Cloud Computing
Fundamentals

Minchen Yu

SDS@CUHK-SZ

Fall 2024

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

SCHOOL OF
DATA SCIENCE
數據科學學院

Outline

- Real-world examples of the cloud
- Definitions of cloud computing
- Key cloud concepts and characteristics
- Deployment scenarios
- Cloud pricing

Cloud: Massive Scale

- Facebook [GigaOM, 2012]
  - 30K in 2009 -> 60K in 2010 -> 180K in 2012
- Microsoft [DC knowledge]
  - > 1 million, 2013
- Google [DC knowledge]
  - > 900 K, 2013
- Alibaba
  - > 1 million, 2022

Datacenter: outside

Copyright: Google

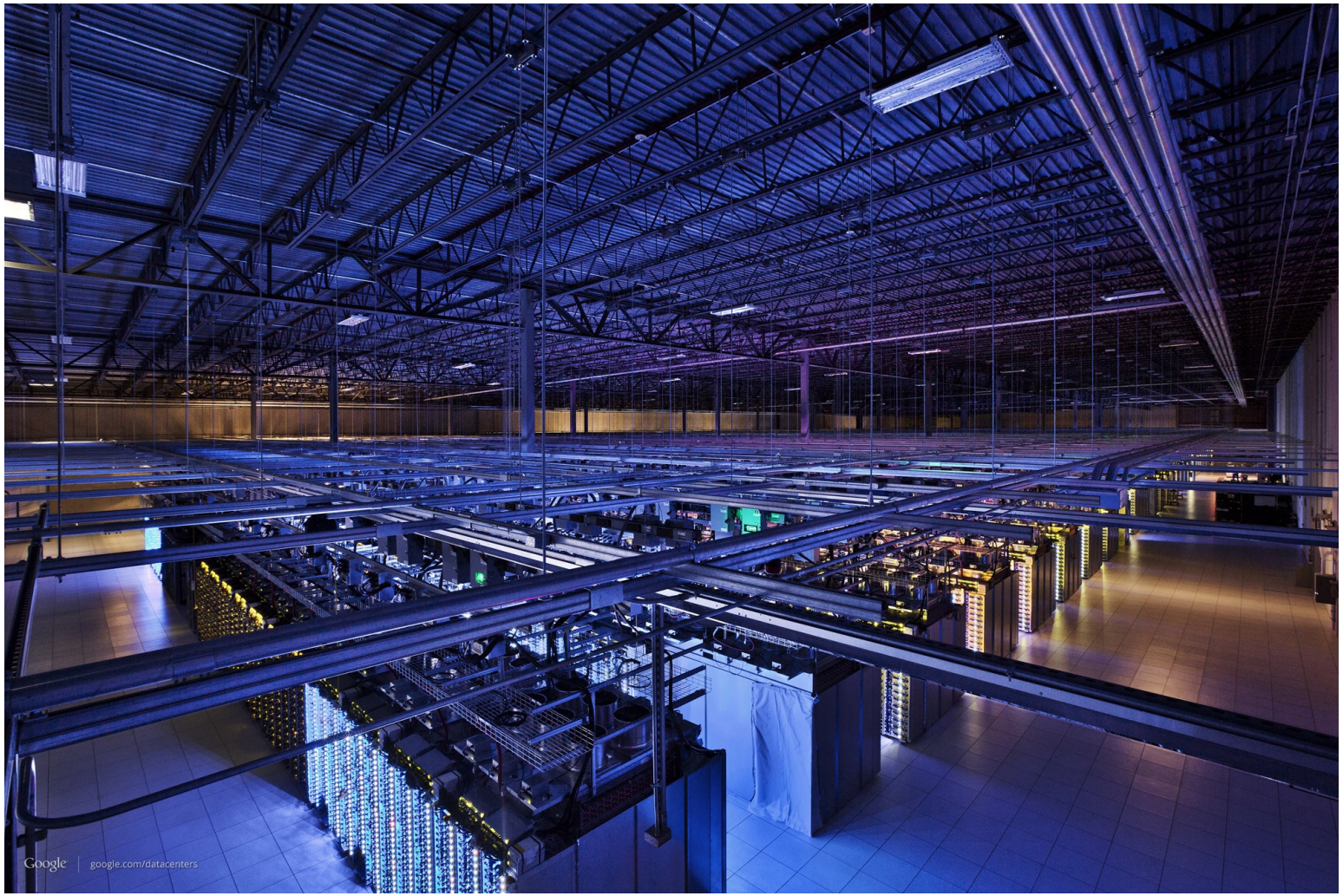
Datacenter: outside

Copyright: Google

A bird’s-eye view of DC

Copyright: Google

Datacenter: inside



Copyright: Google

Server racks

Photo credit: Google

When the nights come…

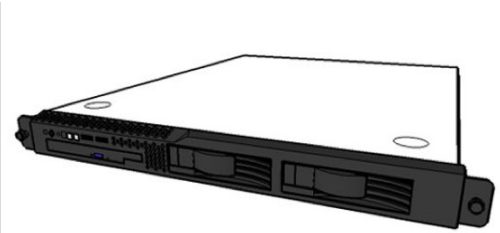
Copyright: Google

Server: inside

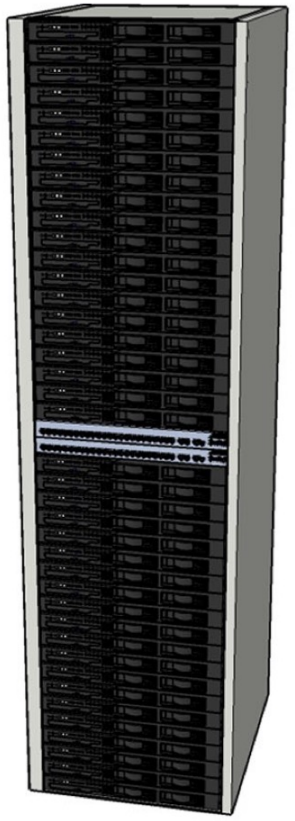
Copyright: xsnet

Server cage

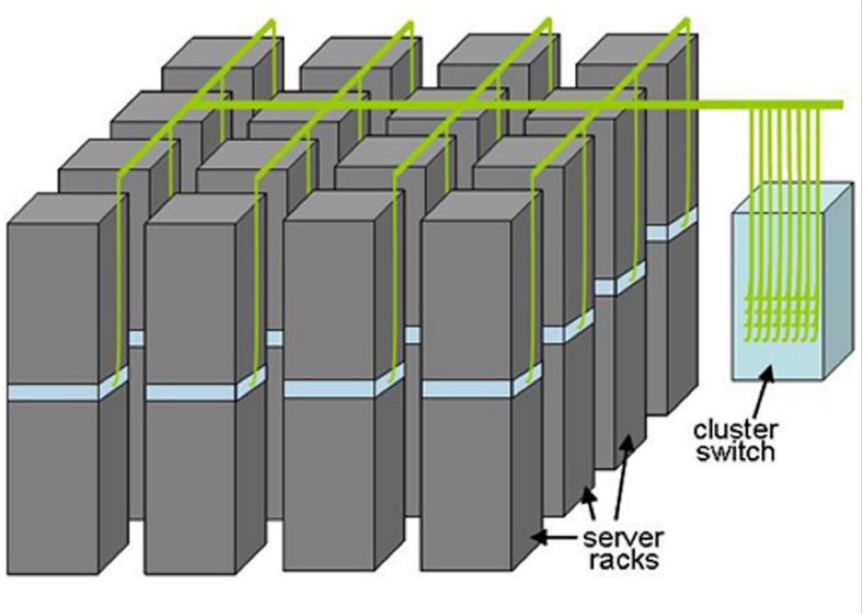
A look into the datacenter



Commodity Server



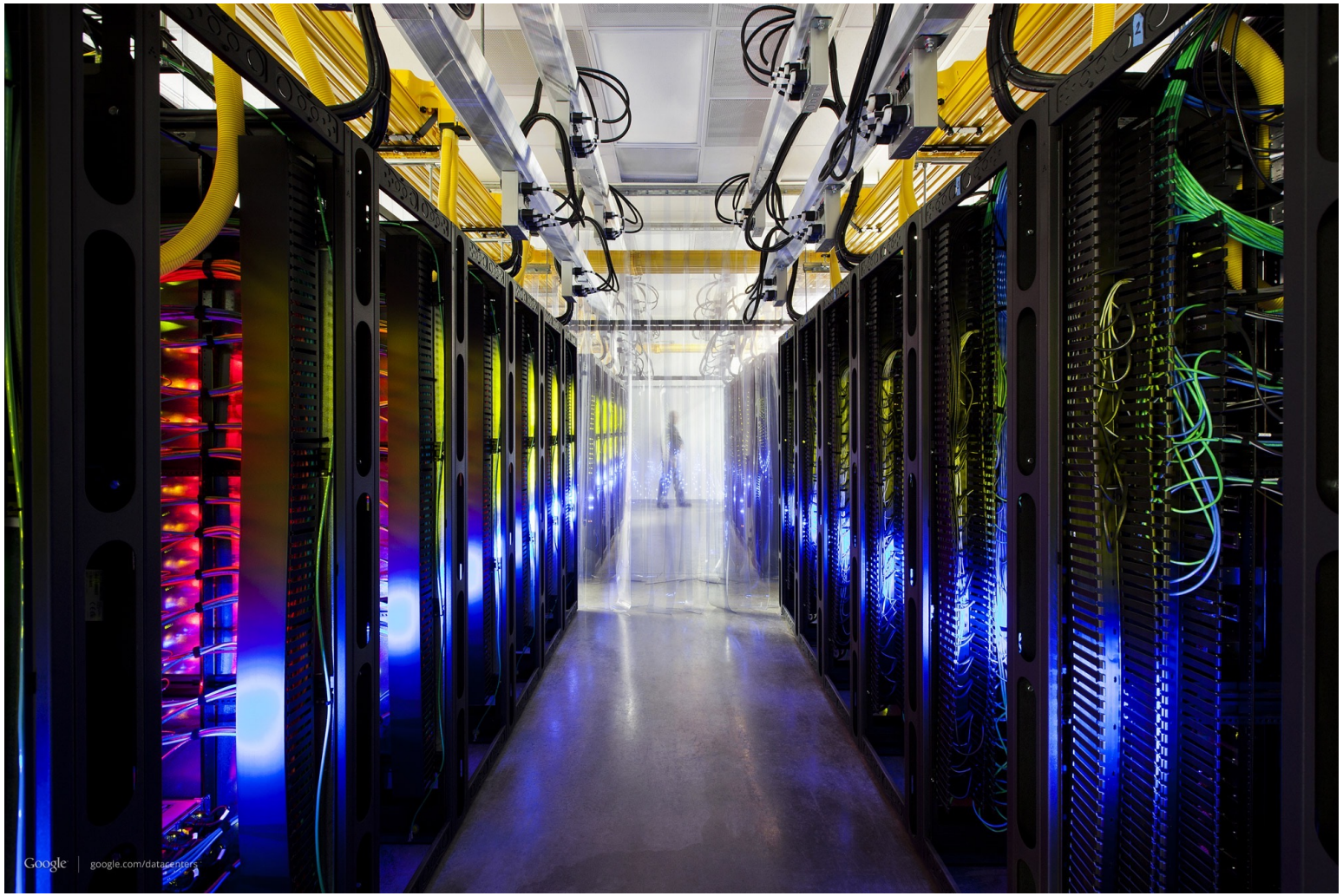
Rack



Cell

Source: L. Barroso et al., “The datacenter as a computer: An introduction to the design of warehouse-scale machines.”

