Cloud Computing Fundamentals

CS6501: Serverless AI

Fall 2025

Lecture 3

Yue Cheng

University of Virginia

Some material taken/derived from:
- Wisconsin CS 320 by Tyler Caraza-Harter

Learning objectives

- Know basic cloud billing models
- Understand concepts of cloud computing paradigms including IaaS, PaaS, and FaaS
- Learn some of the problems of today's clouds (lock-in, cloud resource scaling, cloud economics, pay-as-you-go)

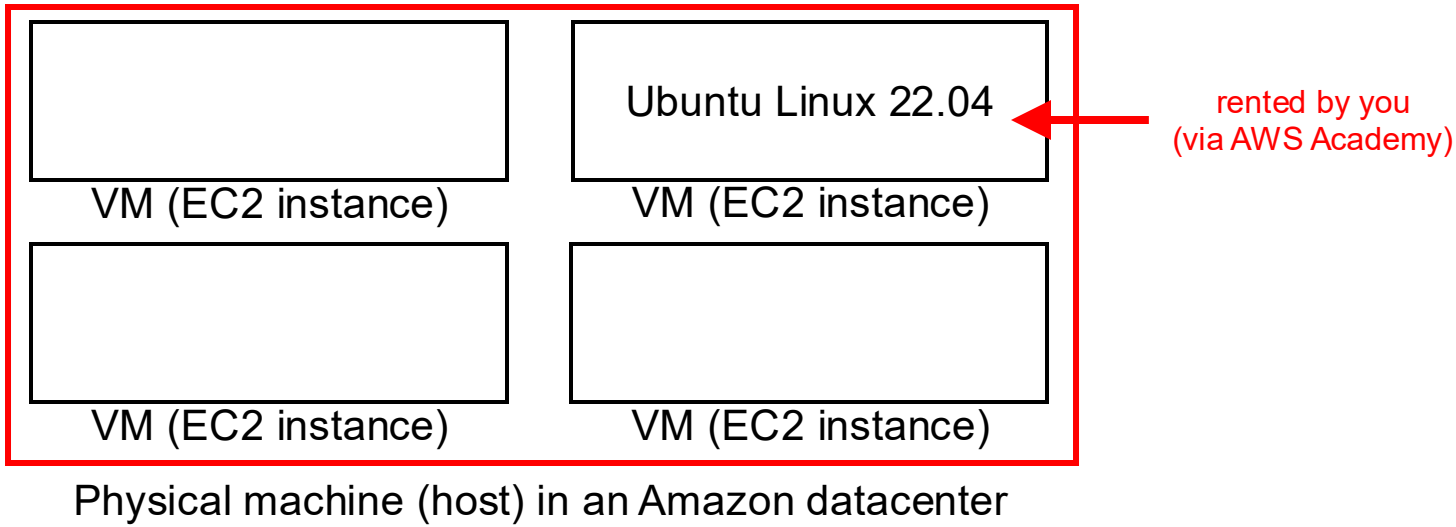
Background

The beginning

"Sometimes you need a lot of processing power; and sometimes you need just a little. Sometimes you need a lot, but you only need it for a limited amount of time."
-- Jeff Barr (https://aws.amazon.com/blogs/aws/amazon_ec2_beta/)

Amazon Web Services (AWS)

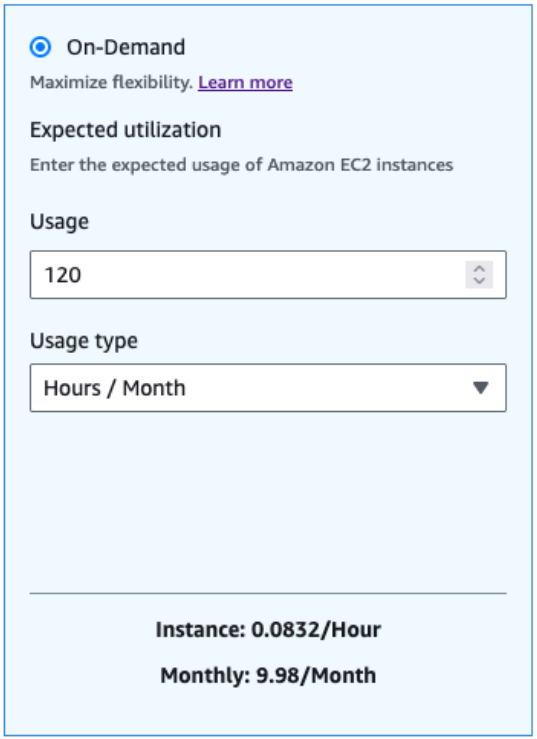
- Elastic Computing Cloud (EC2), rented VMs, launched in 2006
- "**Infrastructure as a Service**" (IaaS): rent infrastructure (compute, storage, network) instead of owning the hardware yourself



VM hours

Pricing summary

t3.large | Family: t3 | 2vCPU | 8 GiB Memory



Amazon EC2 On-Demand instances cost (Monthly): 9.98
Amazon Elastic Block Store (EBS) total cost (Monthly): 1.28

AWS pricing calculator: https://calculator.aws/#/

Pricing comparison

- one VM for a month: about $10
- about 120 hours a month (4*30)
- 120 VMs for an hour: about $10
- same computation + storage resources
- very different wait time

Be careful!!

- programmers previously optimized when things were too slow
- now we need to optimize when it is too expensive
- cost is not always obvious at the moment you're running a job (need to do "back of the envelope" estimates before you deploy the resources)

Region: US East (N. Virginia) | Pricing Unit: Instance | Cost: Hourly | Reserved: 1-year - No Upfront | Currency: United States Dollar ($) | Columns: Columns | Compare | Clear Filters

