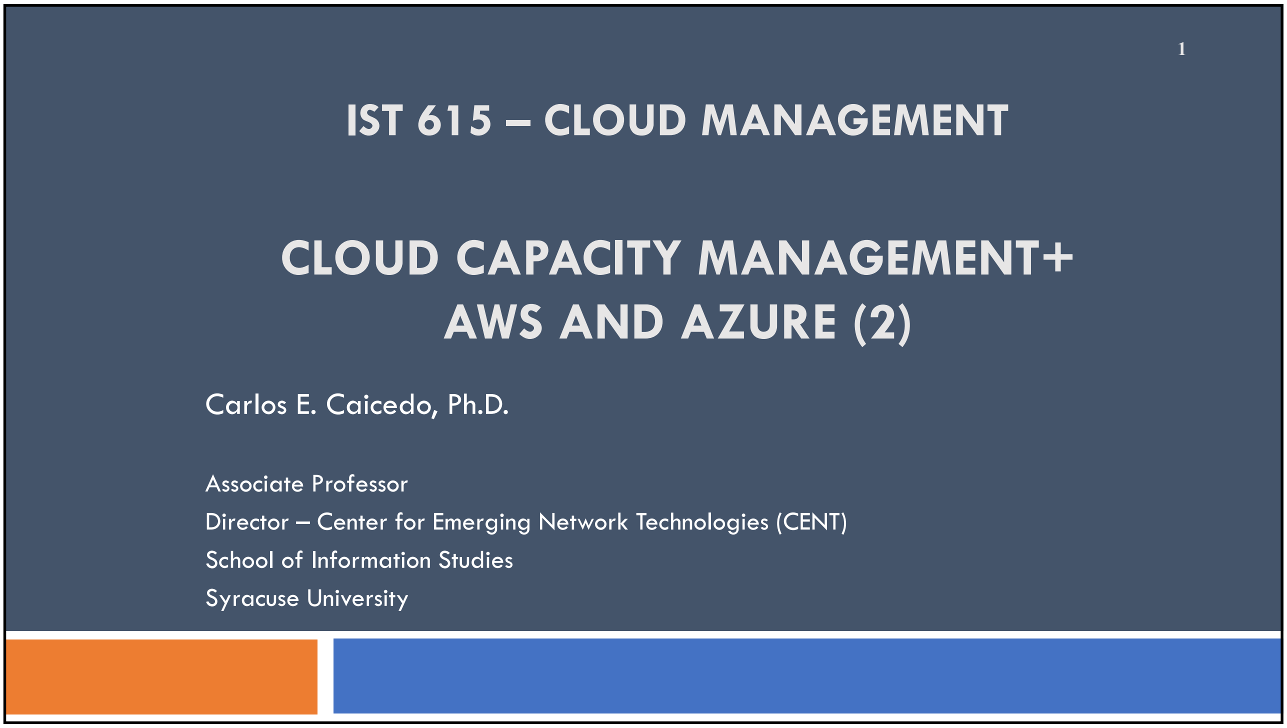


1

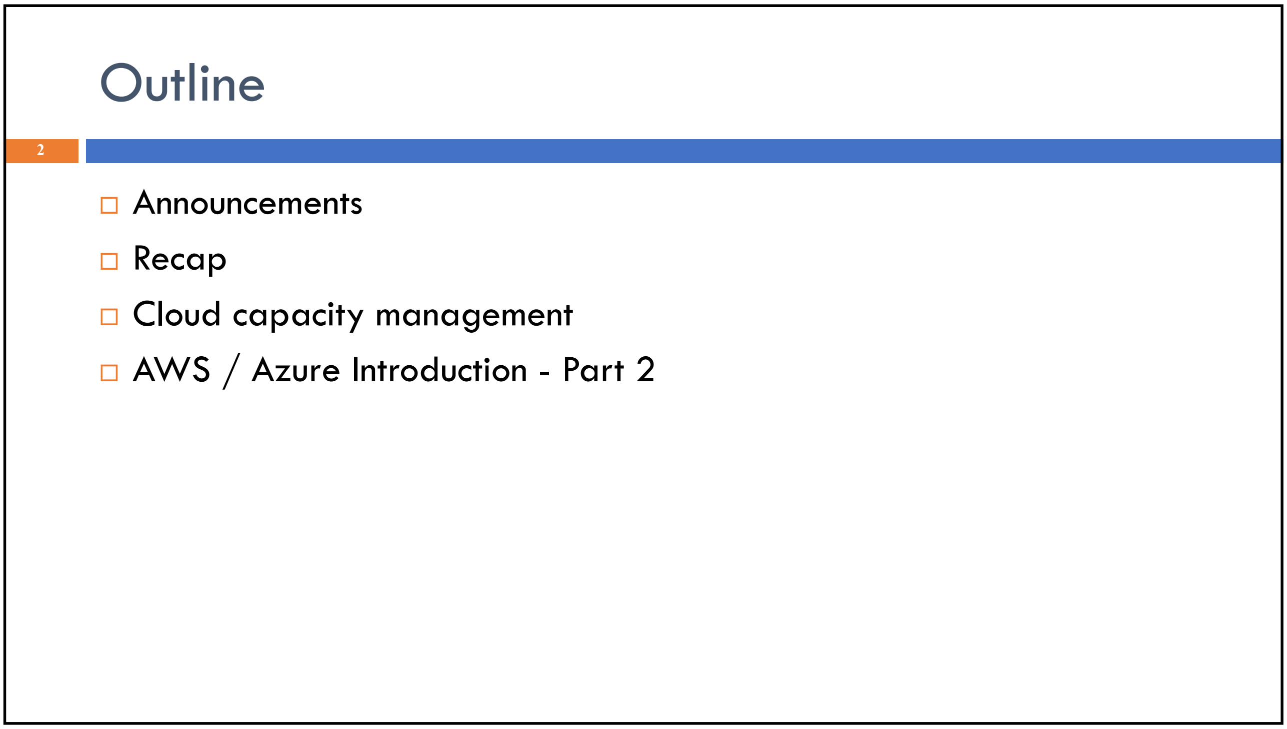
IST 615 – CLOUD MANAGEMENT

CLOUD CAPACITY MANAGEMENT+ AWS AND AZURE (2)

Carlos E. Caicedo, Ph.D.

Associate Professor
Director – Center for Emerging Network Technologies (CENT)
School of Information Studies
Syracuse University

1



Outline

2

- □ Announcements
- □ Recap
- □ Cloud capacity management
- □ AWS / Azure Introduction - Part 2

2

Labs/Class activities

3

- □ Lab 2: AWS
  - Due today
- □ Homework #2
  - Released today, Due next week
- □ Lab 3: Cloud service integration
  - To be released next week

- □ Start exploring services in Azure/AWS and crafting ideas for your mini-project (Lab #8)
- □ October 8: Midterm (during the first half of the class session) + regular lecture session afterwards (2nd half of the class session)

