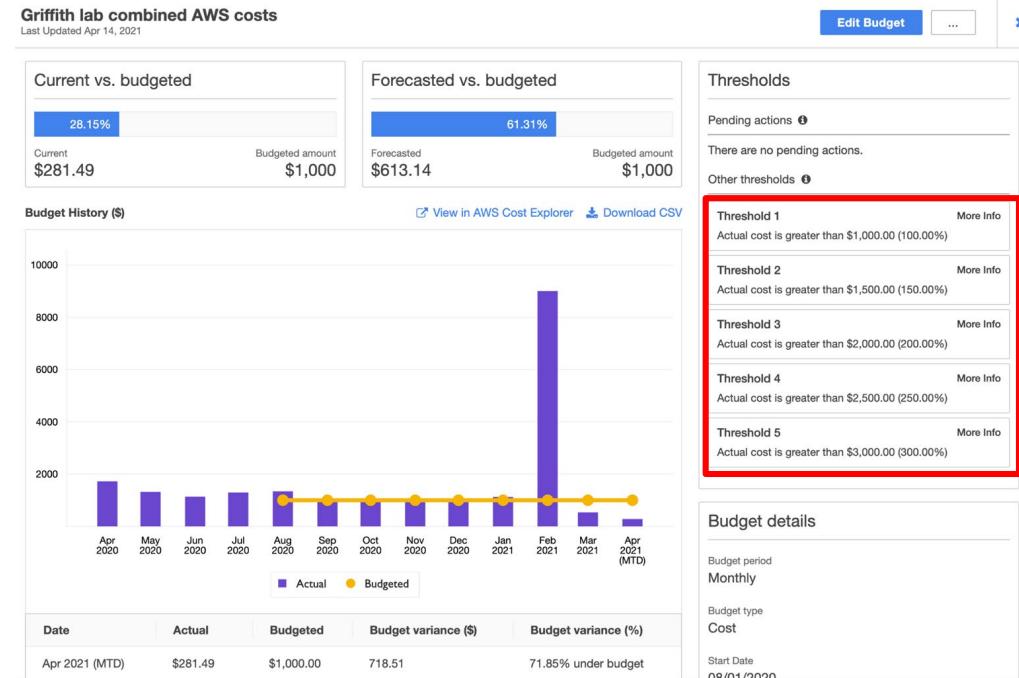
SKU **Service** **SKU ID** **Usage** **Cost** **Discounts** **Promotions and others** **Subtotal**

SKU	Service	SKU ID	Usage	Cost	Discounts	Promotions and others	Subtotal
Custom Instance Core running in Americas	Compute Engine	ACBC-6999-A1C4	7,539.67 hour	\$187.59	-\$9.45	—	\$178.14
Standard Storage US Multi-region	Cloud Storage	0D5D-6E23-4250	6,839.46 gibibyte month	\$133.37	\$0.00	—	\$133.37
Spot Preemptible Custom Instance Core running in Americas	Compute Engine	4A30-9DBE-ECEA	14,825.45 hour	\$103.48	\$0.00	—	\$103.48
Custom Instance Ram running in Americas	Compute Engine	51E2-59BD-7A6E	30,449.28 gibibyte hour	\$101.53	-\$7.15	—	\$94.38
E2 Instance Core running in Americas	Compute Engine	CF4E-A0C7-E3BF	4,221.74 hour	\$69.06	\$0.00	—	\$69.06
Spot Preemptible Custom Instance Ram running in Americas	Compute Engine	AFAD-A1BD-4F9C	56,877.19 gibibyte hour	\$53.46	\$0.00	—	\$53.46

Download CSV

“Budgets” and Billing Alerts in AWS

- Creating a “budget” in AWS allows you to set up email/text alerts when you hit certain amounts each month
- Do this!



“Budgets” and Billing Alerts in GCP

- Creating a “budget” in GCP allows you to set up email/text alerts when you hit certain amounts each month
- Do this!

≡ Google Cloud

budget X Search

Budgets & alerts + CREATE BUDGET DELETE

Budgets track expenses within a Google Cloud Platform project or billing account. Your budget can be a specified amount or based on previous spend. You can set alerts to notify billing admins and users when a budget goes over a specified amount.

i Setting a budget does not cap resource or API consumption. [Learn more.](#)

Filter Enter property name or value

Budget name ↑	Budget period	Budget type	Applies to	Trigger alerts at	Spend and budget amount
<input type="checkbox"/> Google Cloud Griffithlab	Monthly	Specified amount	This billing account	10%, 50%, 100%, 200%, 300%, 400%, 500%, 600%, 700%, 800%, 900%, and 1,000%	<div style="width: 100%;"><div style="width: 100%;">\$80.49 / \$500.00</div></div> <p>Includes -\$11.18 credit</p>

