

# Cloud Computing Fundamentals

*DS5110: Cloud Computing*

*Spring 2025*

Lecture 13

Yue Cheng



Some material taken/derived from:

- Wisconsin CS 320 by Tyler Caraza-Harter

@ 2025 released for use under a [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

# Background

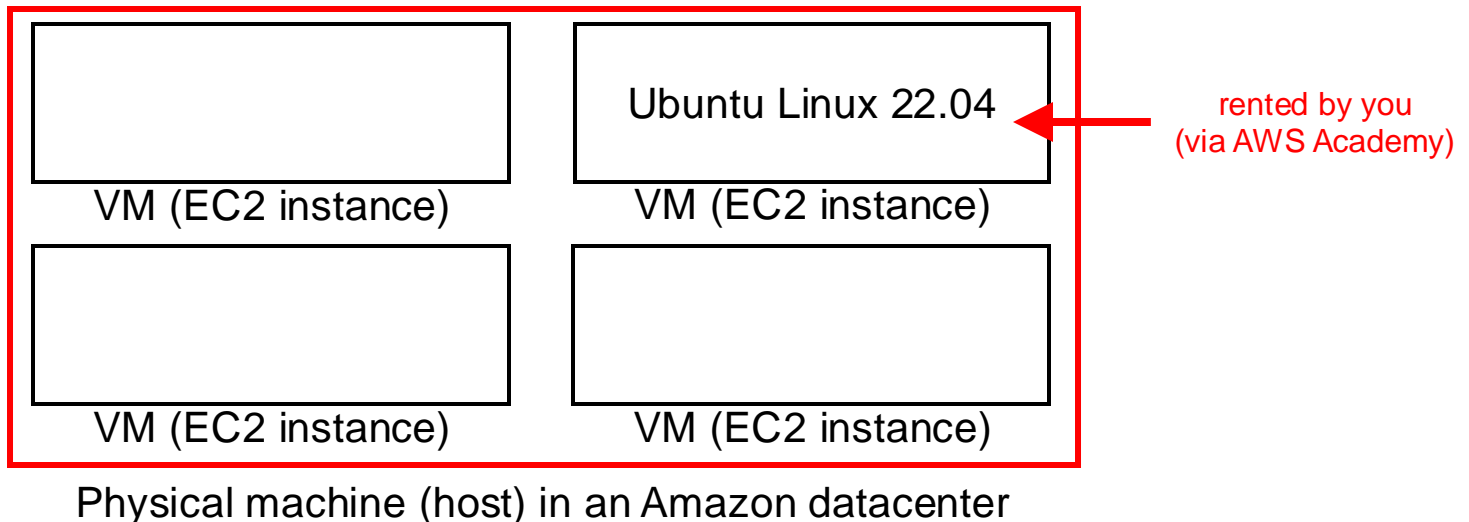
# The beginning

“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/](https://aws.amazon.com/blogs/aws/amazon_ec2_beta/))

## Amazon Web Services (AWS)

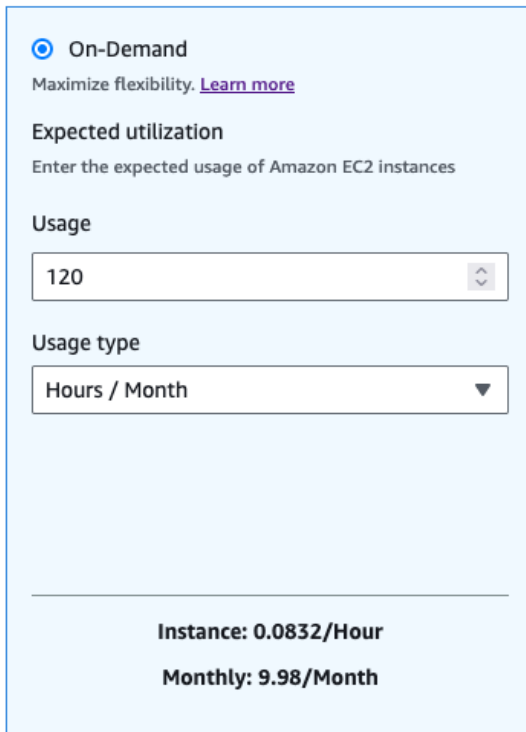
- Elastic Computing Cloud (EC2), rented VMs, launched in 2006
- “**Infrastructure as a Service**” (IaaS): rent infrastructure (compute, storage, network) instead of owning the hardware yourself



# VM hours

## Pricing summary

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



The screenshot shows the AWS Pricing Calculator interface for an Amazon EC2 On-Demand instance. It includes a section for 'Expected utilization' with a 'Usage' input field set to 120 and a 'Usage type' dropdown menu set to 'Hours / Month'. At the bottom, it displays the calculated costs: 'Instance: 0.0832/Hour' and 'Monthly: 9.98/Month'.

☒ On-Demand  
Maximize flexibility. [Learn more](#)

Expected utilization  
Enter the expected usage of Amazon EC2 instances

Usage  
120

Usage type  
Hours / Month

---

Instance: 0.0832/Hour  
Monthly: 9.98/Month

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)

# EC2Instances.info

Easy Amazon EC2 Instance Comparison

EC2

RDS

Region: US East (N. Virginia) -

Cost: Hourly -

Reserved: 1-year - No Upfront -

Columns -

Compare Selected

Clear Filters

CSV

Filter: Min Memory (GiB):

Min vCPUs:

Min Storage (GiB):