3

Labs – Troubleshooting forum

4

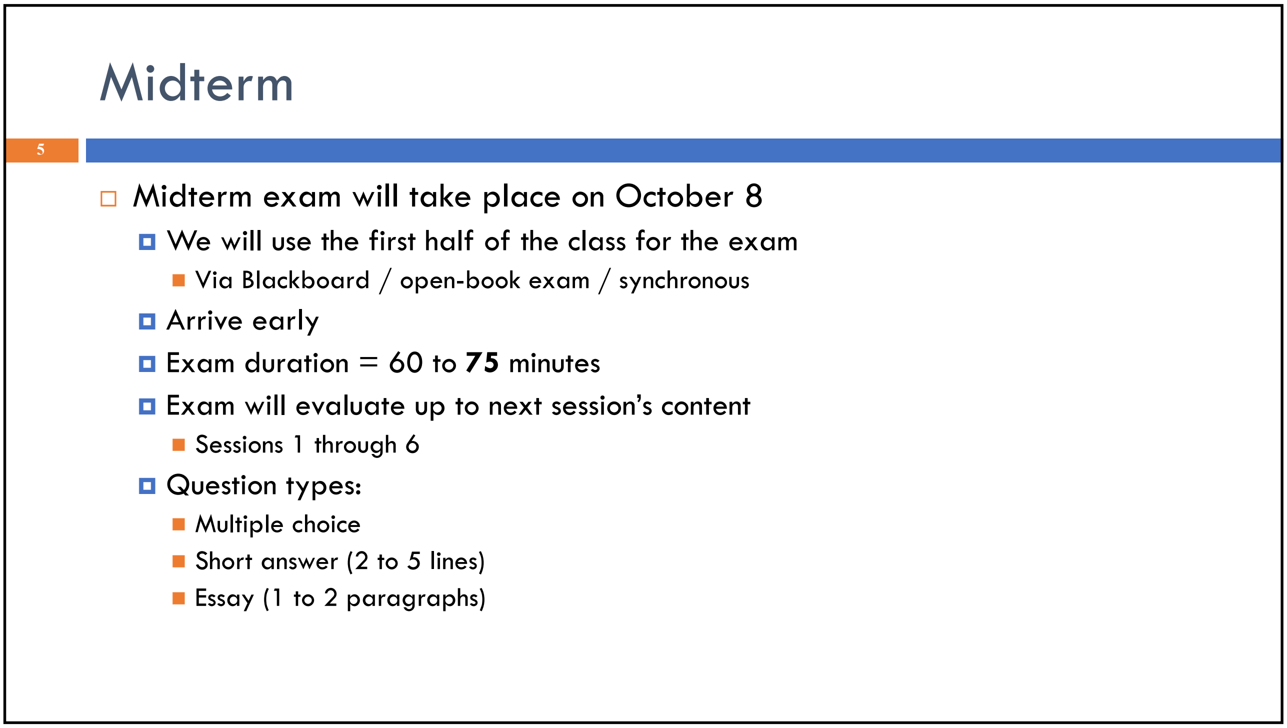
- □ A discussion forum to submit issues that you may have while executing a lab has been created in Blackboard
- Please describe your problem clearly and mention at least 2 approaches to solve the problem that you have already tried.
- After posting an issue, e-mail the instructor if you want him to get involved
- Other students can also proceed to provide help and it would be counted as participation in class (see Participation points)
- If possible, include a Kaltura Media video capture of the problem and try not to include private credentials in the video.

- □ **Participation points**

Participation in this forum will get you participation points through the following rules:

- Help requests posted in "Lab Troubleshooting" discussion board (5 points per req, 10 points max)
- Solutions to help requests posted by others (10 points per solution, 30 points max.)
- First two solutions (if correct) that are different/complimentary to each other will get the points