Name	API Name	Instance Memory	vCPUs	Instance Storage	Network Performance	On Demand	Linux
Filter Name...	Filter API Name...	>=0	>=0	>=0	Filter Network Performan	Filter On Demand	Filte
A1 Double Extra Large	a1.2xlarge	16 GiB	8 vCPUs	EBS only	Up to 10 Gigabit	$0.204 hourly	$0.128
A1 Quadruple Extra Large	a1.4xlarge	32 GiB	16 vCPUs	EBS only	Up to 10 Gigabit	$0.408 hourly	$0.25
A1 Large	a1.large	4 GiB	2 vCPUs	EBS only	Up to 10 Gigabit	$0.051 hourly	$0.03
A1 Medium	a1.medium	2 GiB	1 vCPUs	EBS only	Up to 10 Gigabit	$0.0255 hourly	$0.016
A1 Metal	a1.metal	32 GiB	16 vCPUs	EBS only	Up to 10 Gigabit	$0.408 hourly	$0.25
A1 Extra Large	a1.xlarge	8 GiB	4 vCPUs	EBS only	Up to 10 Gigabit	$0.102 hourly	$0.06
C1 High-CPU Medium	c1.medium	1.7 GiB	2 vCPUs	350 GB HDD	Moderate	$0.13 hourly	$0.09
C1 High-CPU Extra Large	c1.xlarge	7 GiB	8 vCPUs	1680 GB (4×420 GB ...	High	$0.52 hourly	$0.36
C3 High-CPU Double Extra Large	c3.2xlarge	15 GiB	8 vCPUs	160 GB (2×80 GB SSD)	High	$0.42 hourly	$0.29
C3 High-CPU Quadruple Extra Large	c3.4xlarge	30 GiB	16 vCPUs	320 GB (2×160 GB SS...	High	$0.84 hourly	$0.58
C3 High-CPU Eight Extra Large	c3.8xlarge	60 GiB	32 vCPUs	640 GB (2×320 GB S...	10 Gigabit	$1.68 hourly	$1.168
C3 High-CPU Large	c3.large	3.75 GiB	2 vCPUs	32 GB (2×16 GB SSD)	Moderate	$0.105 hourly	$0.07
C3 High-CPU Extra Large	c3.xlarge	7.5 GiB	4 vCPUs	80 GB (2×40 GB SSD)	Moderate	$0.21 hourly	$0.146
C4 High-CPU Double Extra Large	c4.2xlarge	15 GiB	8 vCPUs	EBS only	High	$0.398 hourly	$0.25
C4 High-CPU Quadruple Extra Large	c4.4xlarge	30 GiB	16 vCPUs	EBS only	High	$0.796 hourly	$0.50
C4 High-CPU Eight Extra Large	c4.8xlarge	60 GiB	36 vCPUs	EBS only	10 Gigabit	$1.591 hourly	$1.008
C4 High-CPU Large	c4.large	3.75 GiB	2 vCPUs	EBS only	Moderate	$0.10 hourly	$0.06
C4 High-CPU Extra Large	c4.xlarge	7.5 GiB	4 vCPUs	EBS only	High	$0.199 hourly	$0.126
C5 High-CPU 12xlarge Extra Large	c5.12xlarge	96 GiB	48 vCPUs	EBS only	12 Gigabit	$2.04 hourly	$1.285
C5 High-CPU 18xlarge	c5.18xlarge	144 GiB	72 vCPUs	EBS only	25 Gigabit	$3.06 hourly	$1.928
C5 High-CPU 24xlarge	c5.24xlarge	192 GiB	96 vCPUs	EBS only	25 Gigabit	$4.08 hourly	$2.57
C5 High-CPU Double Extra Large	c5.2xlarge	16 GiB	8 vCPUs	EBS only	Up to 10 Gigabit	$0.34 hourly	$0.214
C5 High-CPU Quadruple Extra Large	c5.4xlarge	32 GiB	16 vCPUs	EBS only	Up to 10 Gigabit	$0.68 hourly	$0.42

Other cloud services

- AWS now has > 200 services beyond EC2 (and growing)

Other cloud services

- **IaaS** (Infrastructure as a Service)
  - EC2, other services that feel closer to raw hardware
  - Virtual disks, virtual network, some storage systems, etc.
  - Cheap + flexible – you can deploy & run anything on it (Spark, Ray, etc.)
- **PaaS** (Platform as a Service)
  - Cloud providers has deployed systems on the infrastructure; you pay to use the deployed system
  - Databases, application framework/platforms, ML training/deployment systems
  - Less flexible, easier to use
  - Often more expensive (though not necessarily more than doing it yourself due to efficiencies available to cloud provider but not you)
- Line between IaaS and PaaS distinction is a bit subjective.

Other cloud services

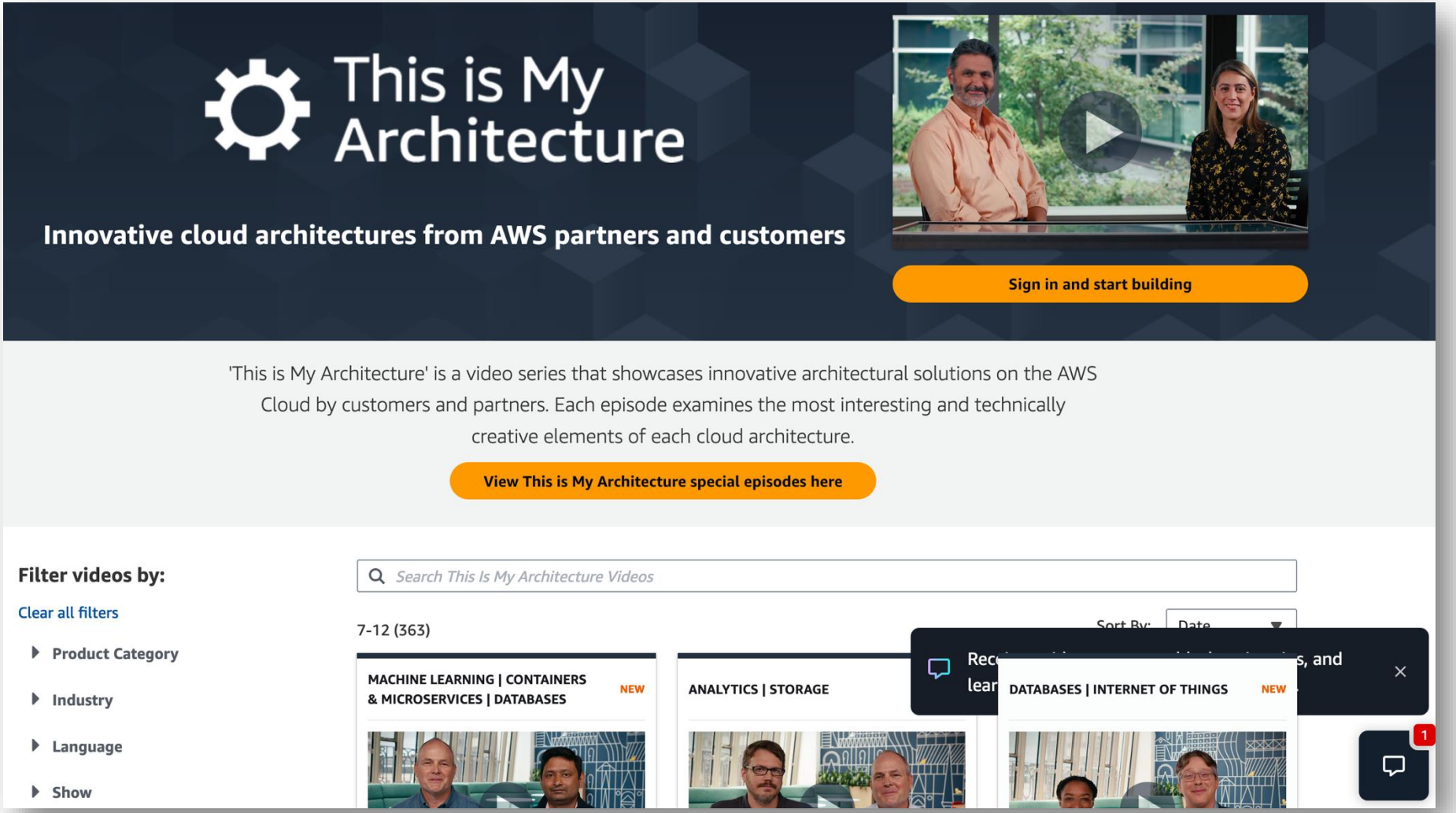
- **FaaS** (Function as a Service)
  - AWS Lambda, the very first FaaS platform across all public cloud providers
  - Users upload code packaged in λ "functions" and AWS helps provision it, auto-scale it, and tear it down
  - Finer-grained billing at millisecond level
  - Bundled CPU+memory resources
  - Cheap but not as flexible – you don't need to worry about deployment

Trends

- What AWS cloud services are most popular today?
- Market share of major cloud providers