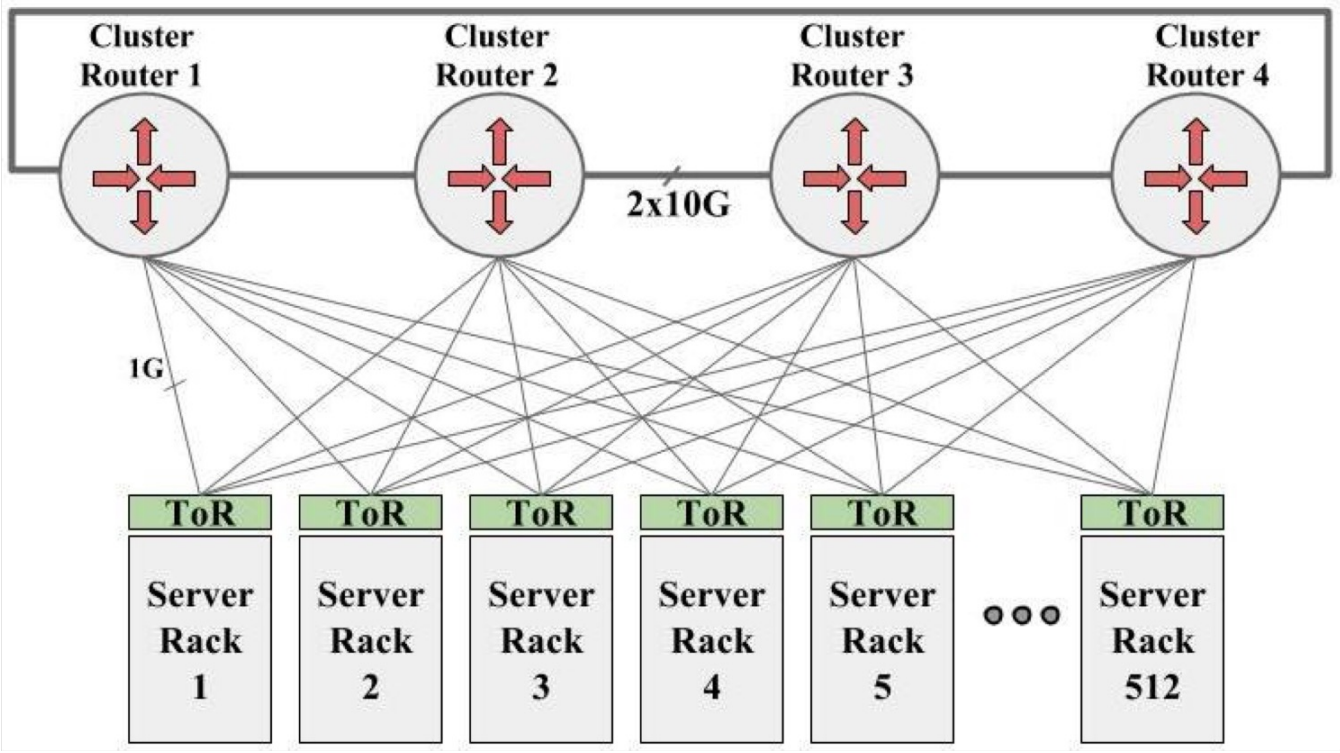
Network room



Copyright: Google

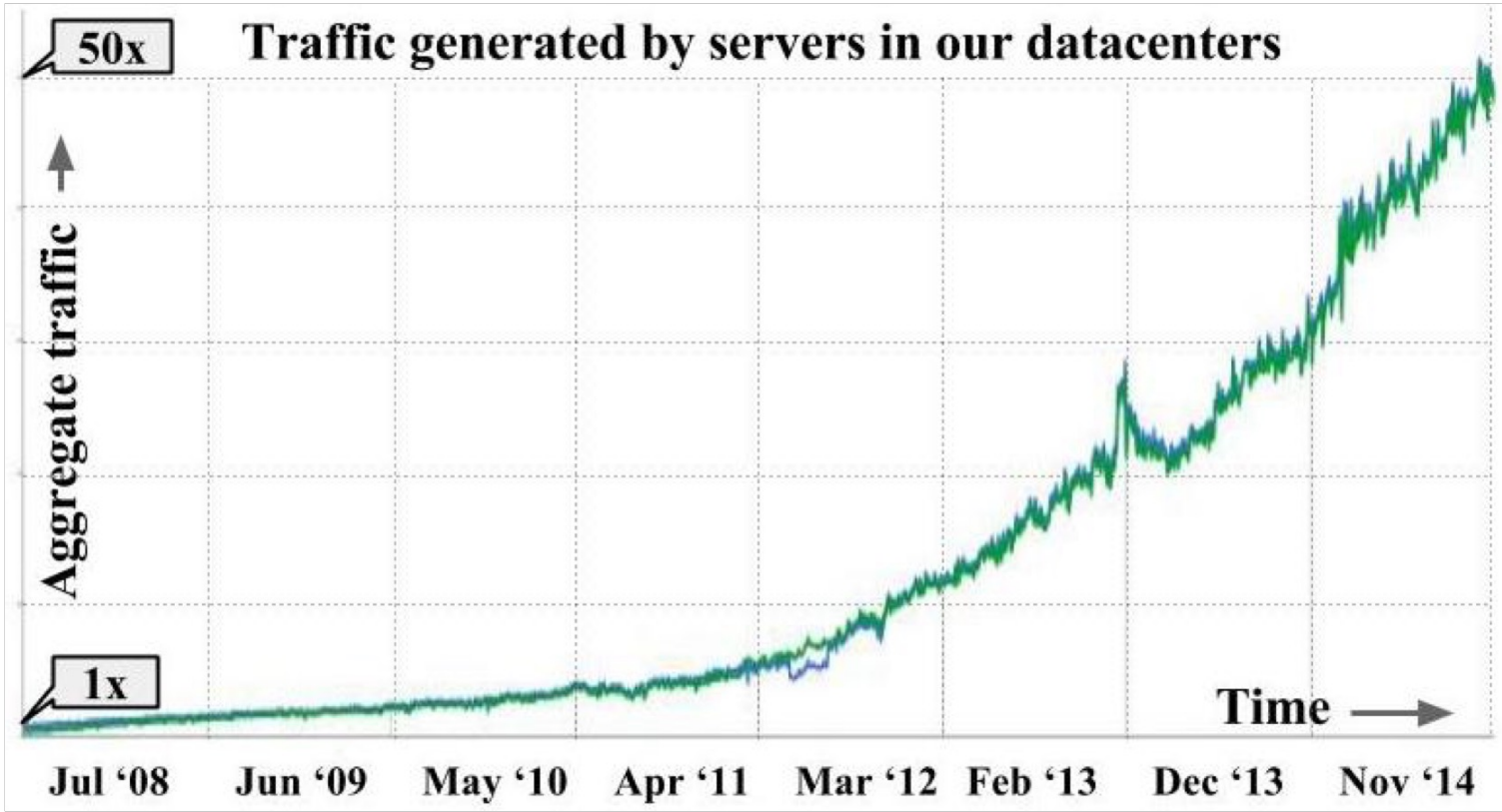
Network for a small-sized cluster

- Back to 2004 when Google has only 20k servers in a datacenter



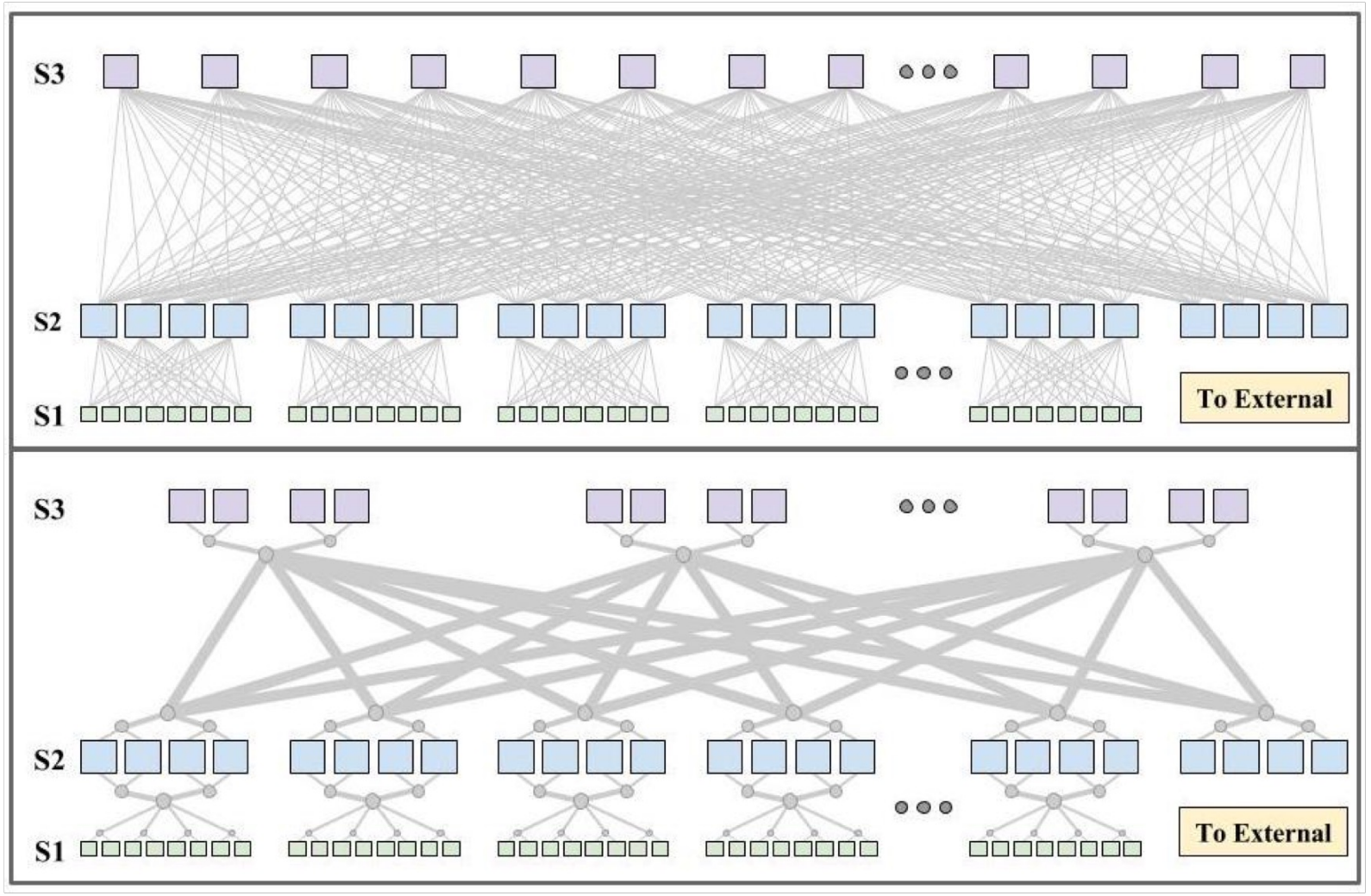
Source: A. Singh et al., "Jupiter rising: A decade of Clos topologies and centralized control in Google's datacenter network," ACM SIGCOMM'15.