Name	API Name	Memory	vCPUs	Instance Storage	Network Performance	Linux On Demand cost	Linux Reserved cost	Windows On Demand cost	Windows Reserved cost
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>
M5DN Extra Large	m5dn.xlarge	16.0 GiB	4 vCPUs	150 GiB NVMe SSD	Up to 25 Gigabit	\$0.272000 hourly	\$0.173000 hourly	\$0.456000 hourly	\$0.357000 hourly
M5A Double Extra Large	m5a.2xlarge	32.0 GiB	8 vCPUs	EBS only	Up to 10 Gigabit	\$0.344000 hourly	\$0.219000 hourly	\$0.712000 hourly	\$0.587000 hourly
R5N 12xlarge	r5n.12xlarge	384.0 GiB	48 vCPUs	EBS only	50 Gigabit	\$3.576000 hourly	\$2.253000 hourly	\$5.784000 hourly	\$4.461000 hourly
R5AD Extra Large	r5ad.xlarge	32.0 GiB	4 vCPUs	150 GiB NVMe SSD	10 Gigabit	\$0.262000 hourly	\$0.166000 hourly	\$0.446000 hourly	\$0.350000 hourly
R5N Extra Large	r5n.xlarge	32.0 GiB	4 vCPUs	EBS only	Up to 25 Gigabit	\$0.298000 hourly	\$0.188000 hourly	\$0.482000 hourly	\$0.372000 hourly
I3EN 12xlarge	i3en.12xlarge	384.0 GiB	48 vCPUs	30000 GiB (4 * 7500 GiB NVMe SSD)	50 Gigabit	\$5.424000 hourly	\$3.694000 hourly	\$7.632000 hourly	\$5.902000 hourly
I3EN Metal	i3en.metal	768.0 GiB	96 vCPUs	60000 GiB (8 * 7500 GiB NVMe SSD)	100 Gigabit	\$10.848000 hourly	\$7.388000 hourly	\$15.264000 hourly	\$11.804000 hourly
R5DN Extra Large	r5dn.xlarge	32.0 GiB	4 vCPUs	150 GiB NVMe SSD	Up to 25 Gigabit	\$0.334000 hourly	\$0.211000 hourly	\$0.518000 hourly	\$0.395000 hourly
I2 Extra Large	i2.xlarge	30.5 GiB	4 vCPUs	800 GiB SSD	Moderate	\$0.853000 hourly	\$0.424000 hourly	\$0.973000 hourly	\$0.565000 hourly
M5N 16xlarge	m5n.16xlarge	256.0 GiB	64 vCPUs	EBS only	75 Gigabit	\$3.808000 hourly	\$2.419000 hourly	\$6.752000 hourly	\$5.363000 hourly
T2 Micro	t2.micro	1.0 GiB	1 vCPUs <u>for a 2h 24m burst</u>	EBS only	Low to Moderate	\$0.011600 hourly	\$0.007200 hourly	\$0.016200 hourly	\$0.011800 hourly
D2 Eight Extra Large	d2.8xlarge	244.0 GiB	36 vCPUs	48000 GiB (24 * 2000 GiB HDD)	10 Gigabit	\$5.520000 hourly	\$3.216000 hourly	\$6.198000 hourly	\$3.300000 hourly
I3EN 3xlarge	i3en.3xlarge	96.0 GiB	12 vCPUs	7500 GiB NVMe SSD	Up to 25 Gigabit	\$1.356000 hourly	\$0.924000 hourly	\$1.908000 hourly	\$1.476000 hourly
Z1D 3xlarge	z1d.3xlarge	96.0 GiB	12 vCPUs	450 GiB NVMe SSD	Up to 10 Gigabit	\$1.116000 hourly	\$0.705000 hourly	\$1.668000 hourly	\$1.257000 hourly
X1E 16xlarge	x1e.16xlarge	1952.0 GiB	64 vCPUs	1920 GiB SSD	10 Gigabit	\$13.344000 hourly	\$8.223000 hourly	\$16.288000 hourly	\$11.167000 hourly
R5N 24xlarge	r5n.24xlarge	768.0 GiB	96 vCPUs	EBS only	100 Gigabit	\$7.152000 hourly	\$4.506000 hourly	\$11.568000 hourly	\$8.922000 hourly
I2 Eight Extra Large	i2.8xlarge	244.0 GiB	32 vCPUs	6400 GiB (8 * 800 GiB SSD)	10 Gigabit	\$6.820000 hourly	\$3.392000 hourly	\$7.782000 hourly	\$4.521000 hourly
R5A Eight Extra Large	r5a.8xlarge	256.0 GiB	32 vCPUs	EBS only	Up to 10 Gigabit	\$1.808000 hourly	\$1.141000 hourly	\$3.280000 hourly	\$2.613000 hourly
A1 Metal	a1.metal	32.0 GiB	16 vCPUs	EBS only	Up to 10 Gigabit	\$0.408000 hourly	\$0.257000 hourly	unavailable	unavailable
I2 Double Extra Large	i2.2xlarge	61.0 GiB	8 vCPUs	1600 GiB (2 * 800 GiB SSD)	High	\$1.705000 hourly	\$0.848000 hourly	\$1.946000 hourly	\$1.131000 hourly
I3EN Double Extra Large	i3en.2xlarge	64.0 GiB	8 vCPUs	5000 GiB (2 * 2500 GiB NVMe SSD)	Up to 25 Gigabit	\$0.904000 hourly	\$0.616000 hourly	\$1.272000 hourly	\$0.984000 hourly
M5A Extra Large	m5a.xlarge	16.0 GiB	4 vCPUs	EBS only	Up to 10 Gigabit	\$0.172000 hourly	\$0.109000 hourly	\$0.356000 hourly	\$0.293000 hourly
P3 Double Extra Large	p3.2xlarge	61.0 GiB	8 vCPUs	EBS only	Up to 10 Gigabit	\$3.060000 hourly	\$2.088000 hourly	\$3.428000 hourly	\$2.456000 hourly
T2 Double Extra Large	t2.2xlarge	32.0 GiB	8 vCPUs <u>for a 4h 4.8m burst</u>	EBS only	Moderate	\$0.371200 hourly	\$0.230000 hourly	\$0.433200 hourly	\$0.292000 hourly
H1 Eight Extra Large	h1.8xlarge	128.0 GiB	32 vCPUs	8000 GiB (4 * 2000 GiB HDD)	10 Gigabit	\$1.872000 hourly	\$1.272000 hourly	\$3.344000 hourly	\$2.744000 hourly
R5D 24xlarge	r5d.24xlarge	768.0 GiB	96 vCPUs	3600 GiB (4 * 900 GiB NVMe SSD)	25 Gigabit	\$6.912000 hourly	\$4.362000 hourly	\$11.328000 hourly	\$8.778000 hourly
I3EN 6xlarge	i3en.6xlarge	192.0 GiB	24 vCPUs	15000 GiB (2 * 7500 GiB NVMe SSD)	25 Gigabit	\$2.712000 hourly	\$1.847000 hourly	\$3.816000 hourly	\$2.951000 hourly
R4 High-Memory Eight Extra Large	r4.8xlarge	244.0 GiB	32 vCPUs	EBS only	10 Gigabit	\$2.128000 hourly	\$1.344000 hourly	\$3.600000 hourly	\$2.816000 hourly
T2 Large	t2.large	8.0 GiB	2 vCPUs <u>for a 7h 12m burst</u>	EBS only	Low to Moderate	\$0.092800 hourly	\$0.057500 hourly	\$0.120800 hourly	\$0.085500 hourly
X1 Extra High-Memory 16xlarge	x1.16xlarge	976.0 GiB	64 vCPUs	1920 GiB SSD	High	\$6.669000 hourly	\$4.110000 hourly	\$9.613000 hourly	\$7.054000 hourly
M5A 16xlarge	m5a.16xlarge	256.0 GiB	64 vCPUs	EBS only	12 Gigabit	\$2.752000 hourly	\$1.751000 hourly	\$5.696000 hourly	\$4.695000 hourly
R5 Metal	r5.metal	768.0 GiB	96 vCPUs	EBS only	25 Gigabit	\$6.048000 hourly	\$3.810000 hourly	\$10.464000 hourly	\$8.226000 hourly
R5A Large	r5a.large	16.0 GiB	2 vCPUs	EBS only	10 Gigabit	\$0.113000 hourly	\$0.071000 hourly	\$0.205000 hourly	\$0.163000 hourly
C3 High-CPU Large	c3.large	3.75 GiB	2 vCPUs	32 GiB (2 * 16 GiB SSD)	Moderate	\$0.105000 hourly	\$0.073000 hourly	\$0.188000 hourly	\$0.165000 hourly
R5A 24xlarge	r5a.24xlarge	768.0 GiB	96 vCPUs	EBS only	20 Gigabit	\$5.424000 hourly	\$3.423000 hourly	\$9.840000 hourly	\$7.839000 hourly
G3 16xlarge	g3.16xlarge	488.0 GiB	64 vCPUs	EBS only	20 Gigabit	\$4.560000 hourly	\$3.112200 hourly	\$7.504000 hourly	\$6.056200 hourly
A1 Double Extra Large	a1.2xlarge	16.0 GiB	8 vCPUs	EBS only	Up to 10 Gigabit	\$0.204000 hourly	\$0.128500 hourly	unavailable	unavailable
C4 High-CPU Extra Large	c4.xlarge	7.5 GiB	4 vCPUs	EBS only	High	\$0.199000 hourly	\$0.126000 hourly	\$0.383000 hourly	\$0.310000 hourly
X1E Quadruple Extra Large	x1e.4xlarge	488.0 GiB	16 vCPUs	480 GiB SSD	Up to 10 Gigabit	\$3.336000 hourly	\$2.056000 hourly	\$4.072000 hourly	\$2.792000 hourly
M5AD Extra Large	m5ad.xlarge	16.0 GiB	4 vCPUs	150 GiB NVMe SSD	Up to 10 Gigabit	\$0.206000 hourly	\$0.132000 hourly	\$0.390000 hourly	\$0.316000 hourly


# 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  “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

# Analyzing AWS' own video series



## This is My Architecture

Innovative cloud architectures from AWS partners and customers



Sign in and start building

'This is My Architecture' is a video series that showcases innovative architectural solutions on the AWS Cloud by customers and partners. Each episode examines the most interesting and technically creative elements of each cloud architecture.