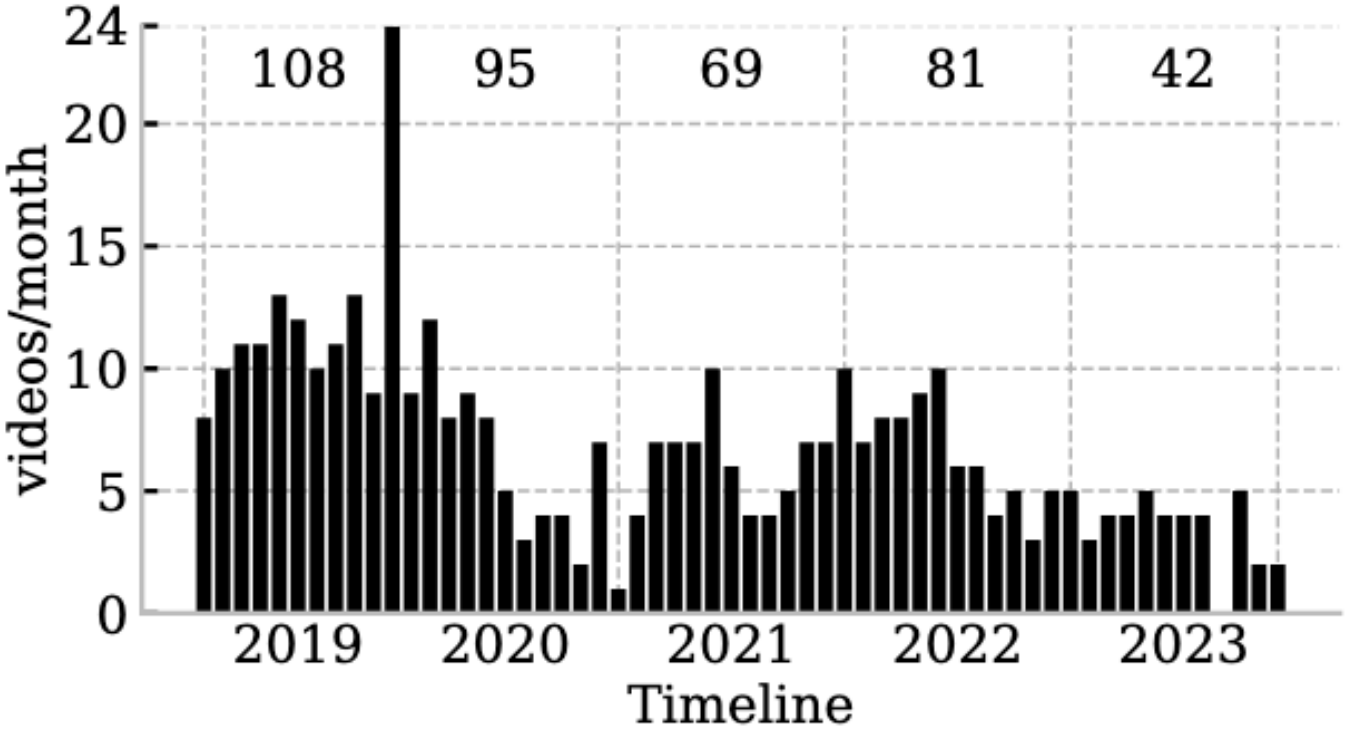
Q: How do we know which AWS services are most popular in today's cloud-native apps?

Analyzing AWS' own video series



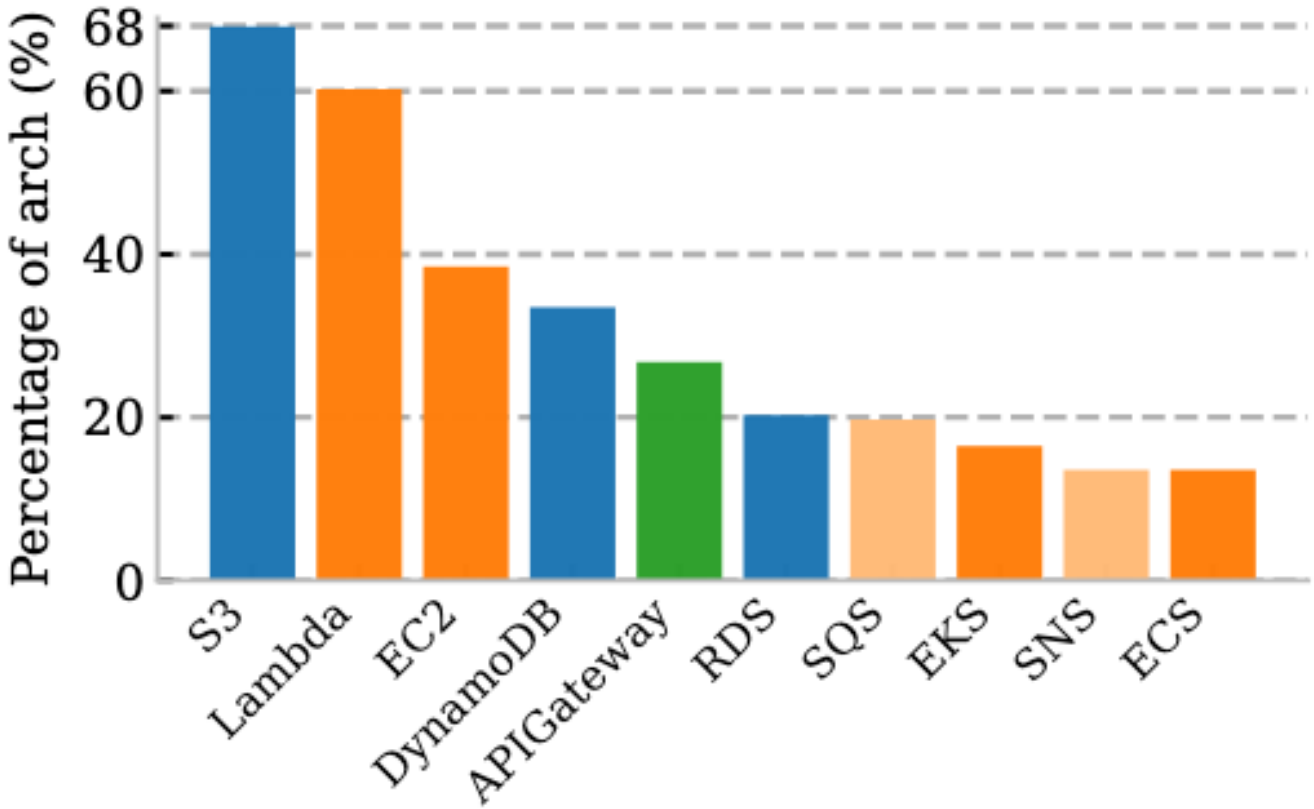
https://aws.amazon.com/architecture/this-is-my-architecture/

Distribution of video release date



* Cloudscape: A Study of Storage Services in Modern Cloud Architectures [USENIX FAST 2025]