Things have changed quite a lot

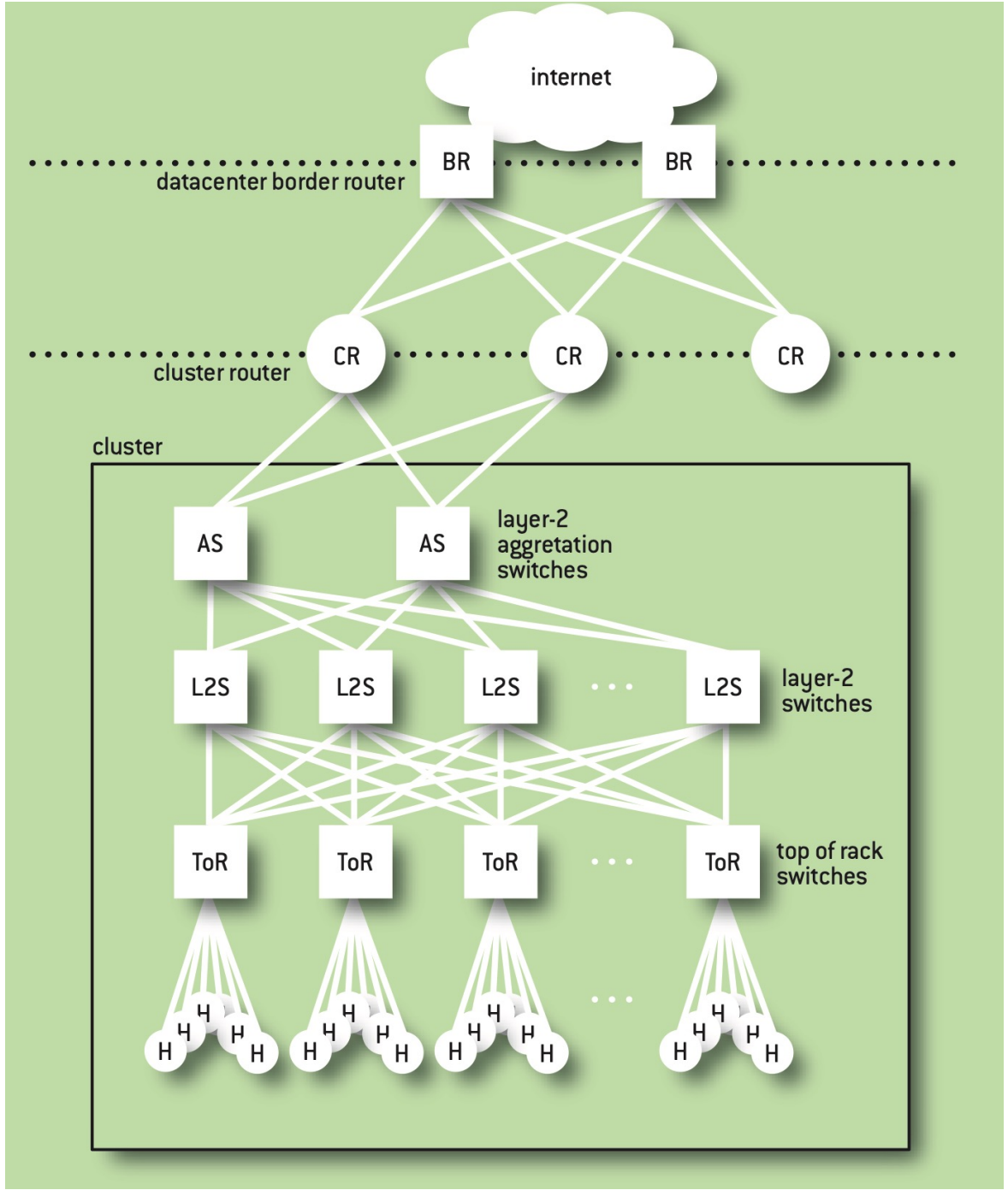


Source: A. Singh et al., "Jupiter rising: A decade of Clos topologies and centralized control in Google's datacenter network," ACM SIGCOMM'15.

When scaled up



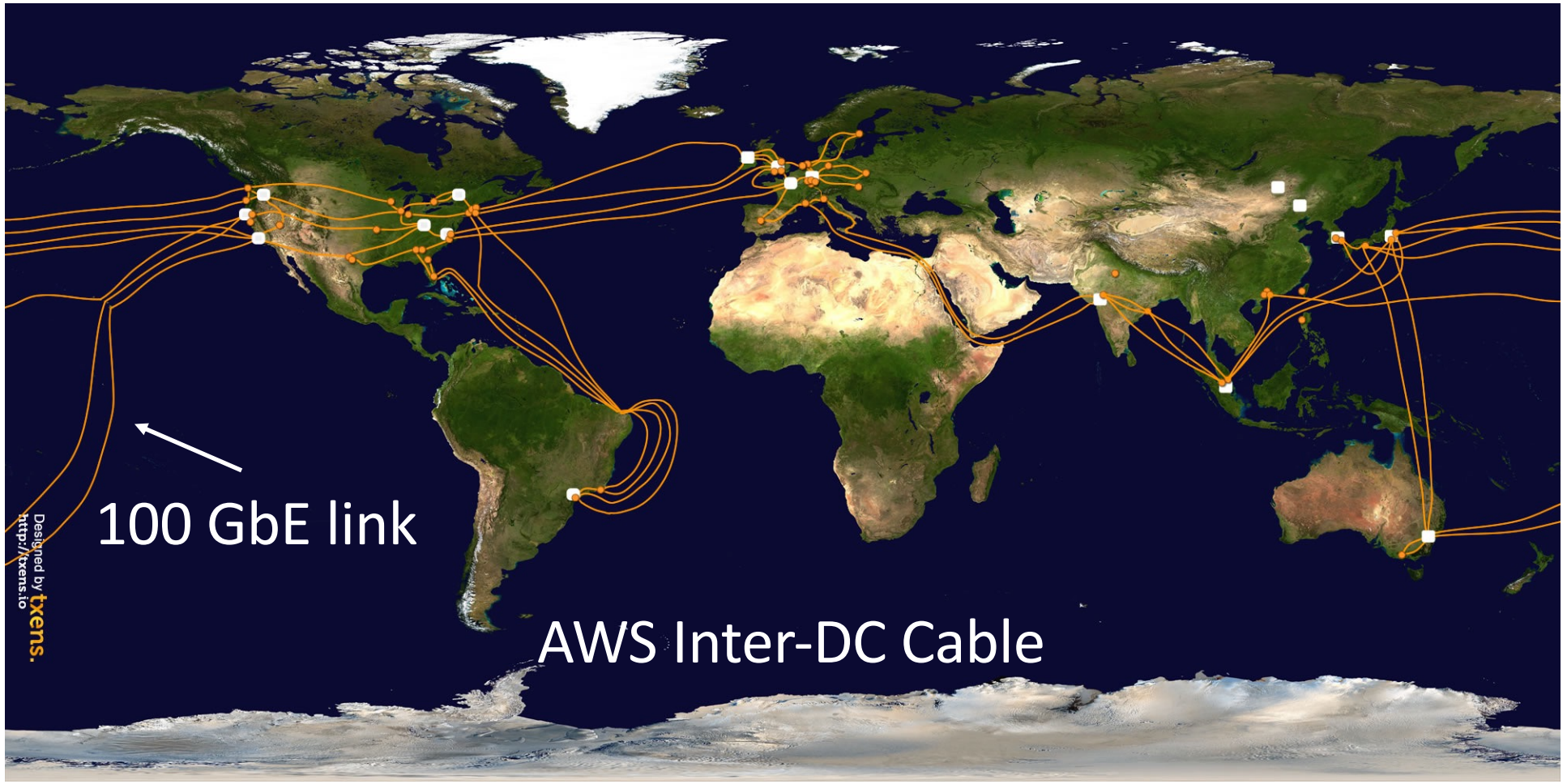
Source: A. Singh et al., “Jupiter rising: A decade of Clos topologies and centralized control in Google’s datacenter network,” ACM SIGCOMM’15.



Tree-like DC Network

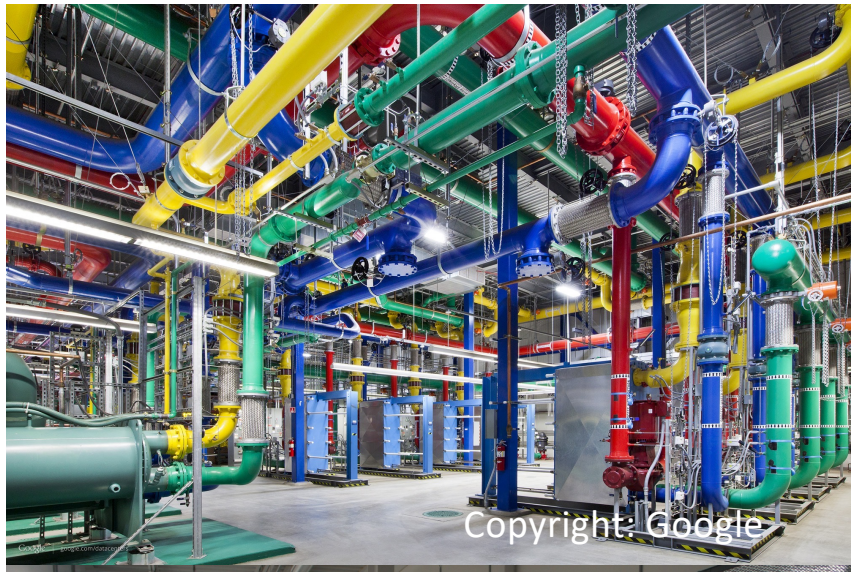
Source: http://queue.acm.org/detail.cfm?id=2208919

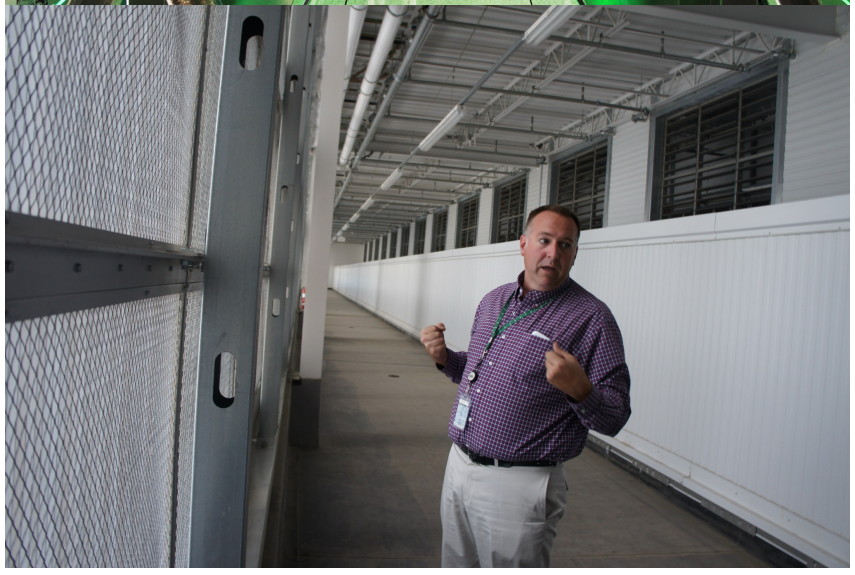
Inter-DC WAN



Source: txens, http://txens.io/index.php/2016/12/16/a-deep-dive-into-the-aws-network/