[View This is My Architecture special episodes here](#)

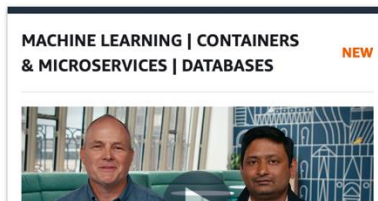
### Filter videos by:

[Clear all filters](#)

- ▶ Product Category
- ▶ Industry
- ▶ Language
- ▶ Show

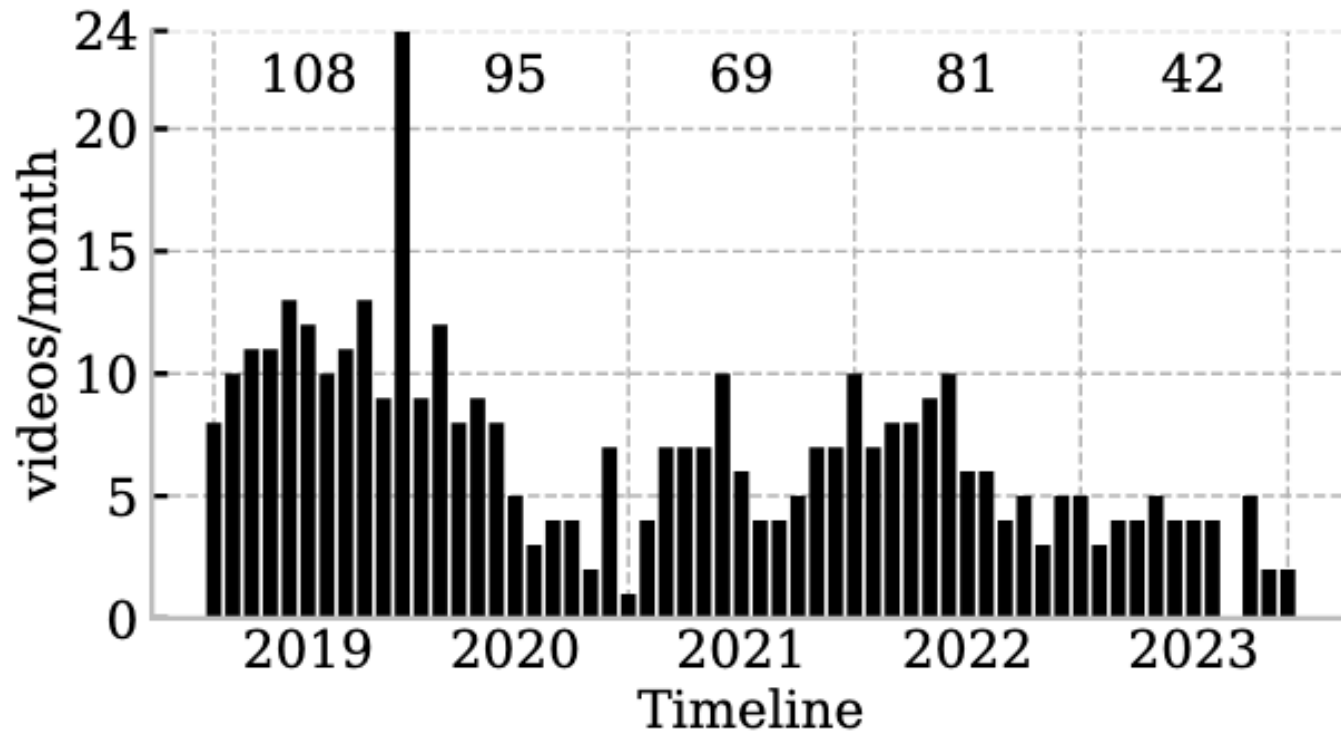
Search This Is My Architecture Videos

7-12 (363)



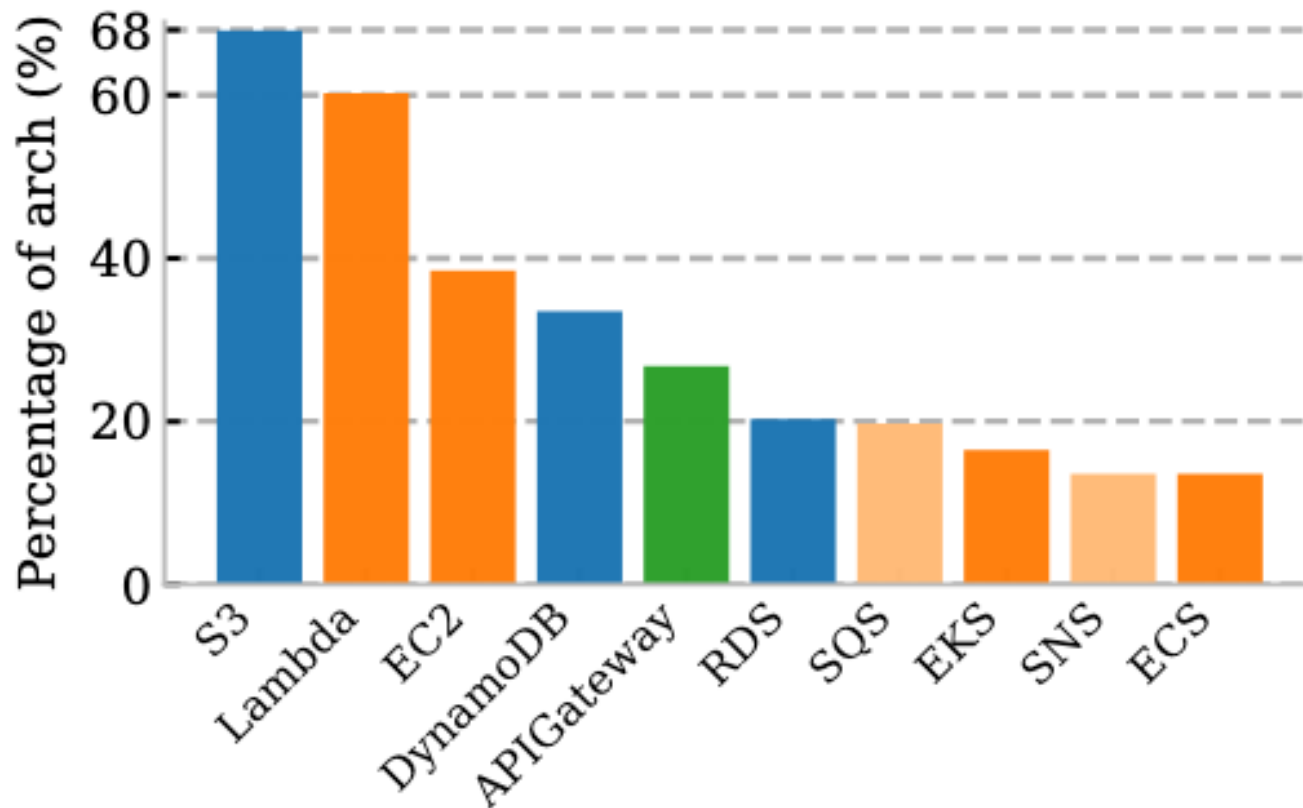
<https://aws.amazon.com/architecture/this-is-my-architecture/>