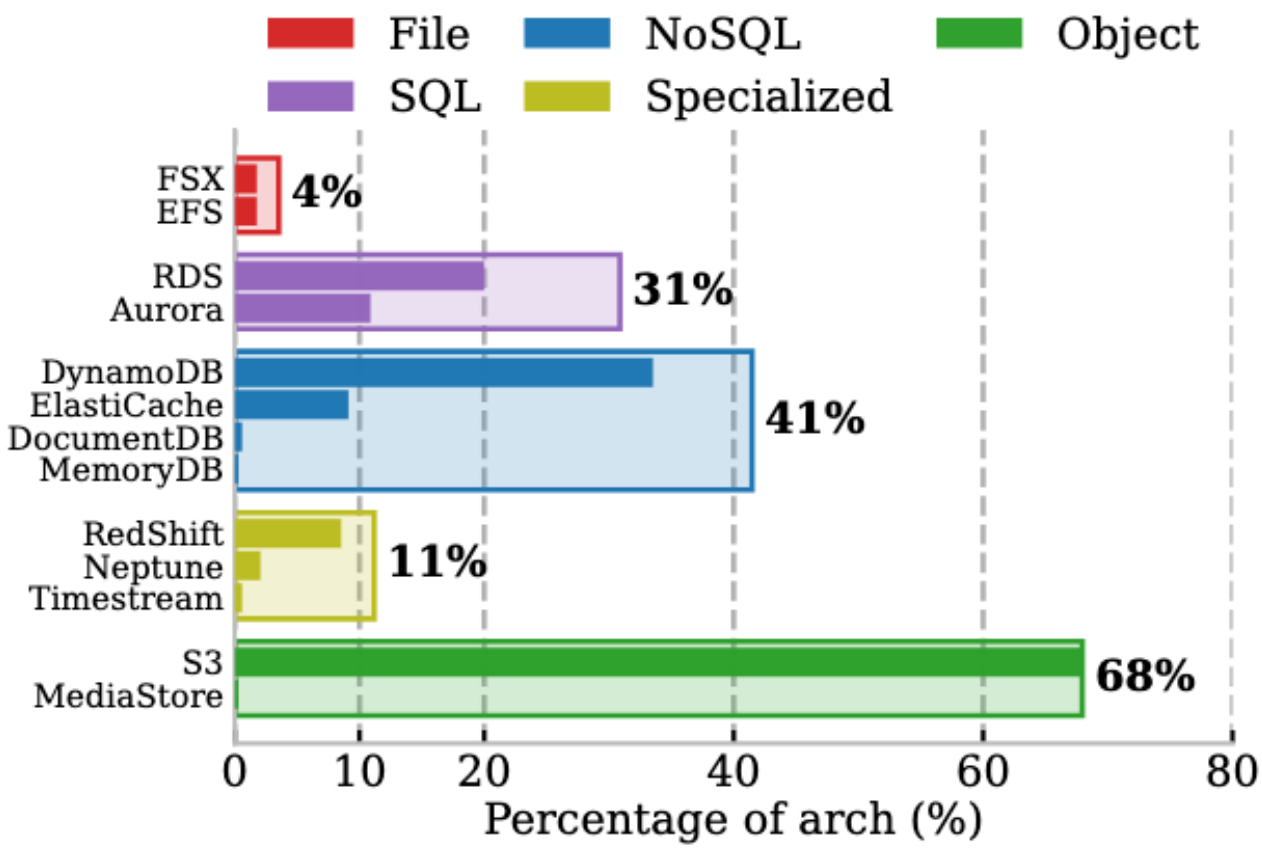
Popularity of different AWS services



All services including compute and storage

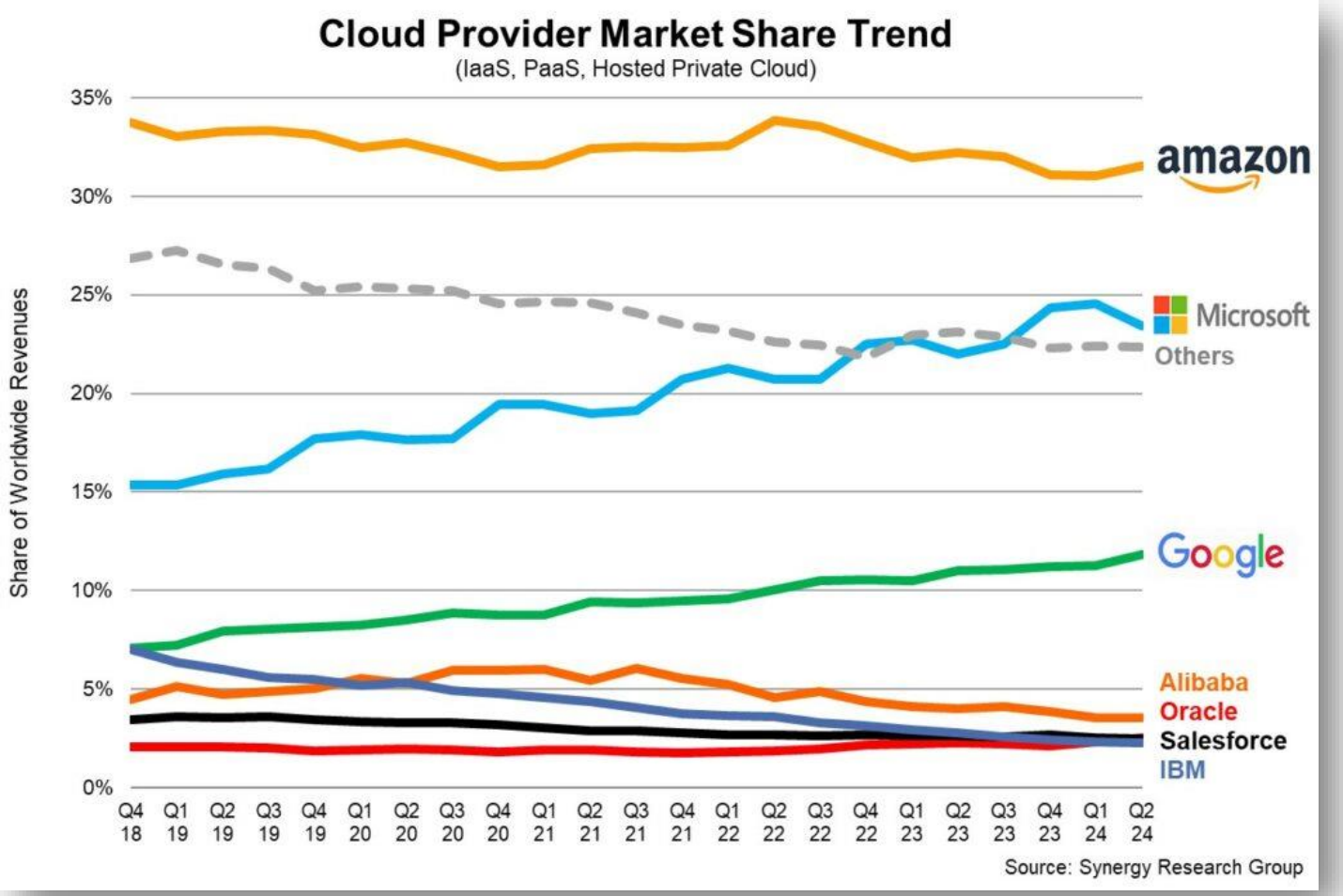
* Cloudscape: A Study of Storage Services in Modern Cloud Architectures [USENIX FAST 2025]

Usage of different storage services



* Cloudscape: A Study of Storage Services in Modern Cloud Architectures [USENIX FAST 2025]

Cloud provider market share trend



https://holori.com/cloud-market-share-2024-aws-azure-gcp/

Lock-in

Lock-in

- Customers (tenants) worry: what if the cloud provider increases the price? If it's hard to move to a competing cloud, you're "**locked in**"!
- **PaaS:** services are often unique, and it would be hard to move to a different cloud providers
- **IaaS:** services like VMs are more uniform – it would be easier to switch to a different cloud to find the cheapest place to rent VMs
- **Data:** cloud providers often make it free to bring data into the cloud (free **ingress**) but expensive to take it out (expensive **egress** $$$$)

Case study: Dropbox

- A data sync startup founded back in 2008
- Became popular so quickly
  - Peak number of users: 500+ Million
  - Overall amount of data stored: 500 PB
- Initially stored all data on public clouds (AWS)
- Seriously considered to move data out of AWS
- Cloud vendor lock in
  - **Enormous** egress cost 
- Now still parts of its data services sitting on AWS

Cloud economics and billing models

Tenants: Pay-as-you-go?