Cooling



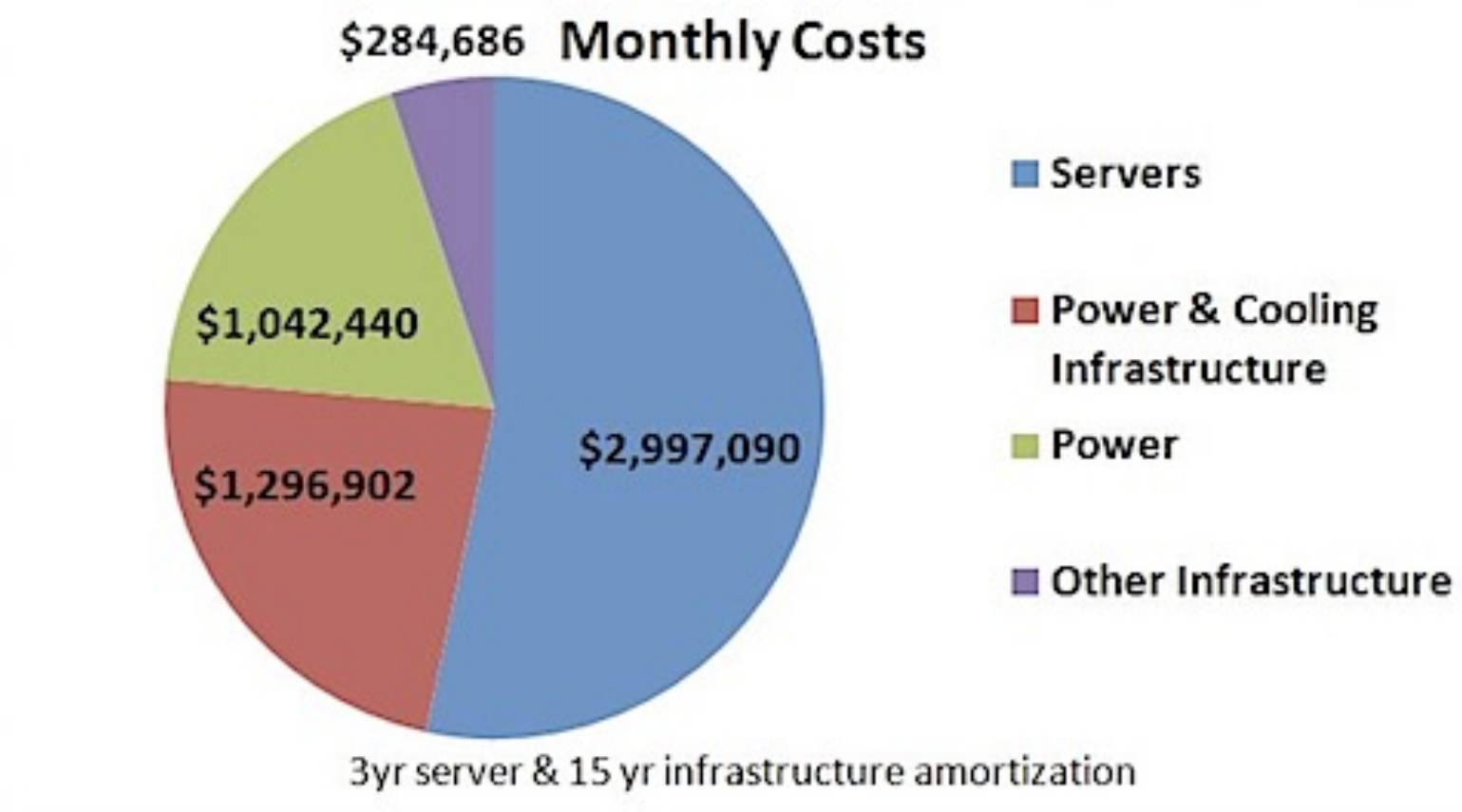


Power

Copyright: GigaOM

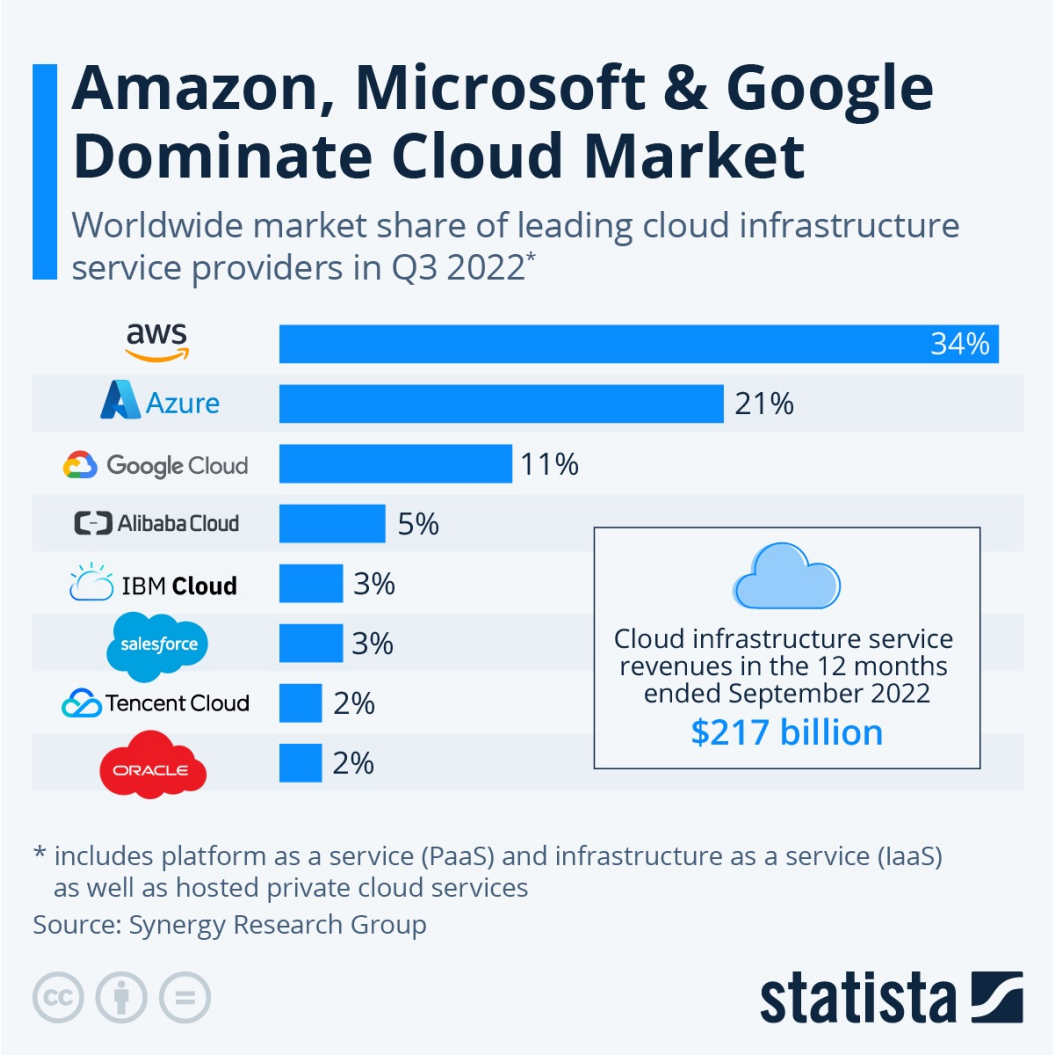
Copyright: Nation of Change

Cost Structure

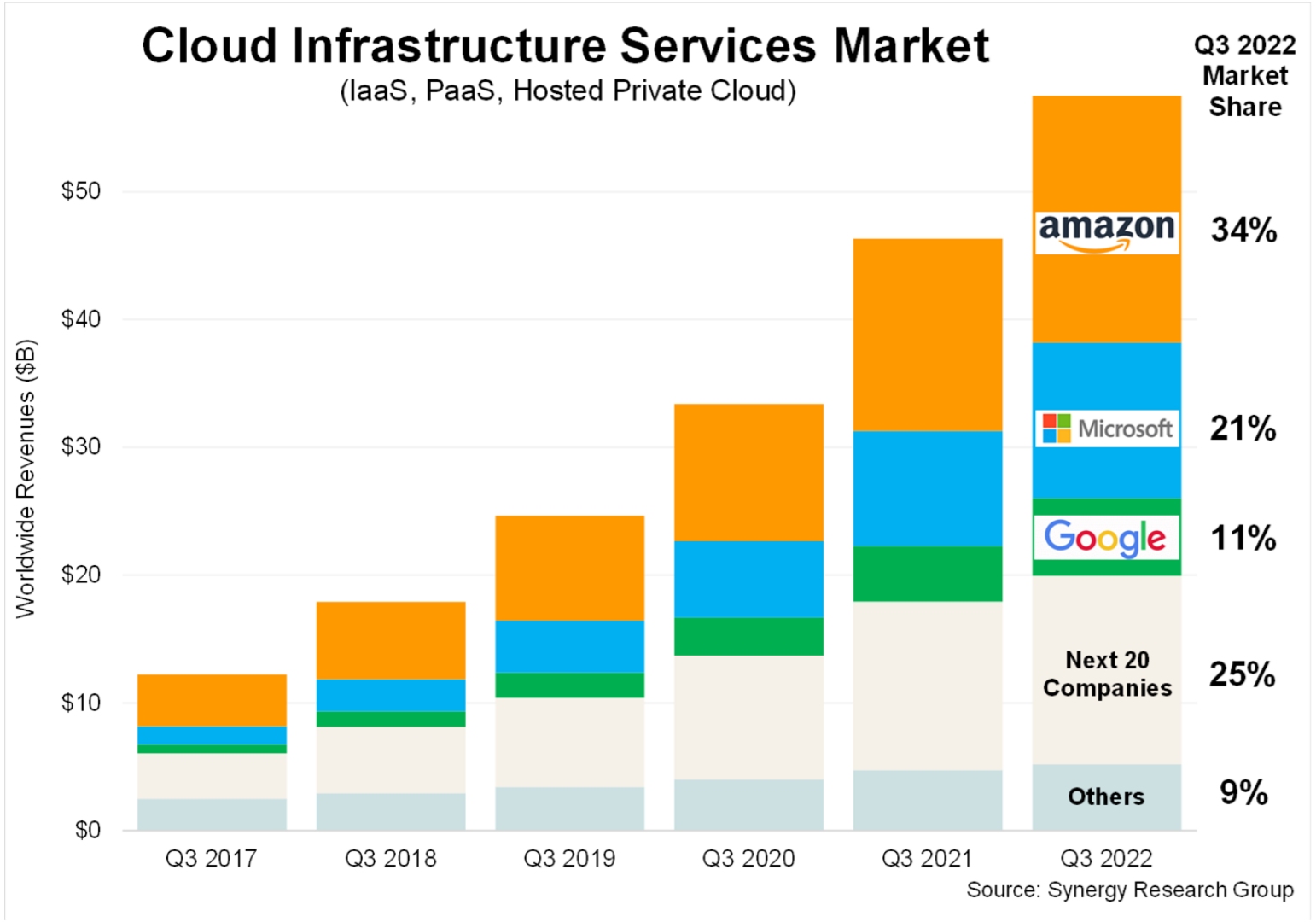


Source: James Hamilton, http://perspectives.mvdirona.com/2008/11/cost-of-power-in-large-scale-data-centers/