# Distribution of video release date



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

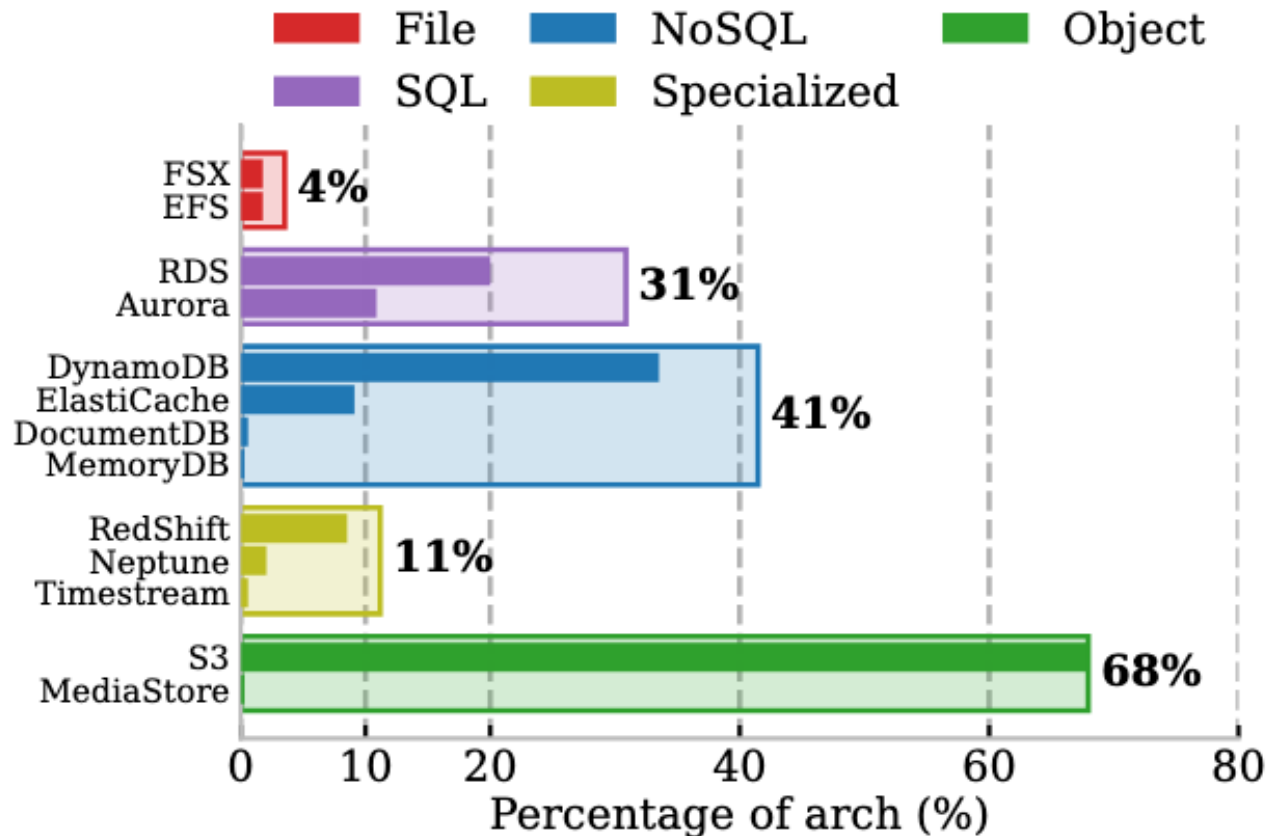
# Popularity of different AWS services



All services including compute and storage

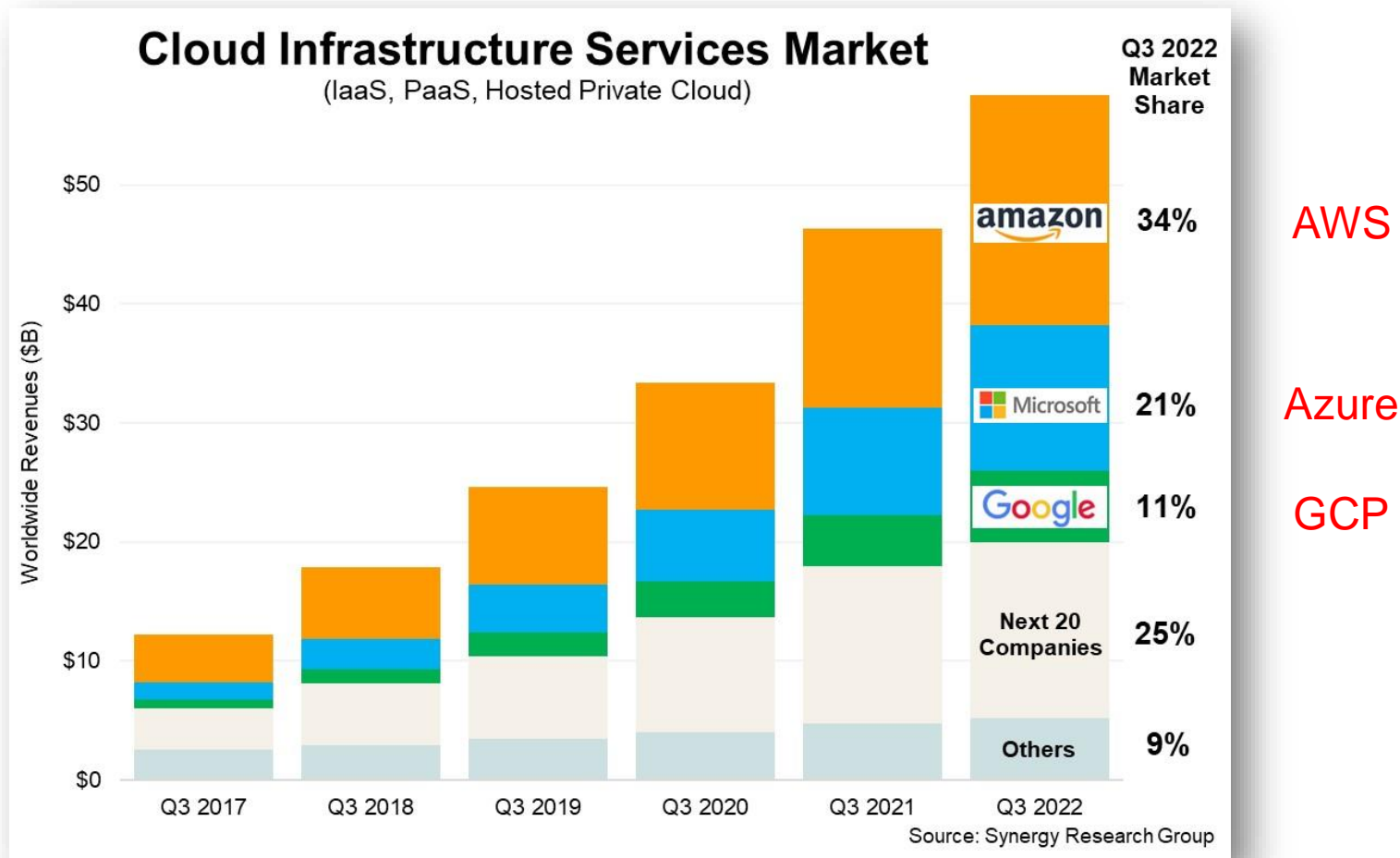
\* Cloudscape: A Study of Storage Services in Modern Cloud Architectures [USENIX FAST 2025]