Long-Term Compute: “Reserved Instances” (AWS) and “Commitments” (GCP)

- If you have persistent compute services running use of Reserved Instances or Commitments can result in huge savings
 - E.g. by reserving instances for 3 years at a cost of ~\$8,000 we project savings of ~\$15,000 compared to if we just used on demand resources for that whole time

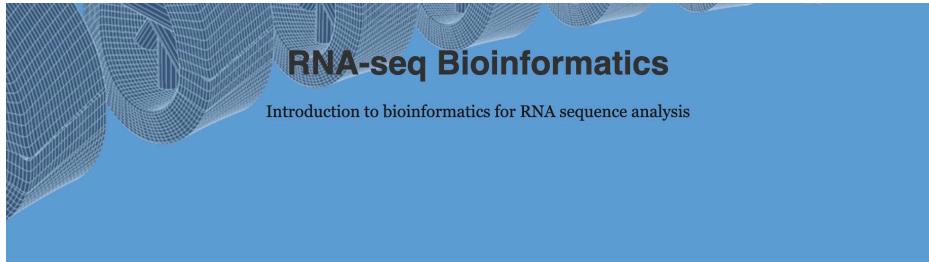
The screenshot shows the AWS EC2 Reserved Instances page. The URL is `us-west-2.console.aws.amazon.com/ec2/v2/home?region=us-west-2#ReservedInstances:`. The page displays a table of reserved instances with the following data:

Instance type	Scope	Availability Zone	Instance count	Start	Expires	Term	Payment option	Offering class
m4.2xlarge	Region	-	1	February 6, 2018, 16:59 (UTC-6:00)	February 5, 2021, 16:59 (UTC-6:00)	3 years	All upfront	Standard
t2.large	Region	-	1	February 5, 2021, 16:59 (UTC-6:00)	February 5, 2024, 16:59 (UTC-6:00)	3 years	All upfront	Standard
m5.2xlarge	Region	-	1	February 5, 2021, 16:59 (UTC-6:00)	February 5, 2024, 16:59 (UTC-6:00)	3 years	All upfront	Standard
m5a.2xlarge	Region	-	1	February 5, 2021, 16:59 (UTC-6:00)	February 5, 2024, 16:59 (UTC-6:00)	3 years	All upfront	Standard

Short-Term Compute: “Spot” (AWS) or “Preemptible” (GCP) Instances

- Meant for “short-term” fault-tolerant jobs
- Much cheaper pricing (up to 90% at times) than regular compute instances
- Drawback is that AWS or GCP reserves the right to take back the machine at any given time depending on the state of the cloud
- Google Preemptible Instances can run at most for 24 hours
- See [GCP Preemptible VM Instances](#) or [Amazon EC2 Spot](#) for more details
- Often useful for running many batch jobs at scale cheaply

Introduction to general AWS cloud computing concepts



Introduction to AWS

[« Prerequisites](#)

[Course](#)

[Log into AWS »](#)

Preamble

Cloud computing allows users to quickly access an arbitrary amount of compute resources from a distance without the need to buy or maintain hardware themselves. There are many cloud computing services. This tutorial describes the use of the Amazon Web Services ([AWS](#)) elastic compute ([EC2](#)) resource. However, the fundamental concepts covered here will generally apply to other cloud computing services such as [Google Cloud](#), [Digital Ocean](#), [Microsoft Azure](#), etc., though with substantial differences in jargon used by each provider.

Table of Contents

- [1. Acknowledgements](#)
- [2. Glossary and abbreviations](#)
- [3. What do I need to perform this tutorial](#)

[3. What do I need to perform this tutorial](#)

[3.1 Creating an account](#)

[3.2 Logging into the AWS console](#)

[4. What is a Region?](#)

[5. How much does it cost to use AWS EC2 resources?](#)

[5.1 How does billing work?](#)

[6. Necessary steps for launching an instance](#)

[6.1 Step 1. Choosing an AMI](#)

[6.2 Step 2. Choosing an instance type](#)

[6.3 Step 3. Configuring instance details](#)

[6.4 Step 4. Adding storage](#)

[6.4.1 Storage volume options](#)

[6.5 Step 5. Tagging the instance](#)

[6.6 Step 6. Configuring a security group](#)

[6.7 Step 7. Reviewing the instance before launch](#)

[6.8 Step 8. Assigning a key pair](#)

[6.9 Step 9. Reviewing launch status](#)

[6.10 Step 10. Examining a new instance in the ec2 console](#)

[6.11 Step 11. Logging into an instance](#)

[7. Trouble-shooting and advanced topics](#)

[7.1 Can not login to EC2 instance - what might have gone wrong?](#)

[7.2 How do storage volumes appear within a linux instance on amazon EC2?](#)

[7.3 Taking stock of compute resources within an ubuntu linux instance](#)

[7.4 Basic setup and administration of an ubuntu linux instance](#)

[7.5 Setting up an Apache web server](#)

[7.6 What is difference between the start, stop, reboot and terminate instance states?](#)

[7.7 How do I create my own AMI, publish as a Community AMI, and what is a snapshot?](#)

https://rnabio.org/module-00-setup/0000/06/01/Intro_to_AWS/

Okay, I'm ready to actually do something on the cloud...

