CS 240

#15: IaaS and Public/Private Clouds

Computer Systems

March 10, 2022 · Wade Fagen-Ulmschneider

Abstractions Using Virtualization Technologies

Modern cloud computing is full of abstractions on top of virtualization technologies. We describe categories of abstractions to understand what kind of technologies we're using and what we're abstracting.

Legend:

	Abstracted by Cloud Provider		Customer Managed Unit of Scale		Customer Managed
--	---------------------------------	--	-----------------------------------	--	------------------

laaS Infrastructure as a Service	CaaS Containers as a Service	PaaS Platform as a Service	FaaS Functions as a Service	SaaS Software as a Service
Data	Data	Data	Data	Data
Functions	Functions	Functions	Functions	Functions
Applications	Applications	Applications	Applications	Applications
Runtime	Runtime	Runtime	Runtime	Runtime
Containers*	Containers	Containers*	Containers*	Containers*
OS	os	os	os	os
Virtualization	Virtualization	Virtualization	Virtualization	Virtualization
Hardware	Hardware	Hardware	Hardware	Hardware

^{*:} May or may not be present on a specific technology stack.

What is the "Cloud"?

Originally just a marketing term, a "cloud" is widely understood to have three features:

1.

2.

3.

A private cloud is	using the
resources.	
•	
_	
•	
•	
Public Clouds	
A public cloud isrevenue model.	and has a
There are dozens of public clou	d providers, but three market leaders:
1.	
2.	
3.	

Abstraction: Infrastructure as a Service (IaaS)

When you choose to host your technology on IaaS:

- You Choose (Unit of Scale):
- Abstracted by Cloud Provider:
- Public Cloud Example:

Amazon EC2 Microsoft Azure Virtual Machine Google Compute Engine

• Why IaaS?

Public Cloud IaaS:

AWS provides IaaS as their EC2 product. Current generation general purpose computing:

• t4g: Uses AWS Graviton2 CPUs (ARM)

• t3a: Uses AMD CPUs (x86-64)

• t3: Uses Intel CPUs

	vCPU	RAM	t4g	t3	t3a		
Hourly Rate:							
nano	2	o.5 GiB	\$0.0042	\$0.0052	\$0.0047		
micro	2	1 GiB	\$0.0084	\$0.0104	\$0.0094		
small	2	2 GiB	\$0.0168	\$0.0208	\$0.0188		
medium	2	4 GiB	\$0.0336	\$0.0416	\$0.0376		
Monthly Rate (30 days × 24 hours):							
nano	2	o.5 GiB	\$ 3.02	\$ 3.74	\$ 3.38		
micro	2	1 GiB	\$ 6.05	\$ 7.49	\$ 6.77		
small	2	2 GiB	\$ 12.10	\$ 14.98	\$ 13.53		
medium	2	4 GiB	\$ 24.19	\$ 29.95	\$ 27.07		

On-demand general purpose hourly rate, as of March 9, 2022 https://aws.amazon.com/ec2/pricing/on-demand/

Azure provides Linux VMs as "Azure Virtual Machines":

- B1: Uses Intel CPUs (equivalent to AWS t3)
- Different naming schemes, absolutely equal pricing to t3

	vCPU	RAM	B1
B1ls	1	o.5 GiB	\$0.0052
B1s	2	1 GiB	\$0.0104
B1ms	2	2 GiB	\$0.0207
B1s	2	4 GiB	\$0.0416

On-demand general purpose hourly rate, as of March 9, 2022 https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/

Google provides Linux VMs as "Google Compute Engine" and has both a "shared core option" (similar to AWS, Azure) and general purpose "dedicated core".

	vCPU (fractional share)	RAM	Price
e2-micro	2 (0.25)	1 GiB	\$0.008376
e2-small	2 (0.5)	2 GiB	\$0.016751
e2-medium	2 (1.0)	4 GiB	\$0.033503
e2-standard-2	2	8 GiB	\$0.067006

https://cloud.google.com/compute/vm-instance-pricing

University of Illinois Private Cloud

We have set up a virtual machine for you on the UIUC Private Cloud. This allows you to use your VM without restriction or cost concerns, and you have an on-demand resource to use!

- You will need to "start it" each time you want to use it.
- (It will turn off at 4am each morning.)
- It's your **completely-from-scratch** server that you will use for upcoming MPs and the final project!
- It is identical to the offering from AWS/Azure.

Q: Suppose you want to manage a database server and a web service yourselves and use an IaaS solution:

- The database requires 4 GiB of RAM
- The web server requires 1 GiB of RAM
- Neither server is CPU-intensive

What is the cost of running these servers on Azure for March 2022?

^{*:} While Google is explicit with the fractional share, AWS and Azure also have "CPU Credits" that accumulate at a rate based on the tier of the machine that allows for burst workloads.