# Usage of different storage services



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

# Major cloud providers today



<https://www.srgresearch.com/articles/q3-cloud-spending-up-over-11-billion-from-2021-despite-major-headwinds-google-increases-its-market-share>

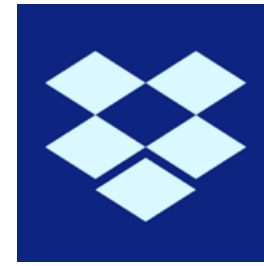
# Lock-in

# 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** (ingress) but **expensive to take it out** (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 costs
- Now still parts of its data services sitting on AWS

# Cloud economics and billing models

# Tenants: Pay-as-you-go?

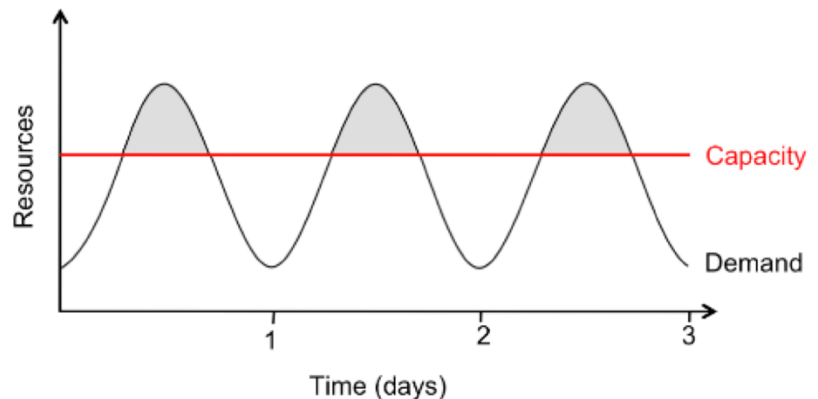
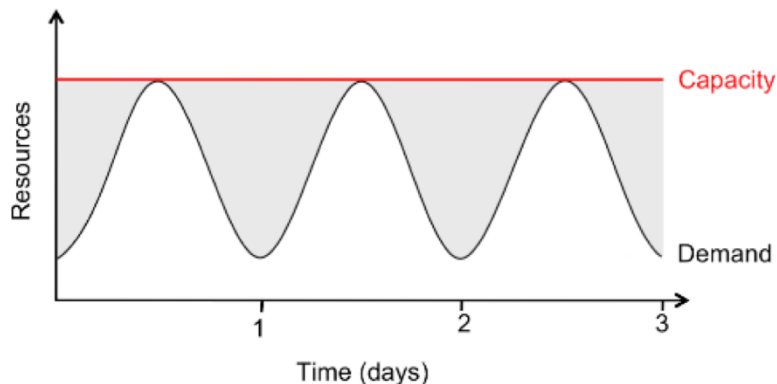
- (**Claimed**) pay-as-you-go pricing
  - Usage-based?
  - Most (compute) services charged per minute
    - Except for Lambda, which is charged per millisecond
  - Storage and network services charged per byte
  - No minimum or upfront fee

# Tenants: Pay-as-you-go?

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

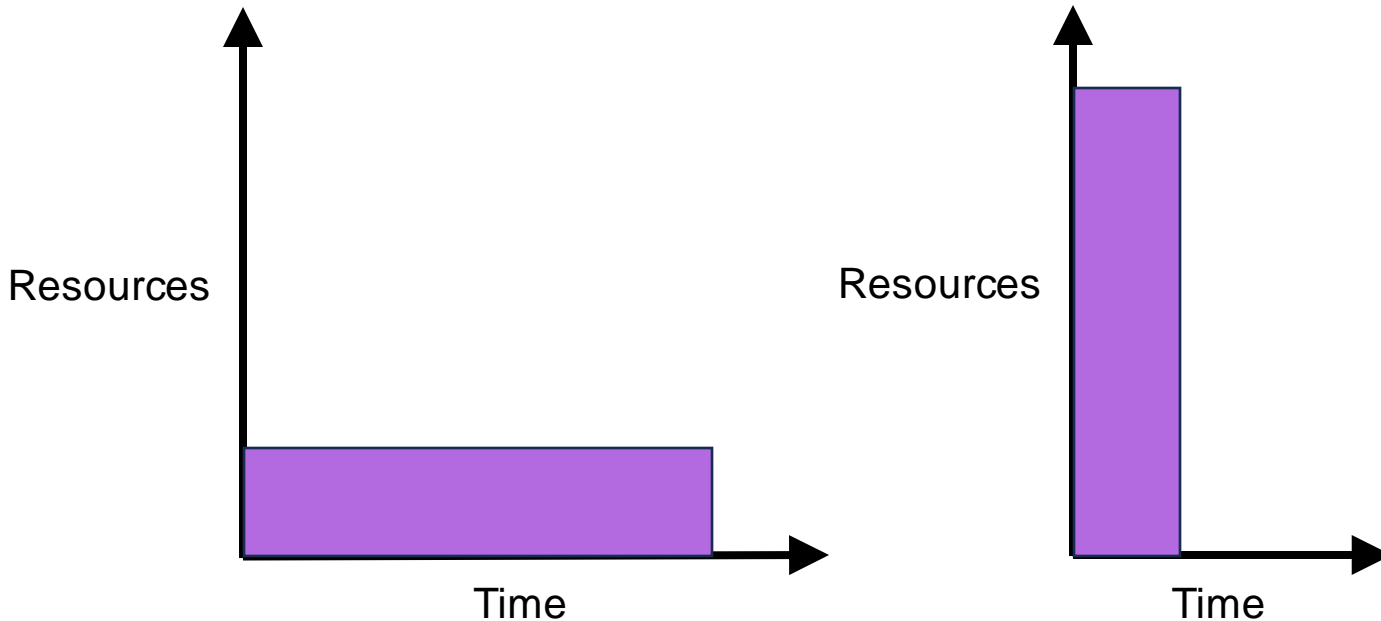
**Q: Is the cloud pricing truly pay-as-you-go?**