 **WUSTL Oncology**
Division of Oncology, Washington University School of Medicine
📍 United States of America ↗ https://oncology.wustl.edu/ 🏠 Part of Washington University in St....

🏠 Overview 📂 Repositories 6 📄 Projects 📥 Packages 🔍 Teams 1 🌐 People 12 🛡️ Security 🛒 Insights

README.md

 Washington University School of Medicine in St. Louis

Code and documentation from the WUSTL Division of Oncology

These repositories represent accumulated tools and knowledge that enable labs to get up and running with cloud computing, mostly using cancer genome analysis as the use case. Some of the contents include:

- How to get your lab signed up for [Google Cloud access through WUSTL IT](#)
- [End-to-end genomic workflows](#) for variant calling, rnaseq analyses, epigenomics, and more.
- Guidance for [running these pipelines \(or others\)](#) on Google Cloud, including complete walkthrough examples of using it to run an immunogenomics pipeline from [data you've already put on the cloud](#) or from [data stored on the local compute1 cluster](#)
- Guides to running workflows on the local compute cluster, using [GMS \(example\)](#) for workflow orchestration or [Cromwell](#)
- [Annotation files pre-loaded on the cloud](#) that are needed for many of these pipelines, as well as [detailed instructions for creating your own](#)
- Links to useful resources for analysis, cloud computing, or running workflows on other providers such as [Terra](#) or [DNAexus](#)

<https://github.com/wustl-oncology>

Launch an instance from the console and login

The screenshot shows the Google Cloud Compute Engine interface. The top navigation bar includes the Google Cloud logo, a project dropdown set to "griffith-lab", a search bar, and a "Search" button. The main menu on the left is collapsed. The "Compute Engine" section is selected, showing the "VM instances" tab is active. Below the tabs are three buttons: "CREATE INSTANCE", "IMPORT VM", and "REFRESH". A sidebar on the left lists "Virtual machines" with "VM instances" selected, and other options like "Instance templates", "Sole-tenant nodes", "Machine images", "TPUs", "Committed use discounts", and "Reservations". The main content area displays "VM instances" with a table. The table has columns: Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. Two instances are listed: "malachi-test" and "tmooney-immuno-test". Both instances are running (green checkmark) and are located in "us-central1-a". The "malachi-test" instance has an internal IP of 10.128.0.57 (nic0) and an external IP of 34.30.193.57 (nic0). The "tmooney-immuno-test" instance has an internal IP of 10.10.0.63 (nic0) and an external IP of 35.224.42.155 (nic0). The "Connect" column for both instances shows "SSH" with a dropdown arrow and three vertical dots. A message at the top of the table area says: "Instance 'tmooney-immuno-test' is underutilized. You can save an estimated \$14 per month by switching to the machine type: e2-custom. [Learn more](#)". Below the table, there's a "Related actions" section.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input checked="" type="checkbox"/>	malachi-test	us-central1-a			10.128.0.57 (nic0)	34.30.193.57 (nic0)	SSH ▾
<input checked="" type="checkbox"/>	tmooney-immuno-test	us-central1-c	💡 Save \$14 / mo		10.10.0.63 (nic0)	35.224.42.155 (nic0)	SSH ▾

<https://console.cloud.google.com/>

The Google Cloud SDK – working with data in buckets

<https://cloud.google.com/sdk/docs/install> (OR <https://hub.docker.com/r/google/cloud-sdk/>)

- gcloud auth login
- gcloud config list
- gsutil ls
- gsutil cp

```
mgriffit@3071-AL-05015 ~ % gsutil ls
gs://griffith-lab-malachи-adhoc/
gs://griffith-lab-test-kartik/
gs://griffith-lab-test-layth/
gs://griffith-lab-test-malachи/
gs://griffith-lab-test-tmooney/
gs://griffith-lab-workflow-inputs/
gs://griffith-lab-zls_dev-immuno-pipeline/
mgriffit@3071-AL-05015 ~ %
```