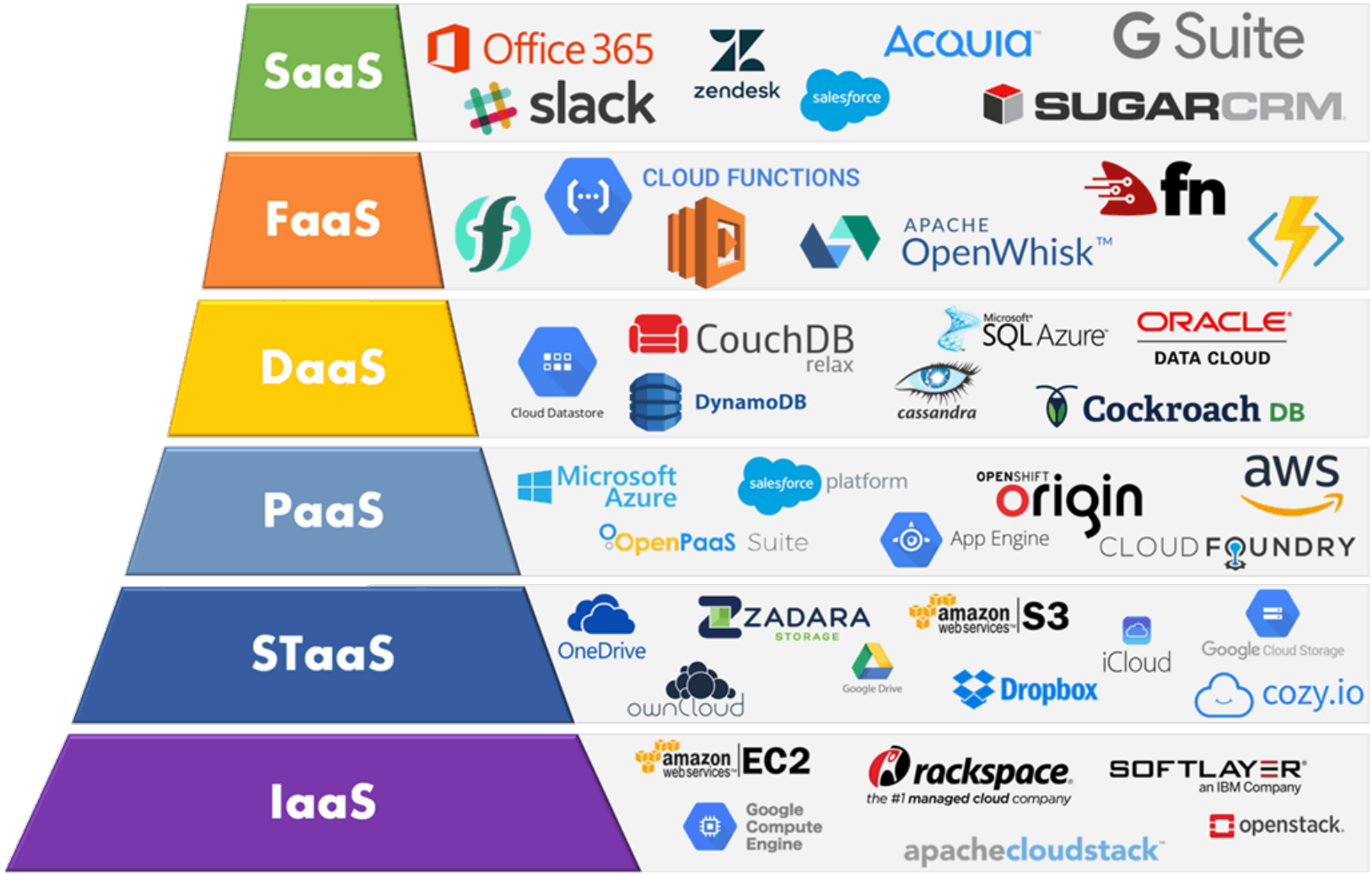
Dominant Cloud Providers



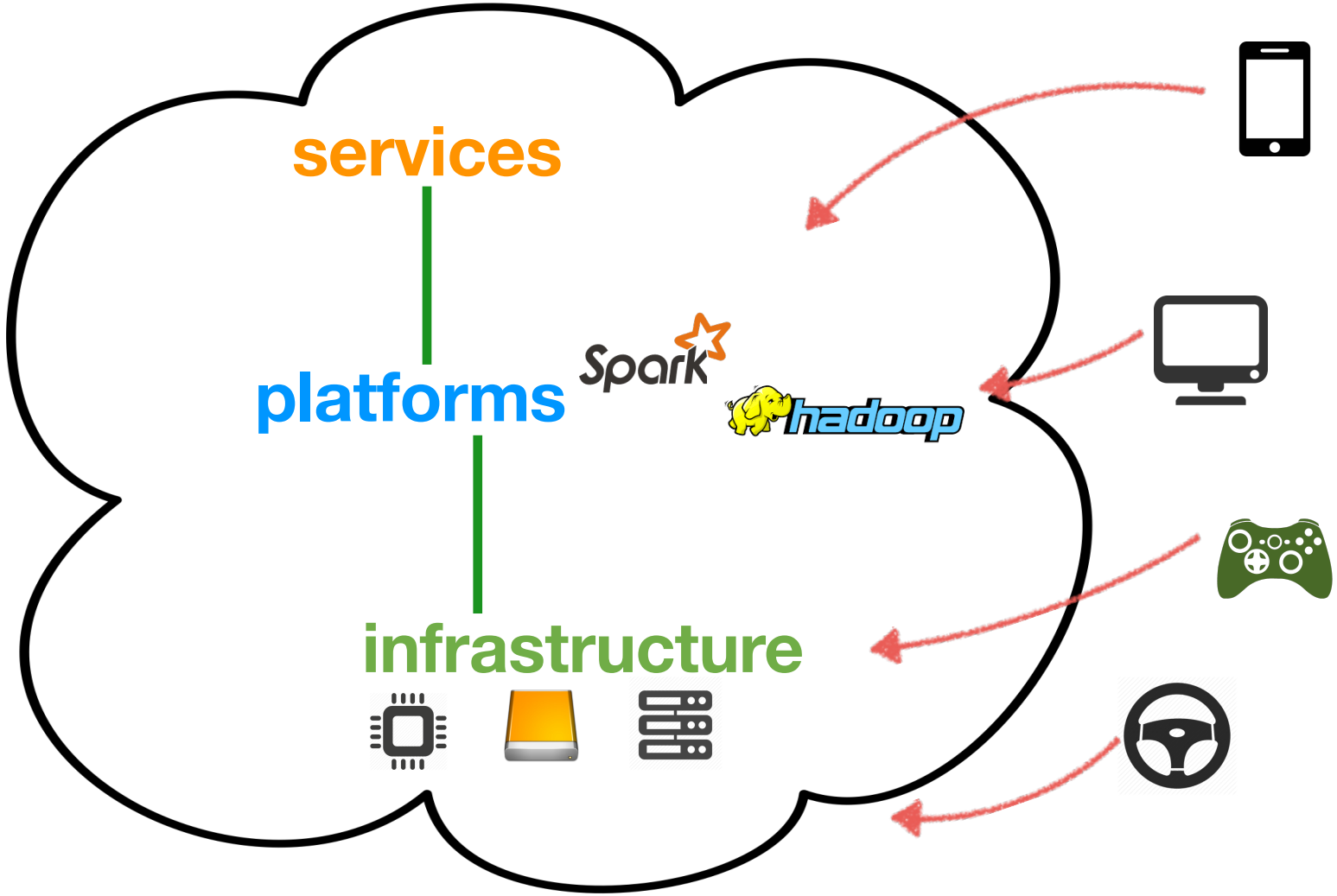
The Booming Cloud Market



Cloud-based services



So what is a cloud?



A definition

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

National Institute of Standards and Technology (NIST), U.S. Department of Commerce

On-demand computing delivered over the Internet

Utility Computing

- Computing as the 5th utility (after water, electricity, gas, and telephony)
  - Applications and computing resources delivered as a service over the Internet
  - Pay-as-you-go
- Provided by the hardwares and system softwares hosted in the datacenters

Visions

- The illusion of infinite computing resources available on demand
- The elimination of an up-front commitment by Cloud users
- The ability to pay for use of computing resources on a short-term basis as needed

Amazon EC2

https://youtu.be/TsRBftzZsQo

1 instance runs 1000 h = 1000 instances run 1 h

Revolutionary!

Suppose you open a startup and need 100 servers

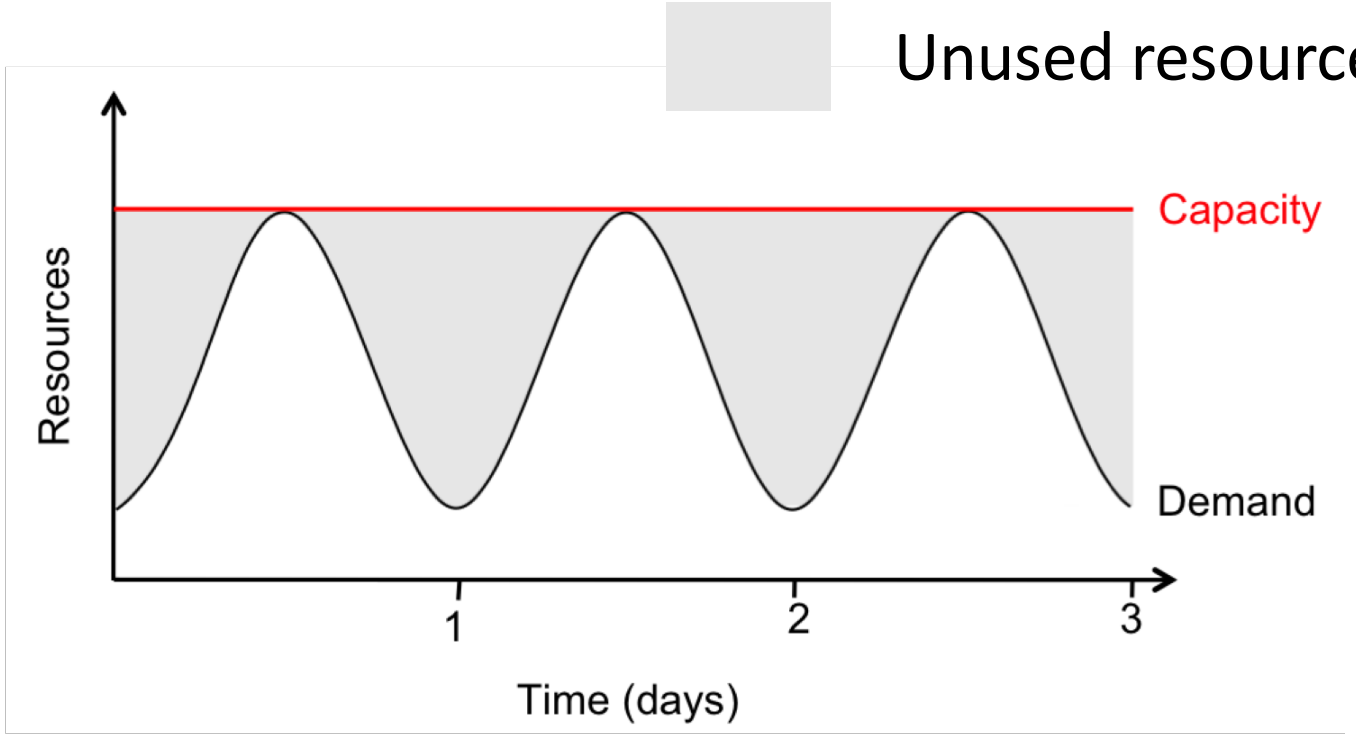
What would you do traditionally, in a pre-cloud era?