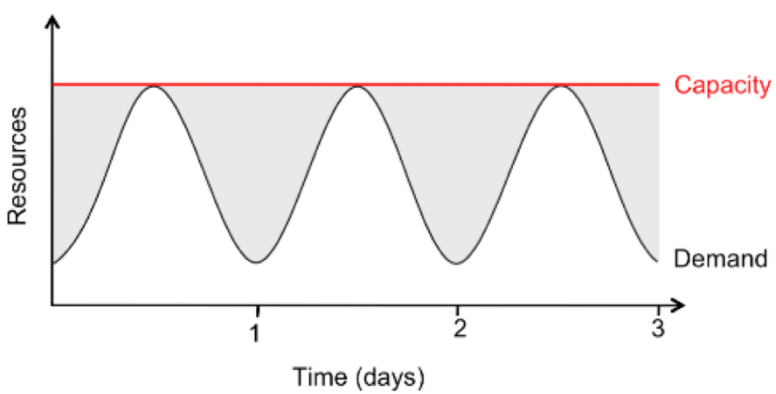
- (**Claimed**) pay-as-you-go pricing
  - Usage-based?
  - Many (compute) services charged per minute
    - Except for Lambda, which is charged per millisecond
  - EC2 charged per second
  - Storage and network services charged per byte
  - No minimum or upfront fee

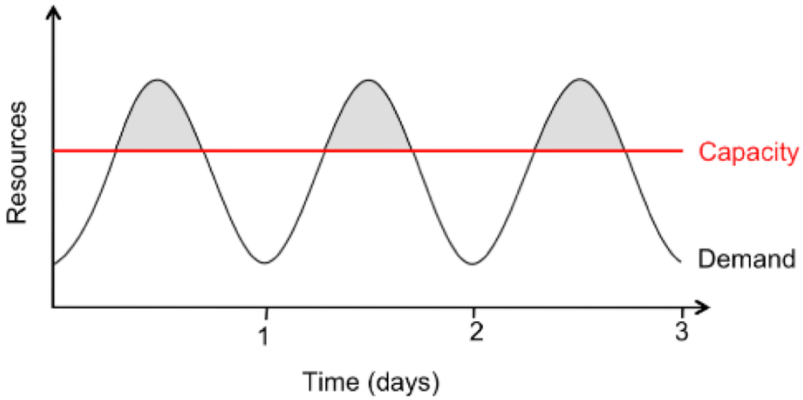
Tenants: Pay-as-you-go?

- (**Claimed**) pay-as-you-go pricing
  - Usage-based?
  - Many (compute) services charged per minute
  - Storage and network services charged per byte
  - No minimum or upfront fee

Q: Is the cloud pricing truly pay-as-you-go?

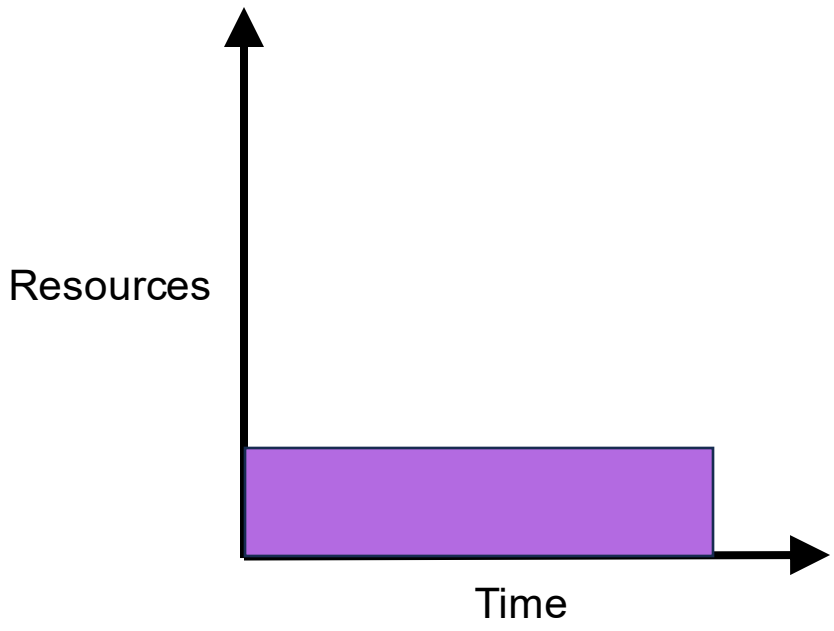
- **Problem**: How to perform strategic planning?

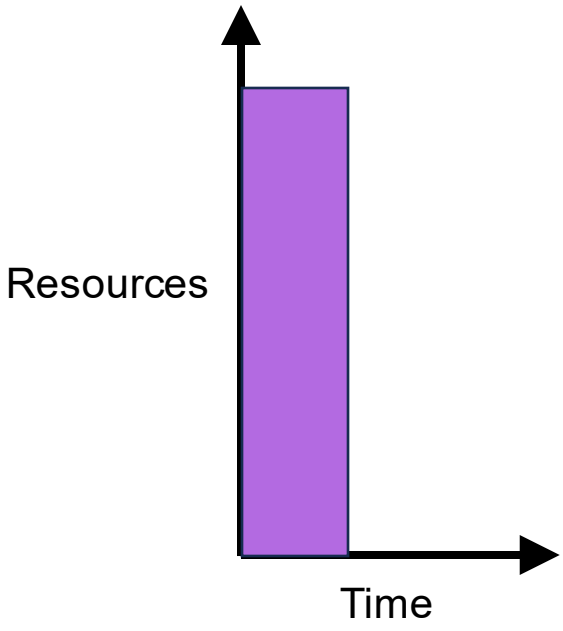




Tenants: Scalability gained?