Perform an ad hoc analysis in a single Google VM

https://github.com/wustl-oncology/immuno_gcp_wdl_compute1/blob/main/AdHoc.md

Running WDL pipelines on the Google Cloud

subworkflows
tools
alignment_exome.wdl
alignment_exome_nonhuman.wdl
alignment_wgs.wdl
alignment_wgs_nonhuman.wdl
bisulfite.wdl
detect_variants.wdl
detect_variants_nonhuman.wdl
detect_variants_wgs.wdl
downsample_and_recall.wdl
germline_exome.wdl
germline_exome_gvcf.wdl
germline_exome_hla_typing.wdl
germline_wgs.wdl
germline_wgs_gvcf.wdl
immuno.wdl
rnaseq.wdl
rnaseq_star_fusion.wdl
rnaseq_star_fusion_with_xenosplit.wdl
somatic_exome.wdl
somatic_exome_cle.wdl
somatic_exome_nonhuman.wdl
somatic_wgs.wdl
tumor_only_detect_variants.wdl
tumor_only_exome.wdl
tumor_only_wgs.wdl

bam_readcount.wdl
bam_to_trimmed_fastq_and_biscuit_alignments.wdl
bgzip_and_index.wdl
cnvkit_single_sample.wdl
docm_cle.wdl
docm_germline.wdl
filter_vcf.wdl
filter_vcf_nonhuman.wdl
fp_filter.wdl
gatk_haplotypecaller_iterator.wdl
generate_fda_metrics.wdl
generate_fda_metrics_for_bam_or_fastqs.wdl
germline_detect_variants.wdl
germline_filter_vcf.wdl
hs_metrics.wdl
merge_svs.wdl
mutect.wdl
phase_vcf.wdl
pindel.wdl
pindel_cat.wdl
pvacseq.wdl
qc_exome.wdl
qc_exome_no_verify_bam.wdl
qc_wgs.wdl
qc_wgs_nonhuman.wdl
sequence_to.bam_nonhuman.wdl
sequence_to.bqsr.wdl
add_strelka_gt.wdl
add_string_at_line.wdl
add_string_at_line_bgzipped.wdl
add_vep_fields_to_table.wdl
agfusion.wdl
aligned_seq_fda_stats.wdl
annotsv.wdl
annotsv_filter.wdl
bam_readcount.wdl
bam_to_bigwig.wdl
bam_to_cram.wdl
bam_to_fastq.wdl
bcftools_merge.wdl
bedgraph_to_bigwig.wdl
bgzip.wdl
biscuit_align.wdl
biscuit_markdup.wdl
biscuit_pileup.wdl
bisulfite_qc.wdl
bisulfite_vcf2bed.wdl
bqsr.wdl
cat_all.wdl
cat_out.wdl
cnvkit_batch.wdl
cnvkit_vcf_export.wdl
cnvator.wdl
collect_alignment_summary_metrics.wdl

Running a WDL pipeline on Google, in simple terms involves:

- Identifying the specific WDL pipeline you wish to use
- Create a YAML file that describes the inputs/parameters needed for that pipeline (including your data)
- Upload input data to a Google bucket
- Start a Google Virtual Machine (VM) running Cromwell and log into it
- Use Cromwell to run the WDL pipeline with your YAML inputs
- Cromwell manages all resources on Google Cloud
- Once complete results can be stored in a cloud bucket or downloaded locally
- We have created tools (python) to facilitate the above step

<https://github.com/wustl-oncology/analysis-wdls>

An end-to-end tutorial that runs immuno.wdl on public data

Running the WASHU Immunogenomics Workflow on Google Cloud - compute1 version

Preamble

This tutorial demonstrates how to run the WASHU immunogenomics pipeline (immuno.wdl) on Google Cloud. This will be done by first setting up the workflow definitions, input data and reference files and a YAML config file on a user's local system. The user will also set up a Google Cloud environment and all the inputs will be staged to this cloud environment. Next Cromwell will be used to execute the pipeline using the specified input and reference files, and finally the results will be pulled back to the local system and cloud resources will be cleaned up.

This version assumes that you are staging your input data files from the WASHU RIS storage1/compute1 cluster. Some steps are run within docker containers on that cluster. It therefore requires that you have access to storage1 disk and ability to launch jobs on compute1 using LSF (bsub).

Source of instructions

This tutorial is a specific example of how to run a specific pipeline (immuno) on a specific example dataset (HCC1395 Tumor/normal cell line pair). The steps below are taken from the following link where you will find a more generic set of documentation that explains in detail how to run any WDL pipeline on the Google Cloud using tools created to assist this process. <https://github.com/wustl-oncology/cloud-workflows/tree/main/manual-workflows>

Prerequisites

- google-cloud-sdk
- git

Security

https://github.com/wustl-oncology/immuno_gcp_wdl_compute1