With cloud

- A consumer can unilaterally provision computing capabilities, such as servers and network storage, as needed automatically without requiring human interaction with each service provider.
- Cloud computing makes the underlying technology, beyond the user device, almost invisible
- Always-on services
- **Advantages for consumers**: flexible, minimal overhead, quick and easy

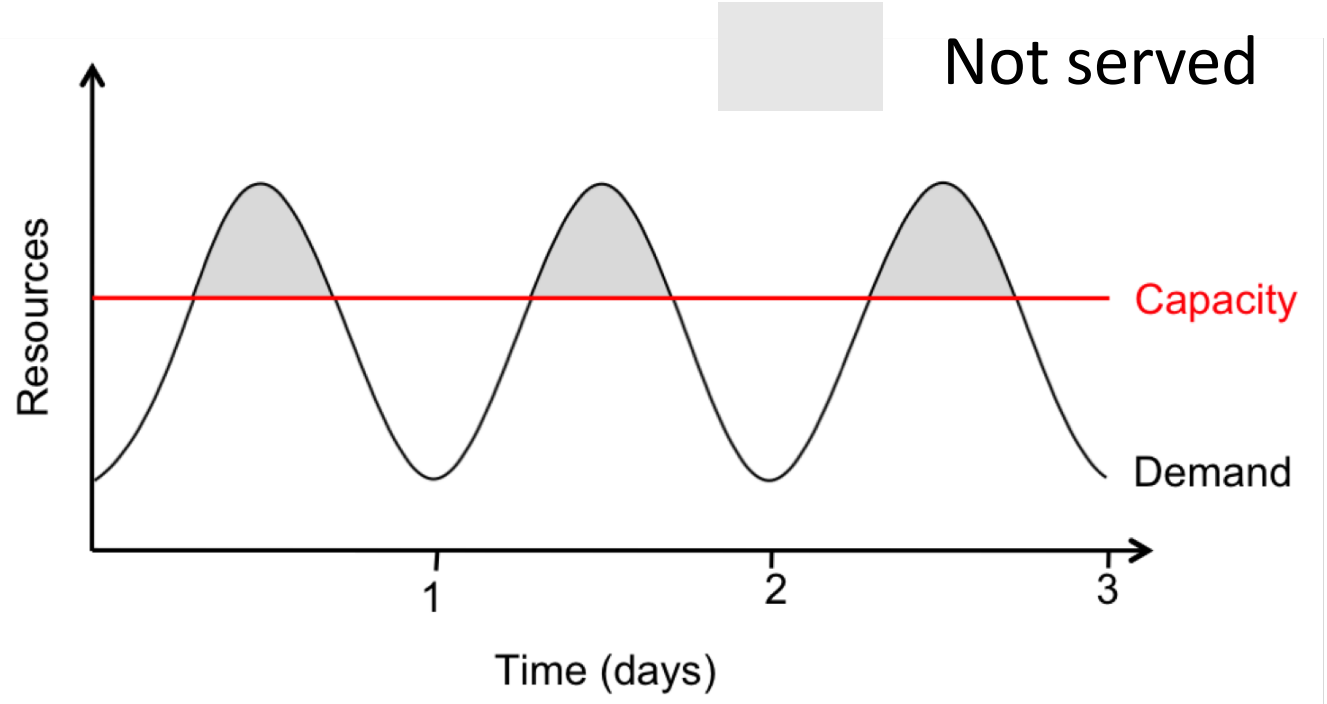
The demands are ***elastic*** — the 100 servers are only needed in peak time

Provisioning for peak load



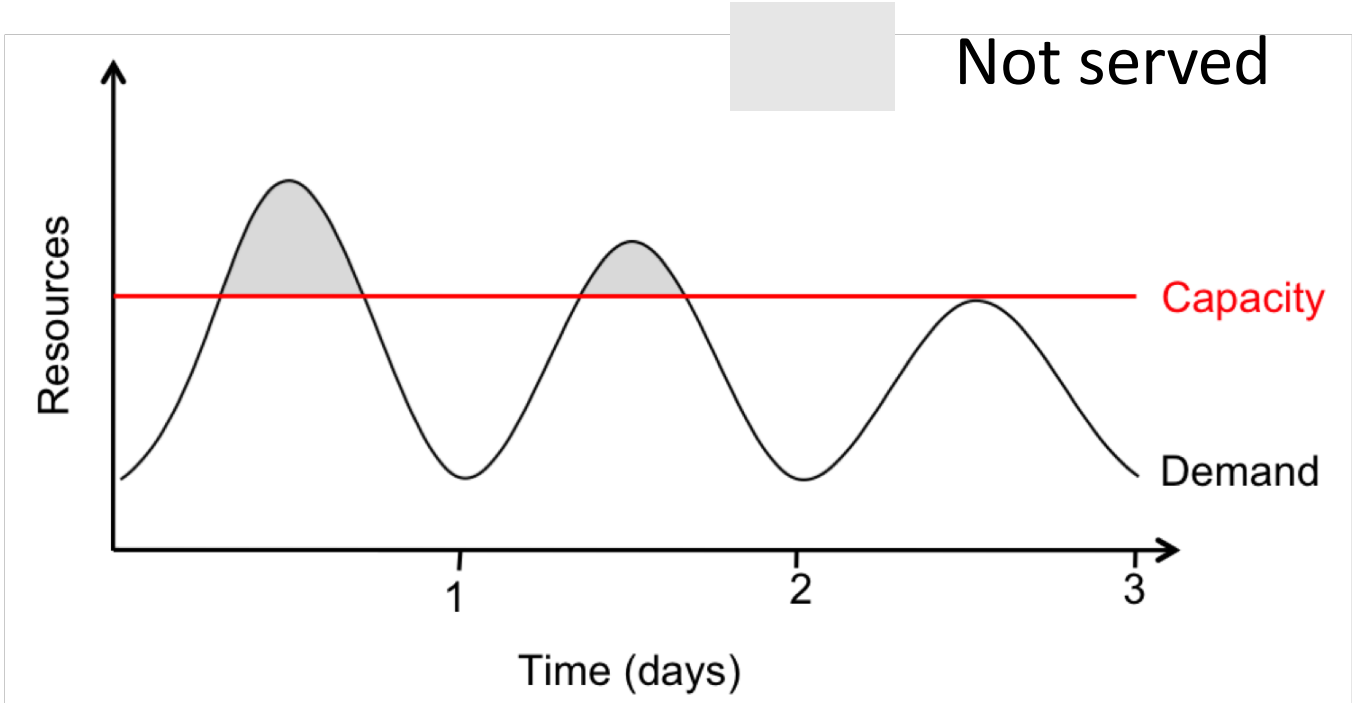
Source: Ambrust et al., “Above the clouds: A Berkeley’s view of Cloud Computing.”

Underprovisioning



Source: Ambrust et al., “Above the clouds: A Berkeley’s view of Cloud Computing.”

Underprovisioning

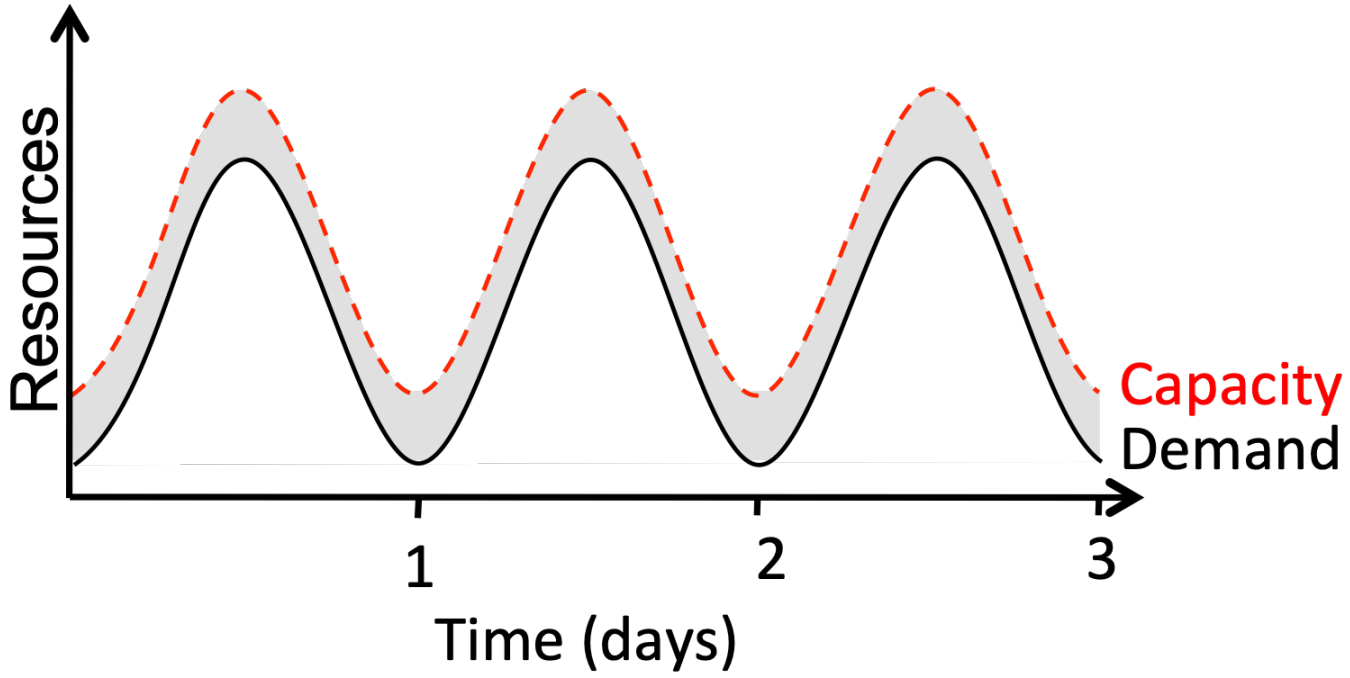


Source: Ambrust et al., “Above the clouds: A Berkeley’s view of Cloud Computing.”