- (**Ideally**) Linear scalability & perfect elasticity
  - Using 1,000 servers for 1 hour costs the same as 1 server for 1,000 hours
  - Same price to get a result faster

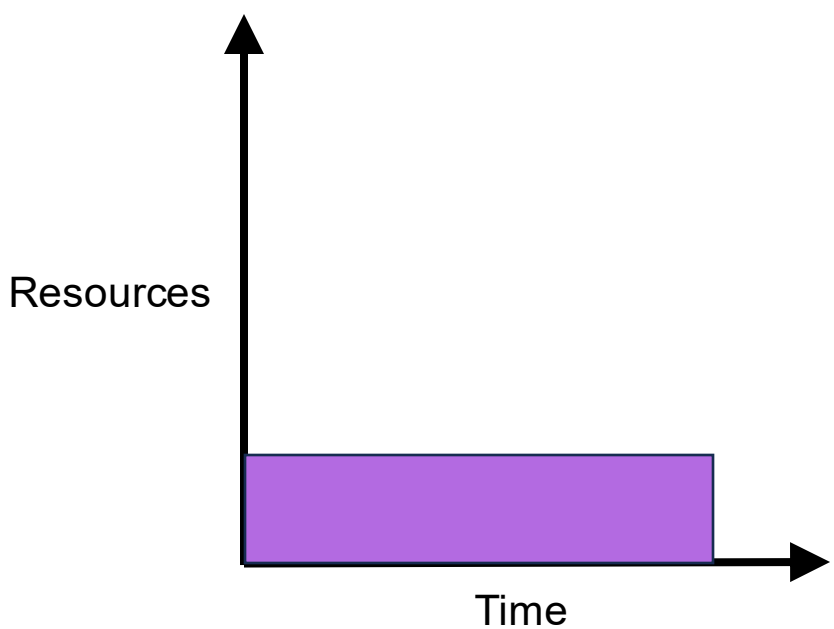


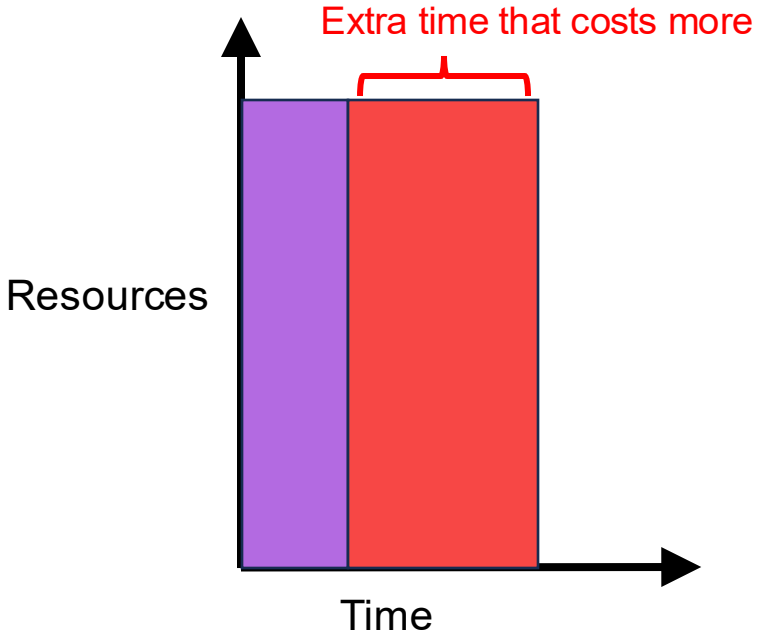


In practice, it really depends, case by case.

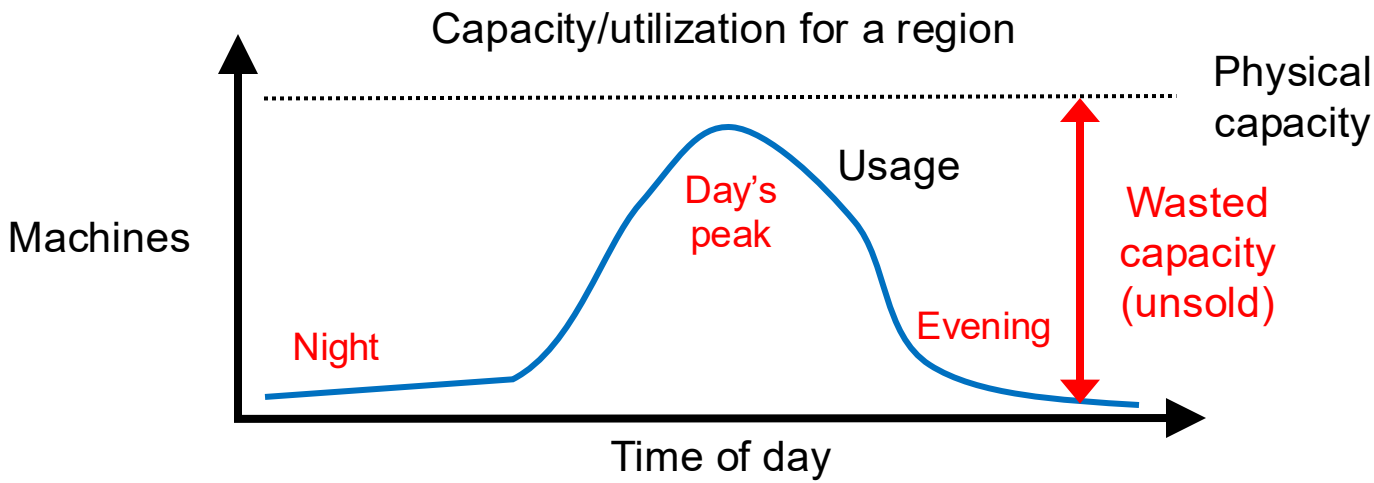
Likely the speedup of the computation is much lower than 1,000X!

- (**Reality**) Scalability is sublinear and VM scaling is slow.
  - Using 1,000 servers for 1+N hour costs N times more than 1 server for 1,000 hours
  - Often higher price to get a result faster





Providers: On-demand vs. spot instances



- How to create incentives for tenants?
  - Use less at peak time
  - Use more at low times

- Two VM deployment options
  - On-demand instances: Constant (high) price. Can generally get a VM. Won't be taken away from you arbitrarily. Used when capacity is needed at specific times.
  - Spot instances: Price varies throughput day. If you're not willing to pay enough, your computation waits for a cheaper price. VM might be interrupted ("preempted") once started. Excellent for once-a-day batch jobs.

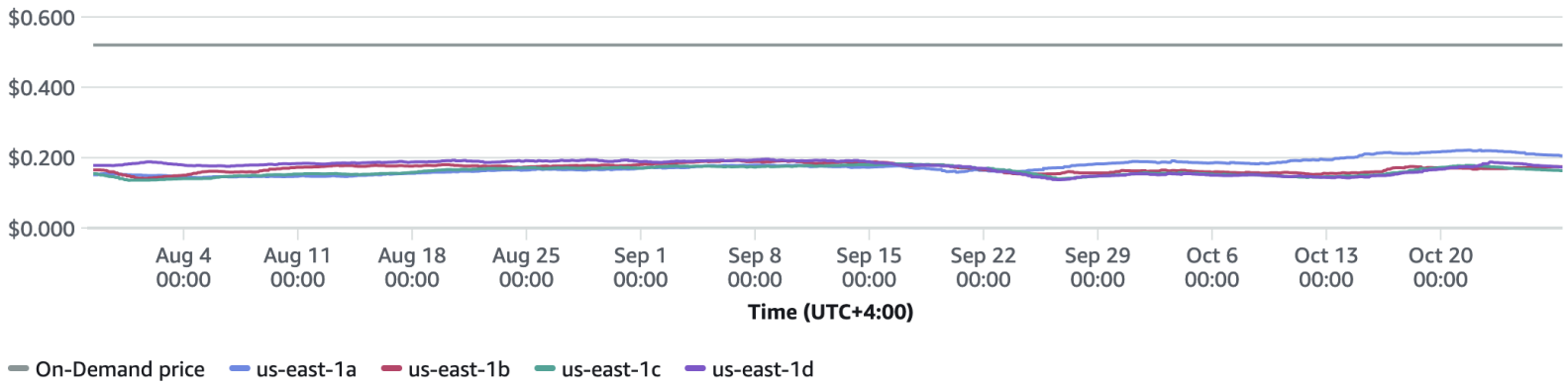
Spot instance pricing (c1.xlarge)