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



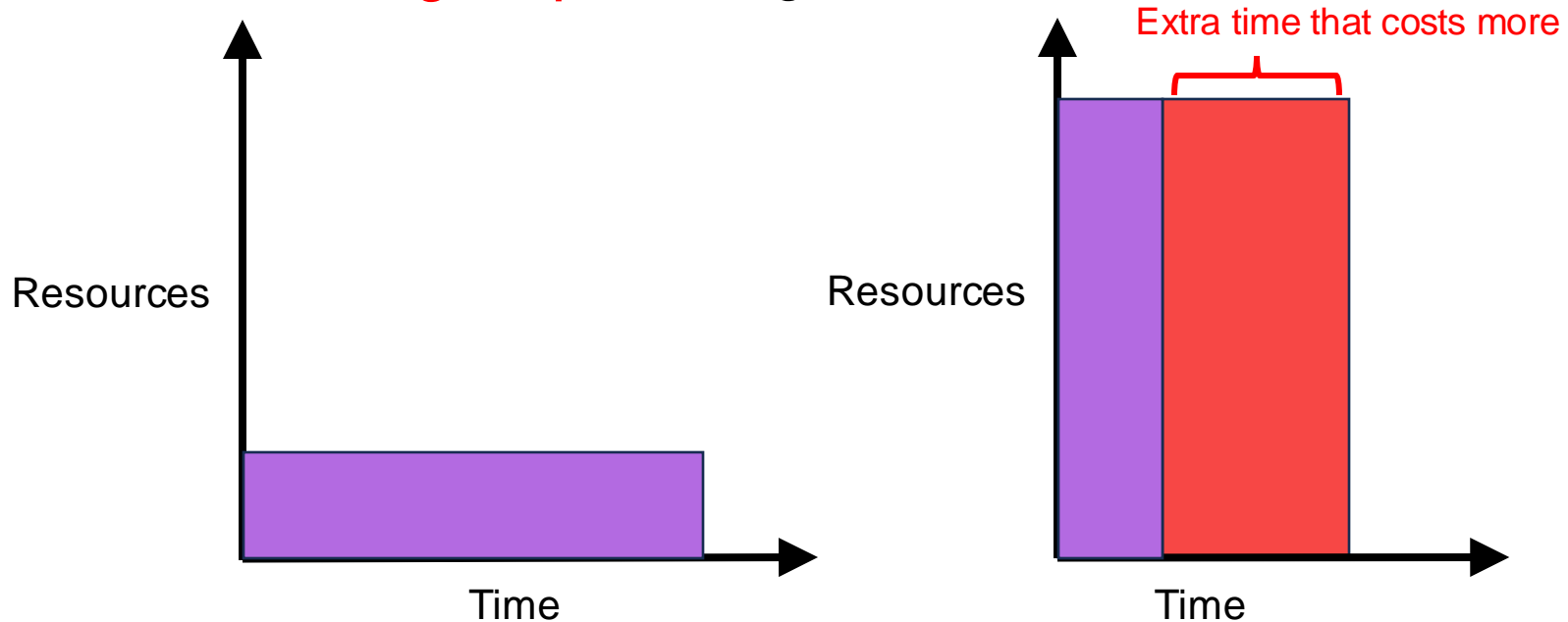
# Tenants: Scalability gained?

- (**Ideally**) Linear scalability & perfect elasticity
  - Using 1000 servers for 1 hour costs the same as 1 server for 1000 hours
  - Same price to get a result faster

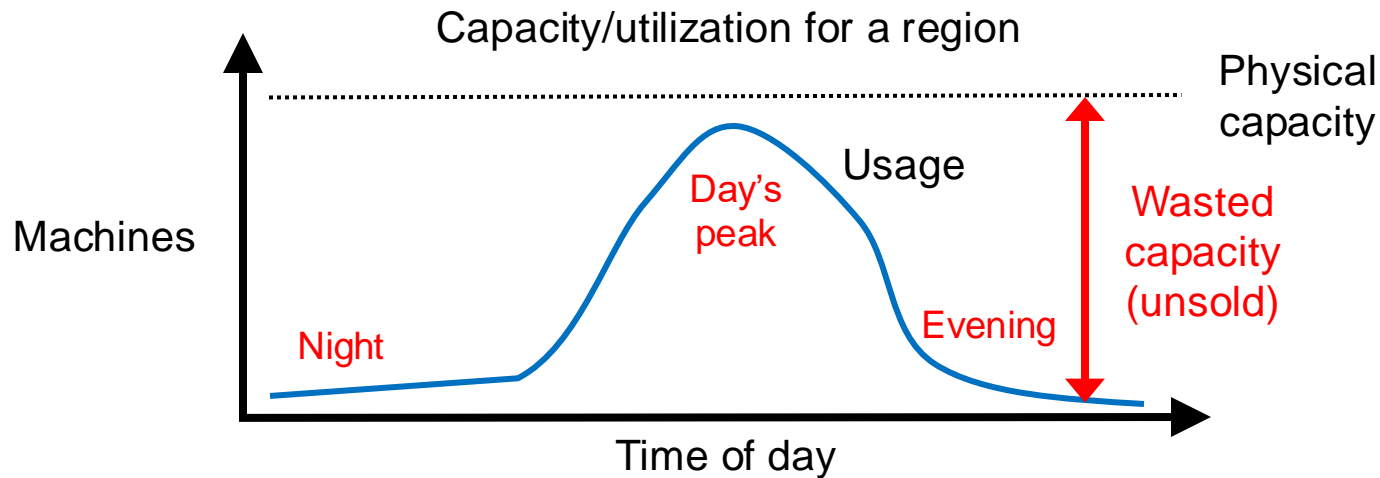


**In practice, it really depends, case by case.**  
**Likely the speedup of the computation is much lower than 1000X!**

- **(In reality)** Scalability is sublinear and VM scaling is slow.
  - Using 1000 servers for 1+N hour costs **N times** more than 1 server for 1000 hours
  - Often **higher price** to get a result faster