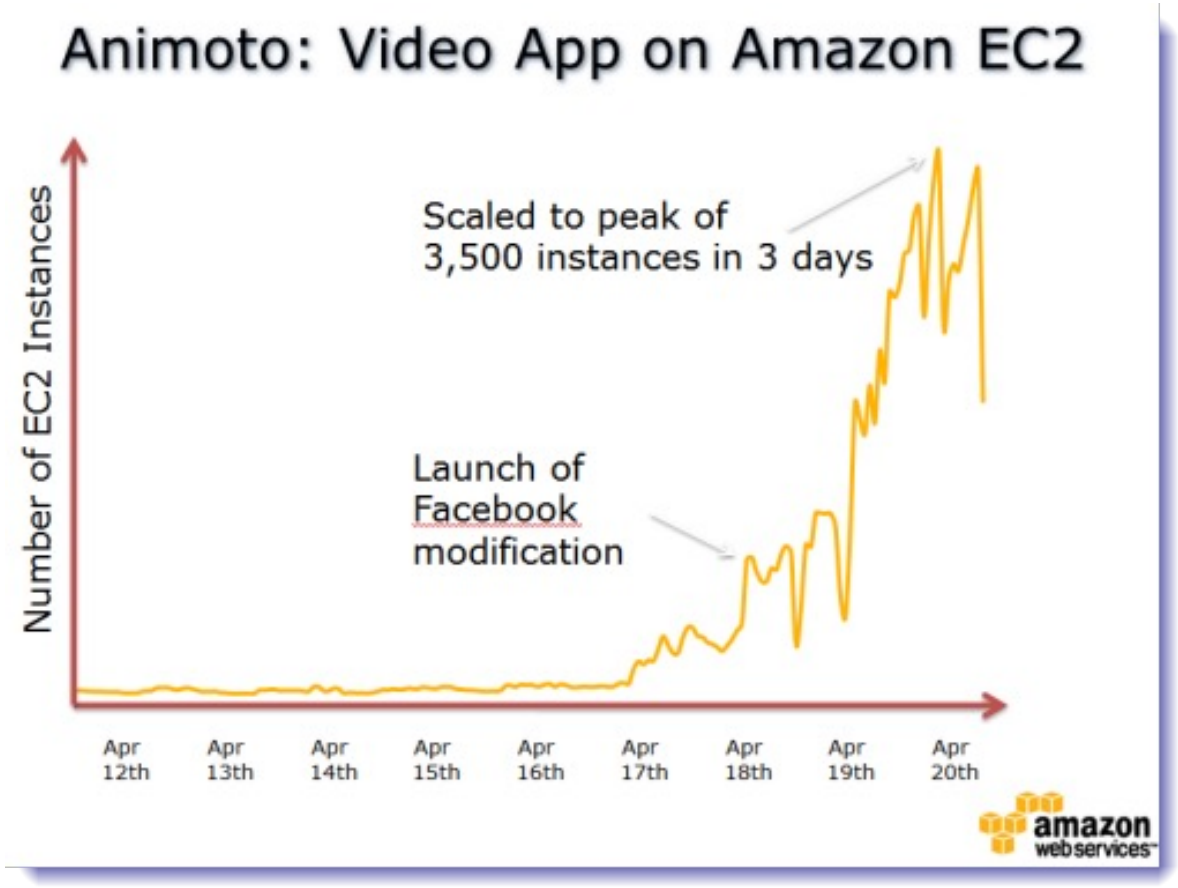
Let’s do it in cloud!

Cloud provisioning on demand

- Pay for what you used



An Animoto Case Study



https://youtu.be/VwDS6MexKEo

Copyright: AWS & Animoto

Cloud Economics: does it make sense?

Shall I move to the Cloud?

- Profit from cloud >= profit from in-house infrastructures

$$\begin{array}{c} \text{UserHours}_{\text{cloud}} \times (\text{revenue} - \text{Cost}_{\text{cloud}}) \\ \text{II} \qquad\qquad\qquad \text{II} \\ \geq \text{UserHours}_{\text{datacenter}} \times (\text{revenue} - \frac{\text{Cost}_{\text{datacenter}}}{\text{Utilization}}) \end{array}$$

$\text{Cost}_{\text{cloud}} << \text{Cost}_{\text{datacenter}} / \text{Util}$

Source: Ambrust et al., “Above the clouds: A Berkeley’s view of Cloud Computing.”

What about the cloud provider?

Resource pooling

- From the provider's perspective

Resource pooling

- The provider's resources are pooled to serve consumers using a multi-tenant model
  - different physical and virtual resources dynamically allocated according to consumer demand
  - creates an illusion of an infinite amount of resources

Resource pooling

- **Location independence:**
  - the customer generally has NO control or knowledge over the exact location of the provided resources
  - but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter)

Resource pooling enables high utilization

Economy of scale

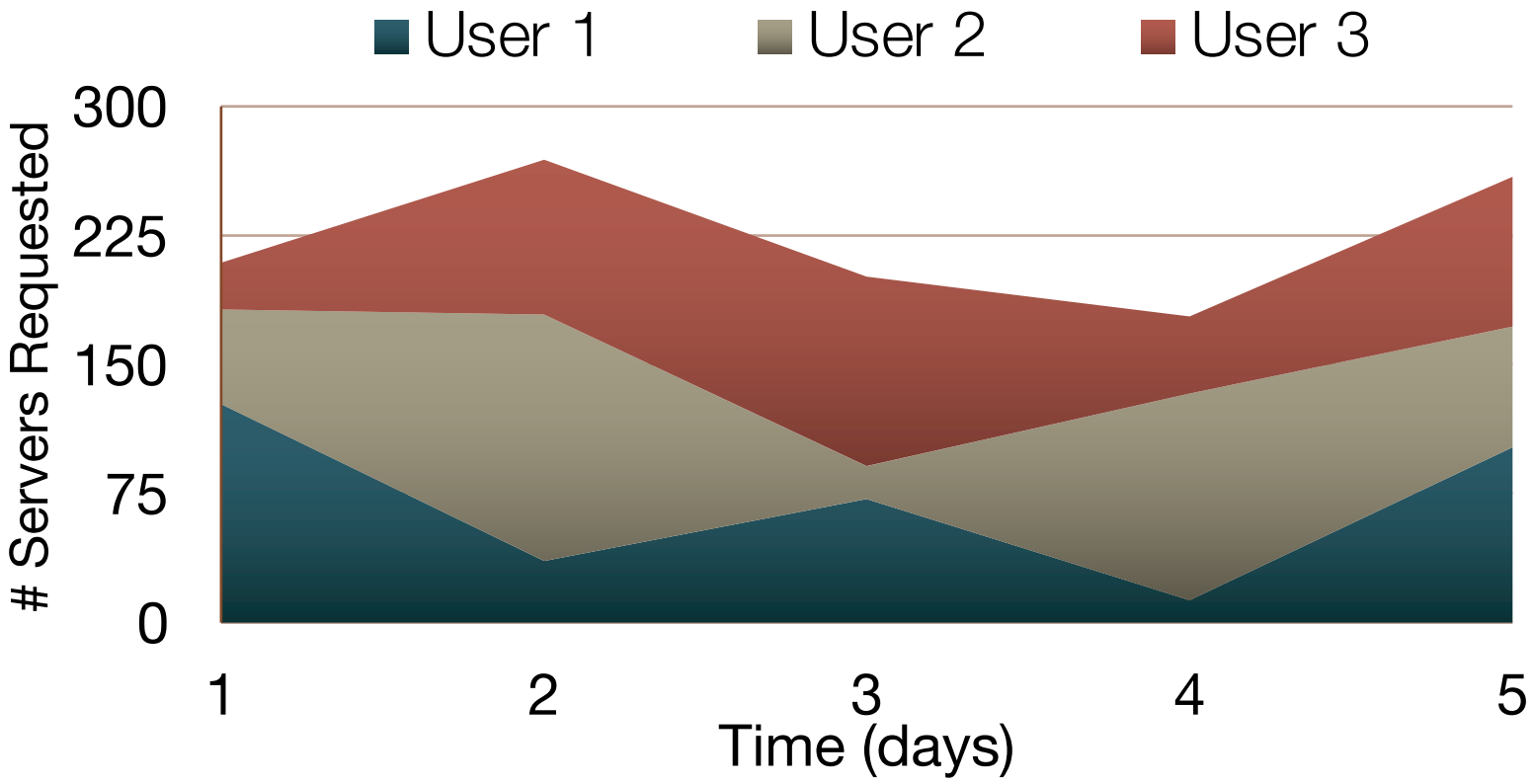
- A medium-sized datacenter (~1k servers) vs. a large datacenter (~50k servers) in 2006

Technology	Cost in Medium-sized DC	Cost in Very Large DC	Ratio
Network	$95 per Mbit/sec/month	$13 per Mbit/sec/month	7.1
Storage	$2.20 per GByte / month	$0.40 per GByte / month	5.7
Administration	≈140 Servers / Administrator	>1000 Servers / Administrator	7.1

5 - 7x decrease of cost!

Source: Ambrust et al., “Above the clouds: A Berkeley’s view of Cloud Computing.”

Statistical multiplexing



Highly profitable business for Cloud providers

Plus…

- **Leverage existing investment**, e.g., Amazon
- **Defend a franchise**, e.g., Microsoft Azure
- **Attack an incumbent**, e.g., Google Cloud Platform
- **Leverage customer relationships**, e.g., IBM
- **Become a platform**, e.g., Facebook, Apple, etc.

Summary: Why cloud?

- Better capital utilization
- The unit cost of on-demand capacity may be higher than the unit cost of fixed capacity; offset by no charge when capacity is not being used
- Elasticity, easy to scale up and down
- Access to complex infrastructure and resources without internal resources
- **Providers**: better resource utilization, lower cost

Cloud Pricing

Fundamental Drivers of Cost

- Compute (EC2)
  - charged per hour/second
  - varies by instance type (VM configurations)
- Storage (S3, EBS)
  - charged typically per GB w/ tiered pricing
- Data transfer
  - outbound is aggregated and charged, typically per GB
  - inbound has no charge (w/ some exceptions)

Let's focus on compute

How to set the unit instance price?

amazon web services | EC2

US East (N. Virginia)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.nano	1	Variable	0.5	EBS Only	$0.0059 per Hour
t2.micro	1	Variable	1	EBS Only	$0.012 per Hour
t2.small	1	Variable	2	EBS Only	$0.023 per Hour
t2.medium	2	Variable	4	EBS Only	$0.047 per Hour
t2.large	2	Variable	8	EBS Only	$0.094 per Hour
t2.xlarge	4	Variable	16	EBS Only	$0.188 per Hour
t2.2xlarge	8	Variable	32	EBS Only	$0.376 per Hour
m4.large	2	6.5	8	EBS Only	$0.108 per Hour
m4.xlarge	4	13	16	EBS Only	$0.215 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.431 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$0.862 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.155 per Hour
m4.16xlarge	64	188	256	EBS Only	$3.447 per Hour

amazon web services | EC2

Asia Pacific (Tokyo)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.nano	1	Variable	0.5	EBS Only	$0.008 per Hour
t2.micro	1	Variable	1	EBS Only	$0.016 per Hour
t2.small	1	Variable	2	EBS Only	$0.032 per Hour
t2.medium	2	Variable	4	EBS Only	$0.064 per Hour
t2.large	2	Variable	8	EBS Only	$0.128 per Hour
t2.xlarge	4	Variable	16	EBS Only	$0.256 per Hour
t2.2xlarge	8	Variable	32	EBS Only	$0.512 per Hour
m4.large	2	6.5	8	EBS Only	$0.139 per Hour
m4.xlarge	4	13	16	EBS Only	$0.278 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.556 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$1.113 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.782 per Hour
m4.16xlarge	64	188	256	EBS Only	$4.45 per Hour