Spot Instance pricing history

Your instance type requirements, budget requirements, and application design will determine how to apply the following best practices for your application. To learn more, see Spot Instance Best Practices

Graph	Instance type	Platform	Date range
Availability Zones	c1.xlarge	Linux/UNIX	3 months

Prices



Average per hour within date range

On-Demand	us-east-1a	us-east-1b	us-east-1c Cheapest	us-east-1d
$0.5200	**$0.1743**	**$0.1698**	**$0.1612**	**$0.1742**
	$0.0218 per vCPU	$0.0212 per vCPU	$0.0201 per vCPU	$0.0218 per vCPU
	66.48% saving	67.34% saving	69.00% saving	66.51% saving

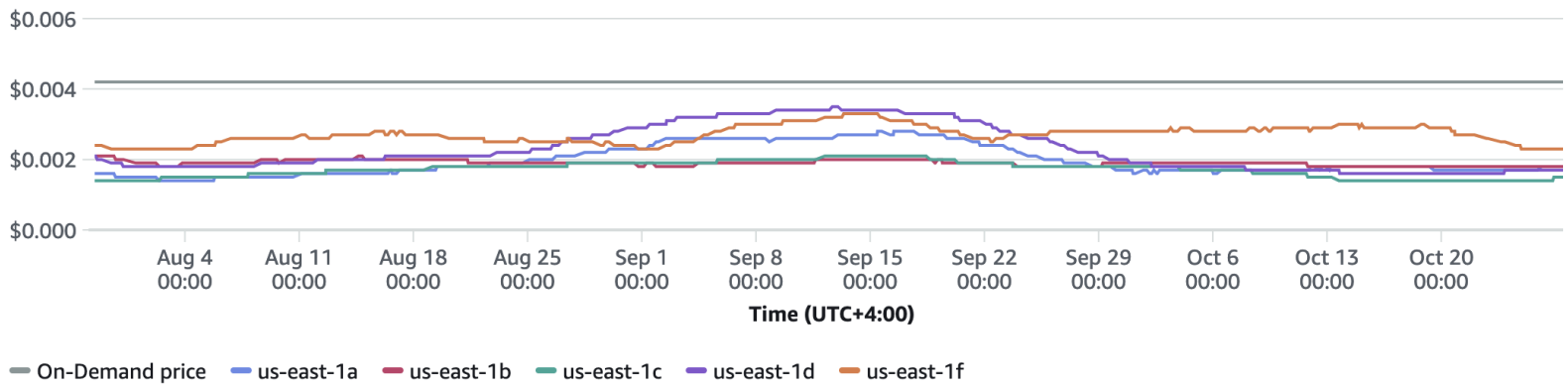
Spot instance pricing (t4g.nano)

Spot Instance pricing history

Your instance type requirements, budget requirements, and application design will determine how to apply the following best practices for your application. To learn more, see Spot Instance Best Practices

Graph	Instance type	Platform	Date range
Availability Zones	t4g.nano	Linux/UNIX	3 months