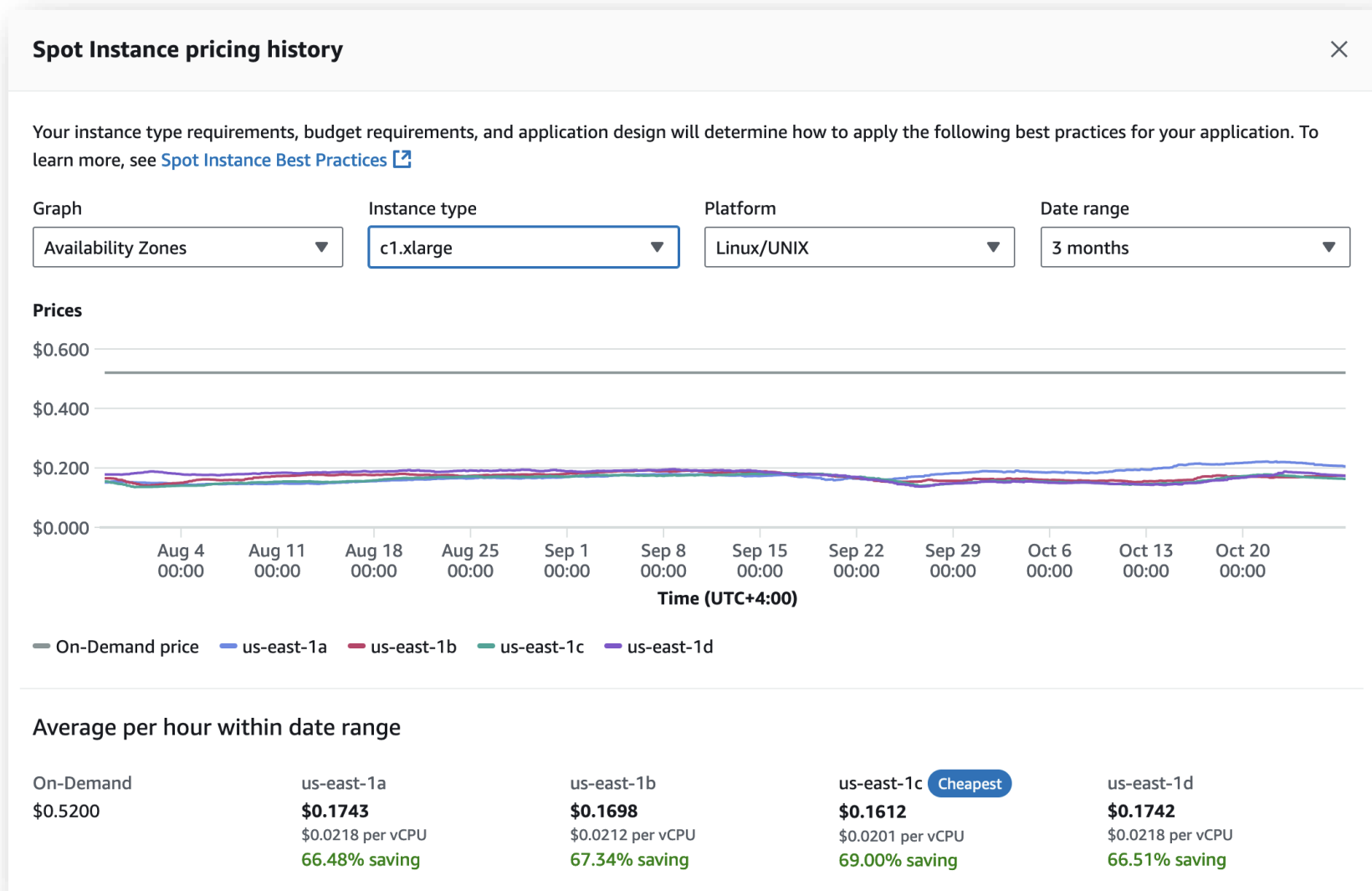


# Providers: On-demand vs. spot instances



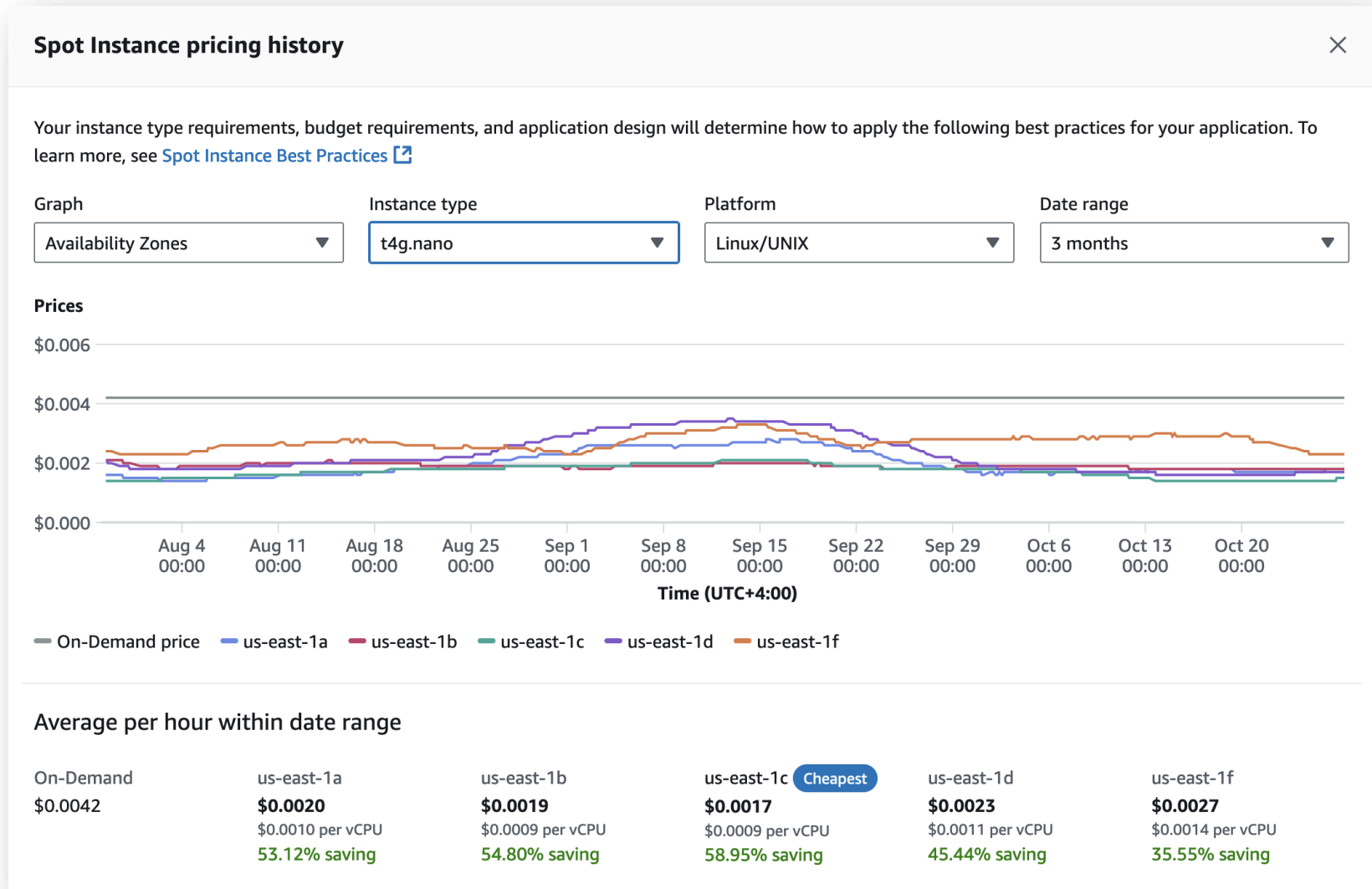
- How to create **incentives** for tenants?
  - Use less at peak time
  - Use more at low times
- Two VM deployment options
  - **On-demand instances**: Constant (high) price. Can generally get a VM. Won't be taken away from you arbitrarily. Used when capacity is needed at specific times.
  - **Spot instances**: Price varies throughout 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)