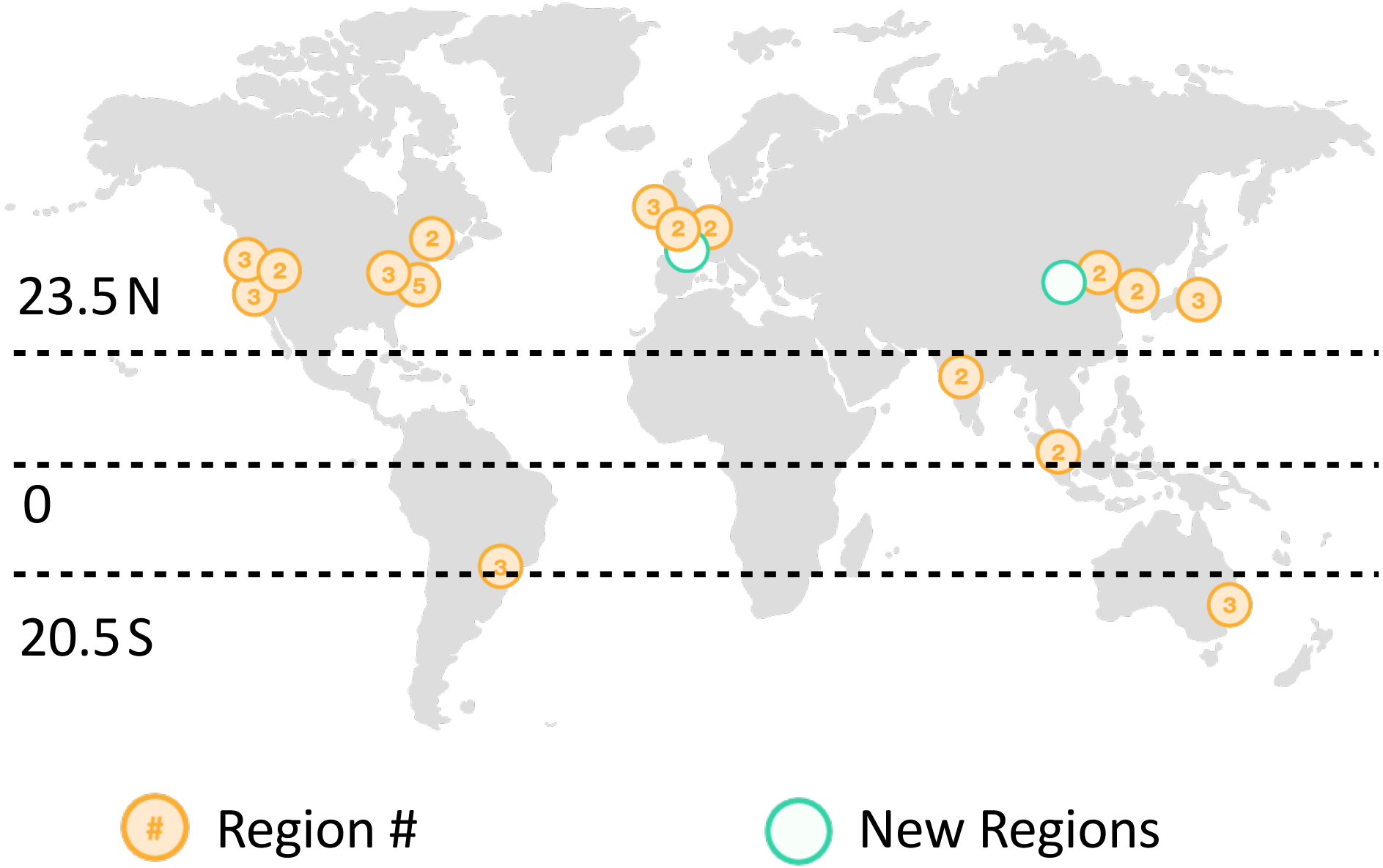
amazon web services™ | EC2

	NOVA		Tokyo
t2.nano	$0.0059 per Hour		$0.008 per Hour
t2.micro	$0.012 per Hour		$0.016 per Hour
t2.small	$0.023 per Hour		$0.032 per Hour
t2.medium	$0.047 per Hour		$0.064 per Hour
t2.large	$0.094 per Hour		$0.128 per Hour
t2.xlarge	$0.188 per Hour		$0.256 per Hour
t2.2xlarge	$0.376 per Hour	<	$0.512 per Hour
m4.large	$0.108 per Hour		$0.139 per Hour
m4.xlarge	$0.215 per Hour		$0.278 per Hour
m4.2xlarge	$0.431 per Hour		$0.556 per Hour
m4.4xlarge	[illegible]		[illegible]
m4.10xlarge	$2.155 per Hour		$2.782 per Hour
m4.16xlarge	$3.447 per Hour		$4.45 per Hour

Why location matters?

Why location matters?

- Cooling cost
- Manpower cost
- Land price
- Policy issues
- …


23.5 N
0
20.5 S
Region #
New Regions

Is on-demand pay-as-you-go pricing enough?

Diverse pricing options

- On-demand
- Reservation-based
- Spot pricing
- …

Reserved pricing

- Pay an up-front reservation fee to reserve an instance for a long period, e.g., 1 to 3 years
- Enjoy a significant discount during the reservation period

$$\text{Cost}(t) = U + \textit{discount} \times R \times t$$

Upfront

On-demand rate

Reserved pricing

- Guaranteed availability
  - Users signed up for the reserved pricing are always serviced, regardless of the DC load
  - Not possible for on-demand pricing

Reserved pricing for t2.xlarge

STANDARD 1-YEAR TERM					
Payment Option	**Upfront**	**Monthly***	**Effective Hourly****	**Savings over On-Demand**	**On-Demand Hourly**
No Upfront	$0	$109.62	$0.150	20%	$0.188 per Hour
Partial Upfront	$562	$46.85	$0.128	32%	
All Upfront	$1102	$0	$0.126	33%	

STANDARD 3-YEAR TERM					
Payment Option	**Upfront**	**Monthly***	**Effective Hourly****	**Savings over On-Demand**	**On-Demand Hourly**
Partial Upfront	$1164	$32.33	$0.089	53%	$0.188 per Hour
All Upfront	$2188	$0	$0.083	56%	

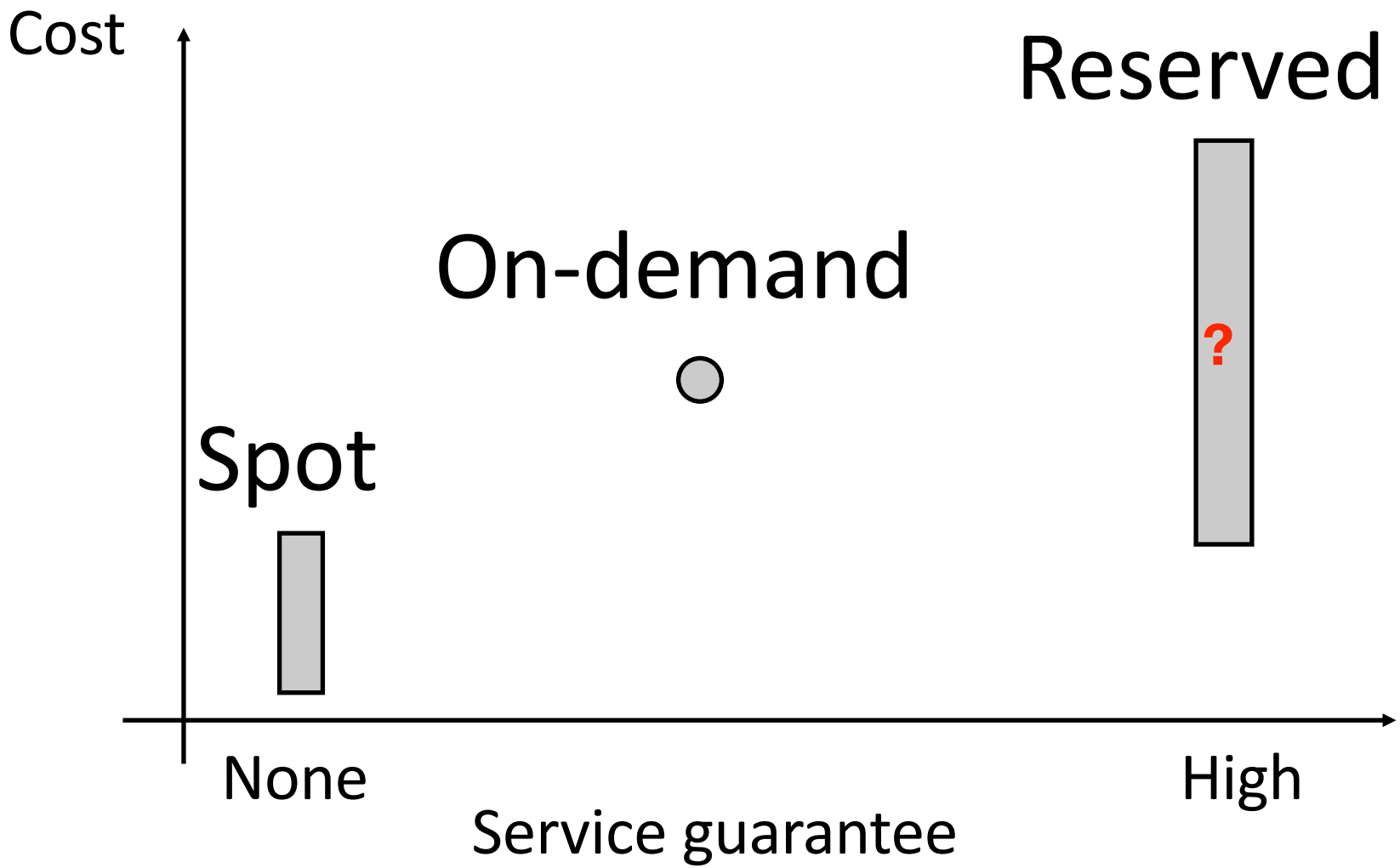
Spot pricing

- Used to be an **auction-like** pricing option
  - users submit bid for instance acquisition
  - cloud posts a spot price periodically
  - users with a higher bid than the spot price wins
    - the spot price is applied until a new one is posted
  - running users with a lower bid get their instances terminated

Spot pricing

- Spot price is usually much cheaper than on-demand
  - Does it make sense to have a higher spot price than on-demand?
- No service guarantee
  - running spot instances get terminated when the spot prices rises above the bid

Summary of pricing



Why so many different pricing models?

Market segmentation

- Reserved pricing
  - locks in long-term users
  - helps predict future demand: better for capacity planning
- On-demand
  - the fundamental cloud business model
- Spot pricing
  - leftover capacity on sale: increase utilization

Provider’s problems

- Datacenter has a limited capacity
- How to allocate the capacity for each pricing model?
  - if not planned well, one model can cannibalize the other
- How to set the price of each model?

User's problems

- How to cut down the cloud bill by combining different pricing models?
  - demand/workload prediction
  - predict spot price: many works try to reverse-engineer how the spot price is set
  - creative use of spot instances
    - periodic checkpointing and recovery upon instance revocation
    - save over 50% compared with on-demand

The rise of brokerage service

- Cloud brokerage service
  - helps users to make instance acquisition strategies
  - trade-in unused instances in a secondary cloud marketplace
  - hybrid cloud: connects to multiple cloud providers to explore the best deal
  - many innovative business models coming…

Credits

- Some slides are adapted from course slides of COMP 4651 in HKUST