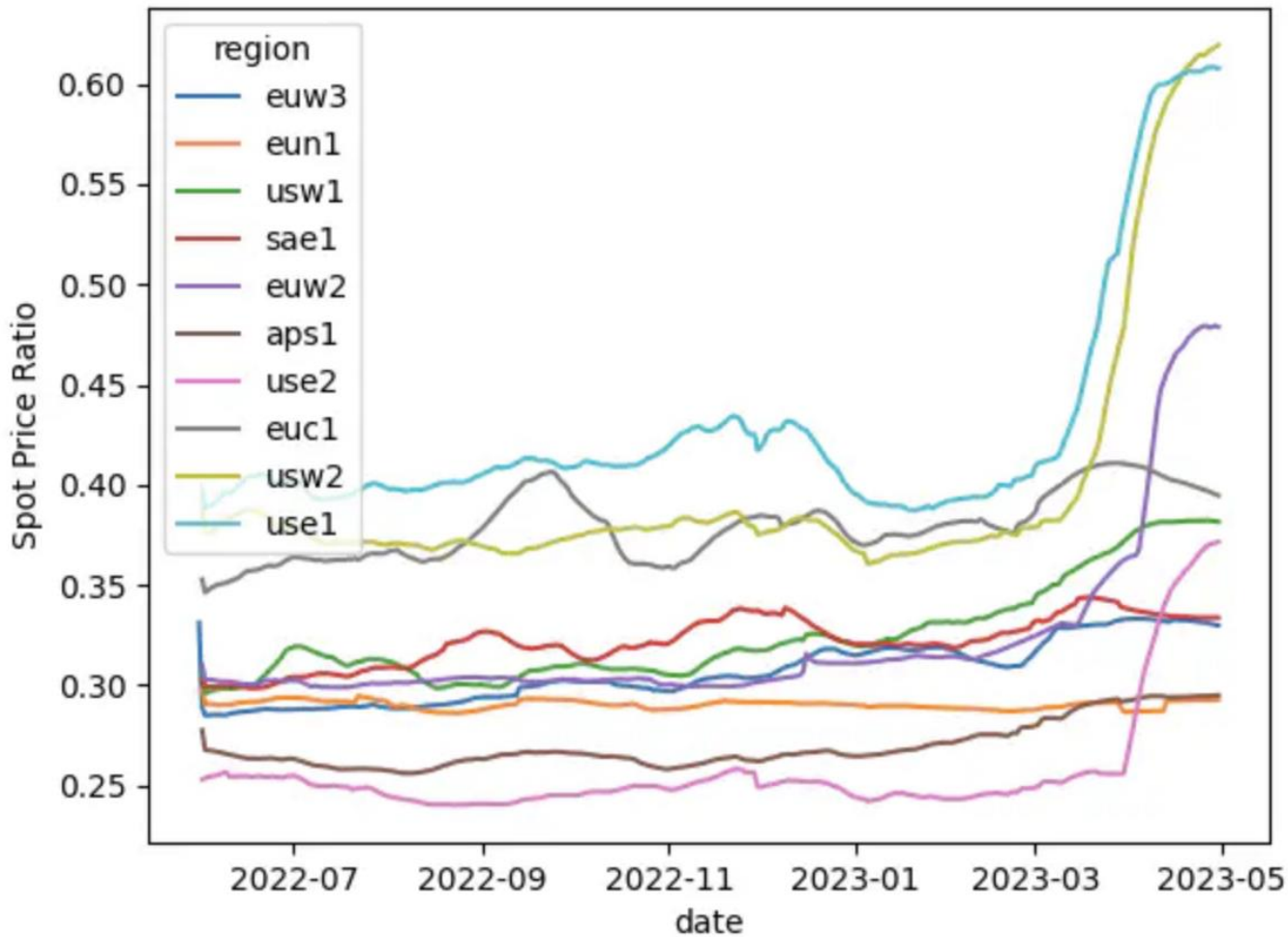
Prices



Average per hour within date range

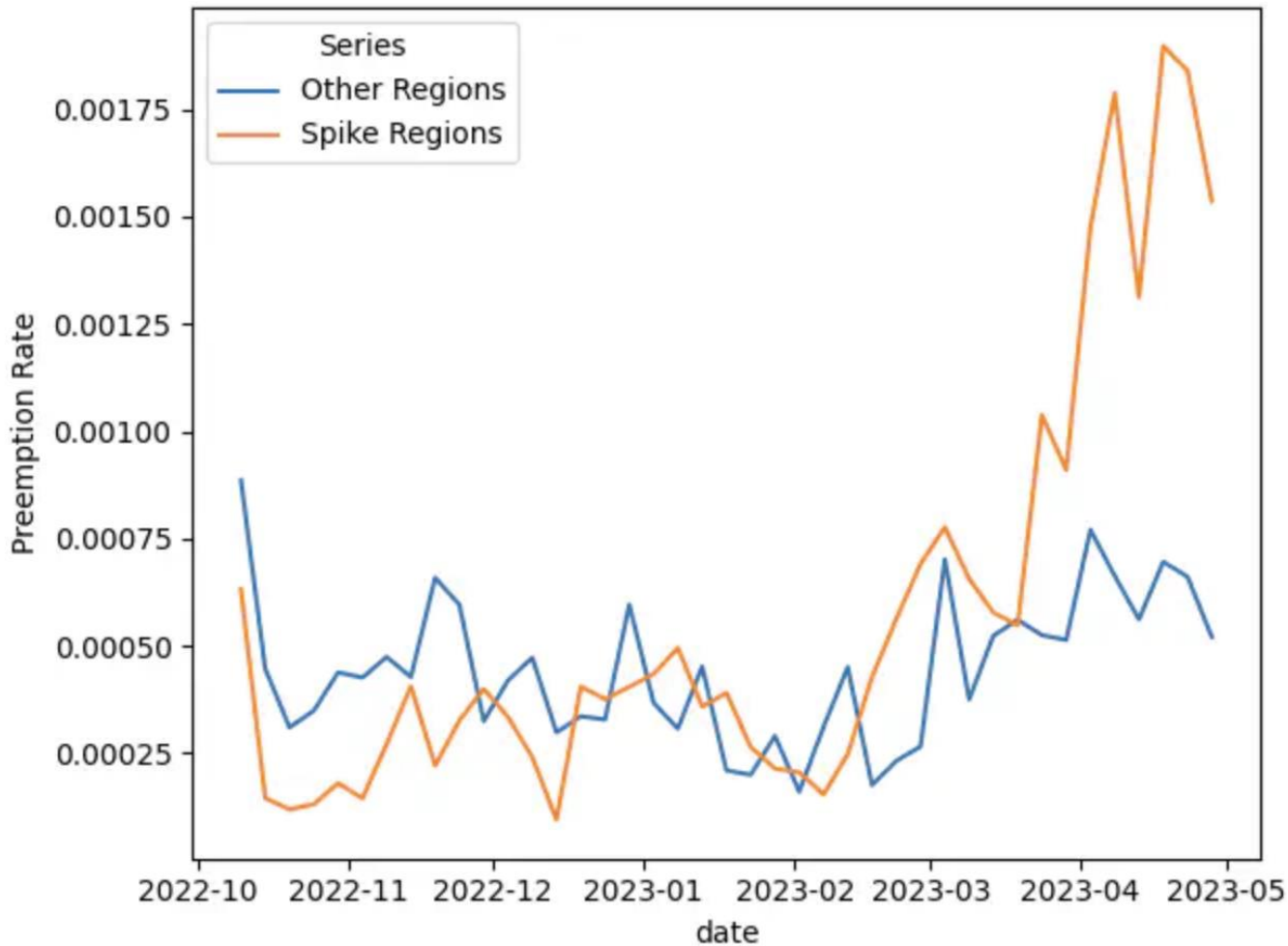
On-Demand	us-east-1a	us-east-1b	us-east-1c Cheapest	us-east-1d	us-east-1f
$0.0042	**$0.0020**	**$0.0019**	**$0.0017**	**$0.0023**	**$0.0027**
	$0.0010 per vCPU	$0.0009 per vCPU	$0.0009 per vCPU	$0.0011 per vCPU	$0.0014 per vCPU
	53.12% saving	54.80% saving	58.95% saving	45.44% saving	35.55% saving

Mean spot price ratios across regions



https://pauley.me/post/2023/spot-price-trends/

Spot instance preemption ratio (t3/t4)



https://pauley.me/post/2023/spot-price-trends/

Providers: Free tier, discounts at scale

AWS Lambda Pricing

Region: US East (N. Virginia)

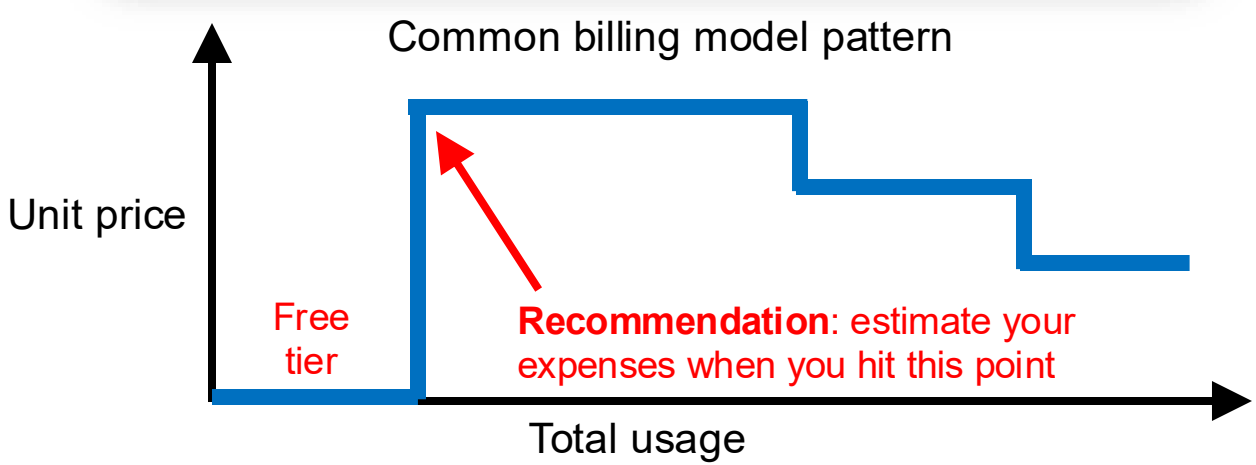
Architecture	Duration
x86 Price	
First 6 Billion GB-seconds / month	$0.0000166667 for every GB-second
Next 9 Billion GB-seconds / month	$0.000015 for every GB-second
Over 15 Billion GB-seconds / month	$0.0000133334 for every GB-second

AWS Lambda example

"The AWS Lambda **free tier** includes one million free requests per month and 400,000 GB-seconds of compute time per month."

(https://aws.amazon.com/lambda/pricing/)

"Duration is calculated from the time your code begins executing until it returns or otherwise terminates, rounded up to the nearest 1 ms."



Recommendation: check if you have a large number of small ops getting rounded up