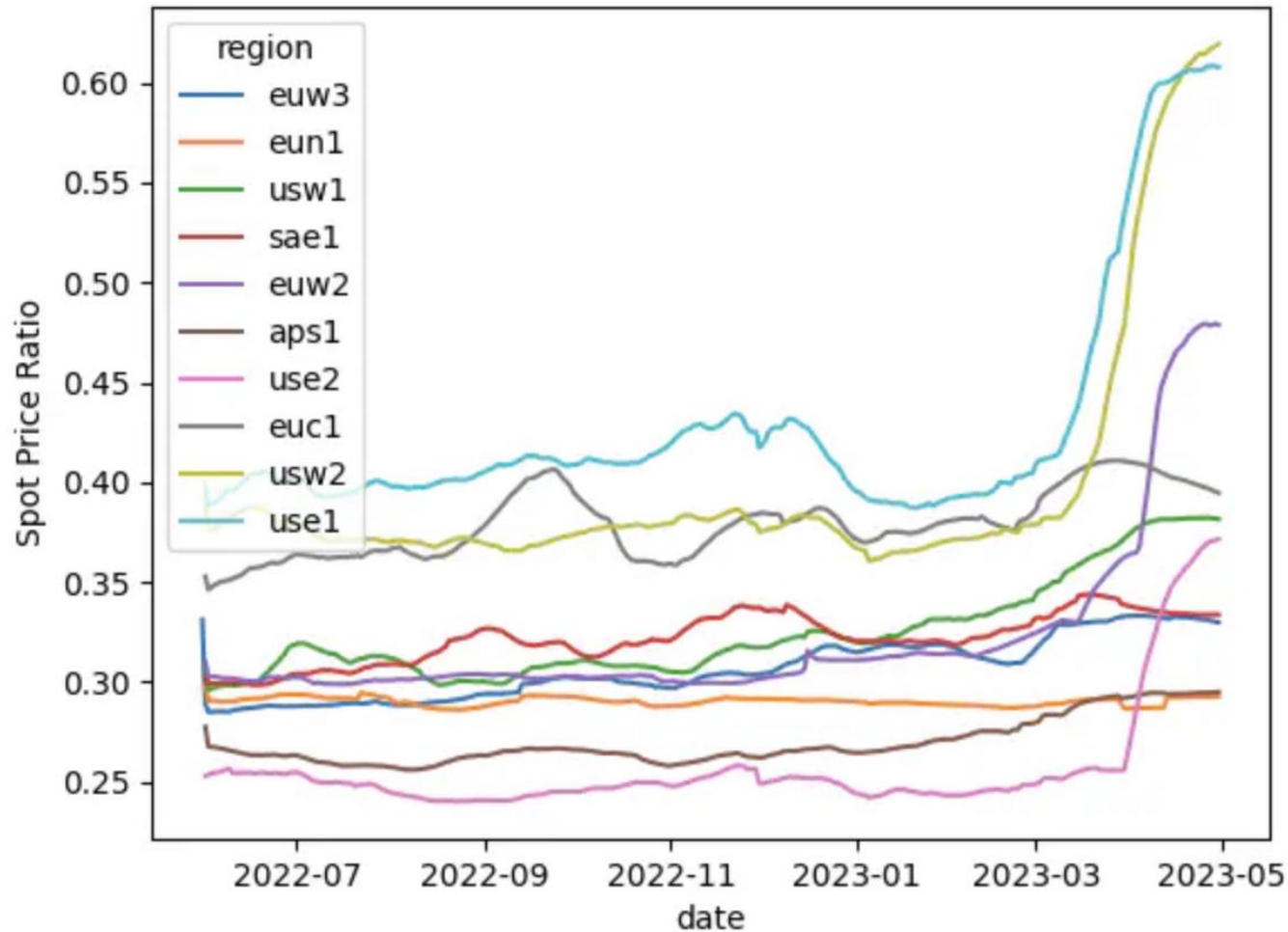




# Spot instance pricing (t4g.nano)

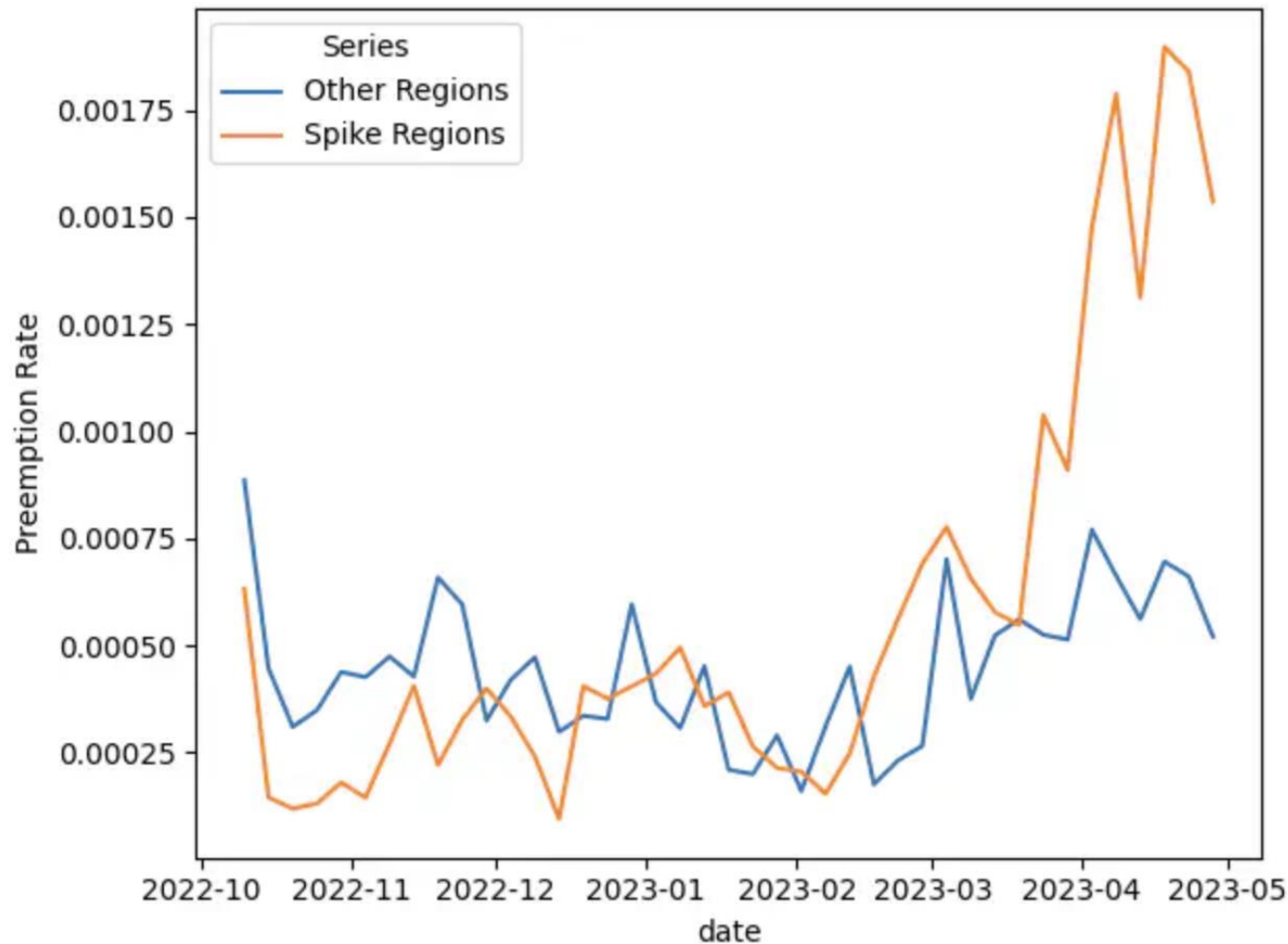


# Mean spot price ratios across regions



<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) ↕

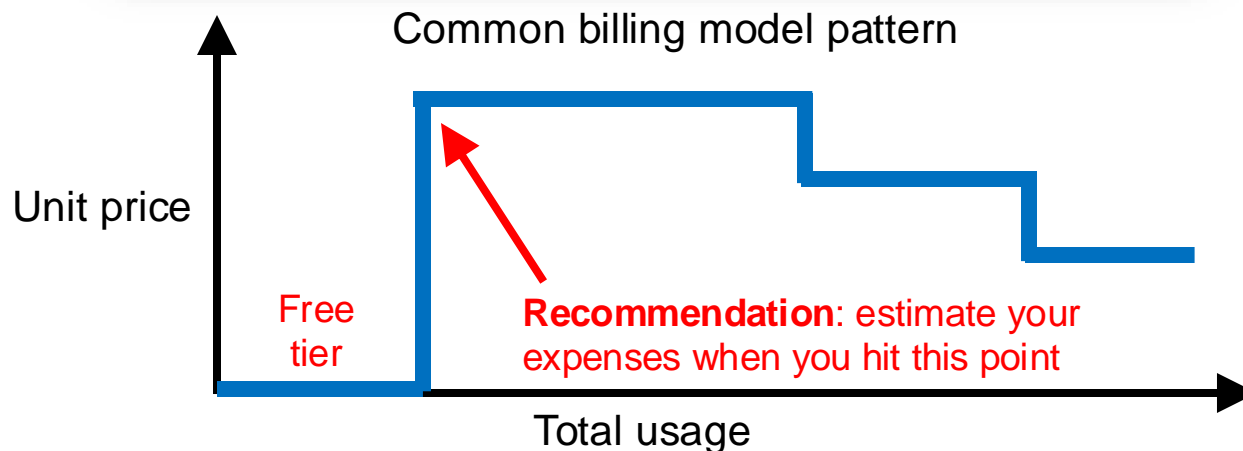
Architecture	Duration
<b>x86 Price</b>	
First 6 Billion GB-seconds / month	\$0.0000166667 for every GB-second
Next 9 Billion GB-seconds / month	\$0.000015 for every GB-second
Over 15 Billion GB-seconds / month	\$0.0000133334 for every GB-second

## 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