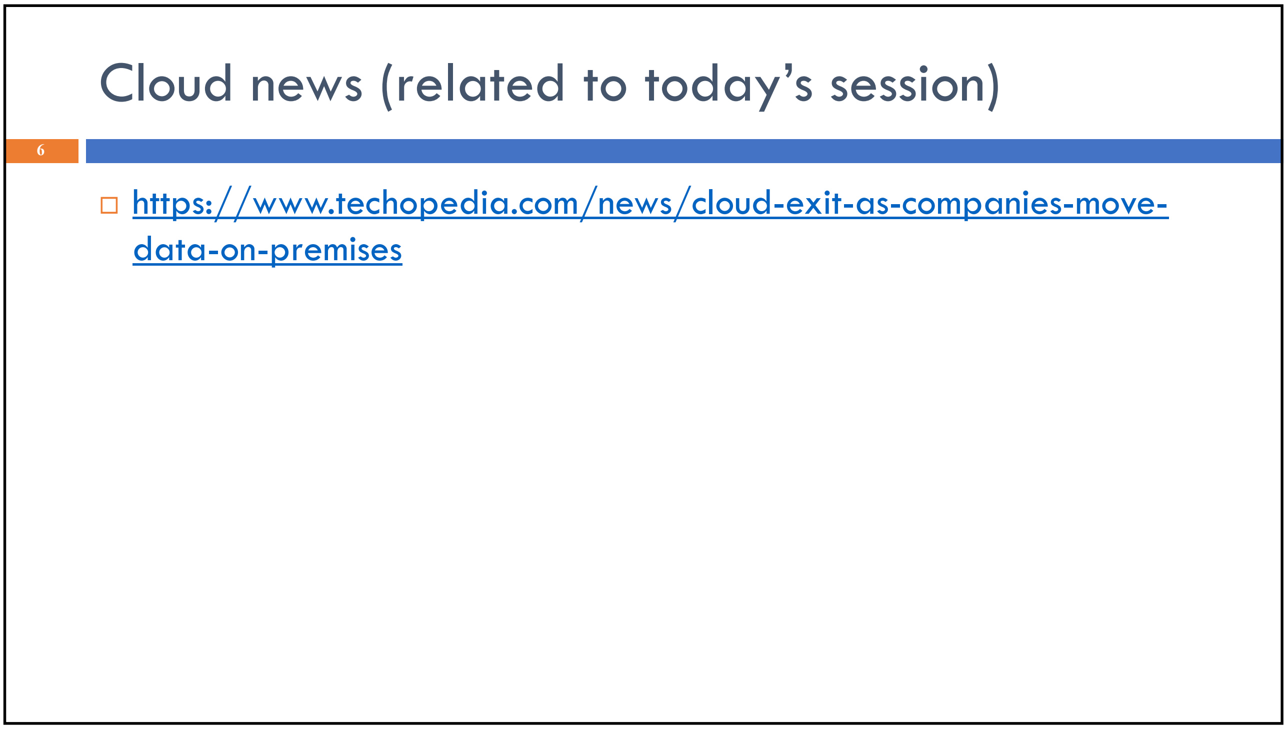
4



Midterm

- Midterm exam will take place on October 8
  - We will use the first half of the class for the exam
    - Via Blackboard / open-book exam / synchronous
  - Arrive early
  - Exam duration = 60 to **75** minutes
  - Exam will evaluate up to next session's content
    - Sessions 1 through 6
  - Question types:
    - Multiple choice
    - Short answer (2 to 5 lines)
    - Essay (1 to 2 paragraphs)

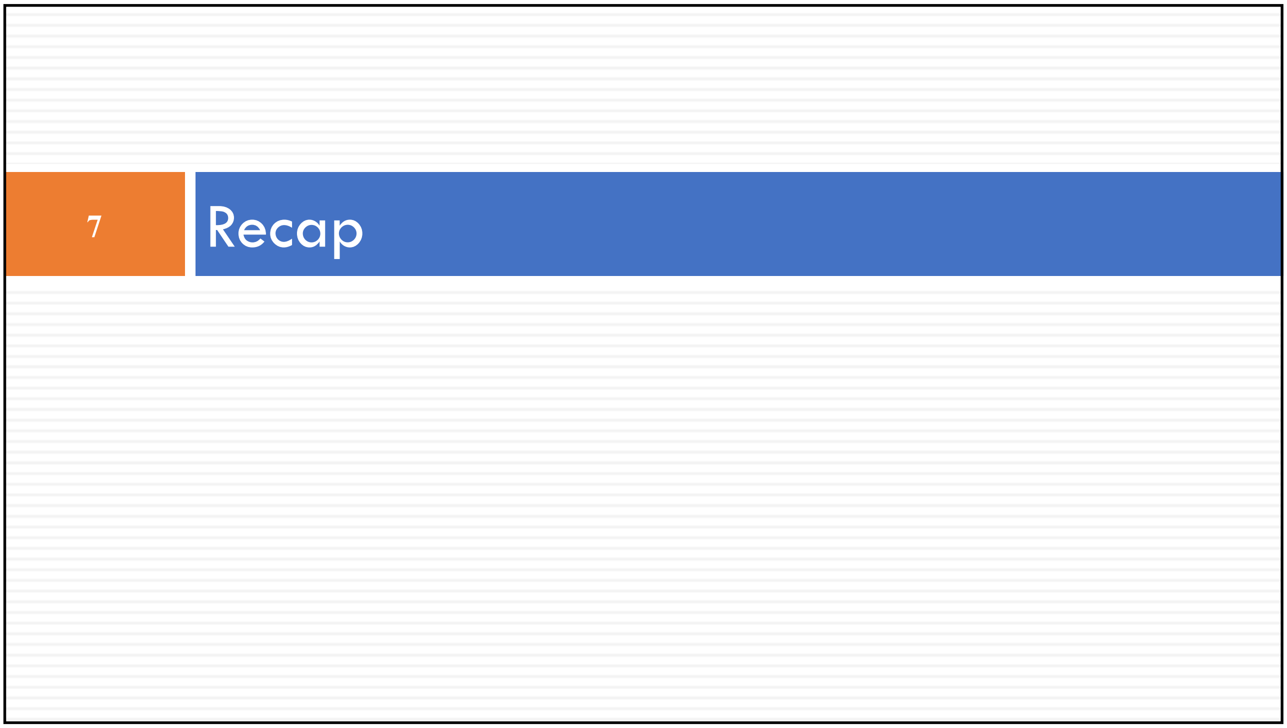
5



Cloud news (related to today's session)

- https://www.techopedia.com/news/cloud-exit-as-companies-move-data-on-premises

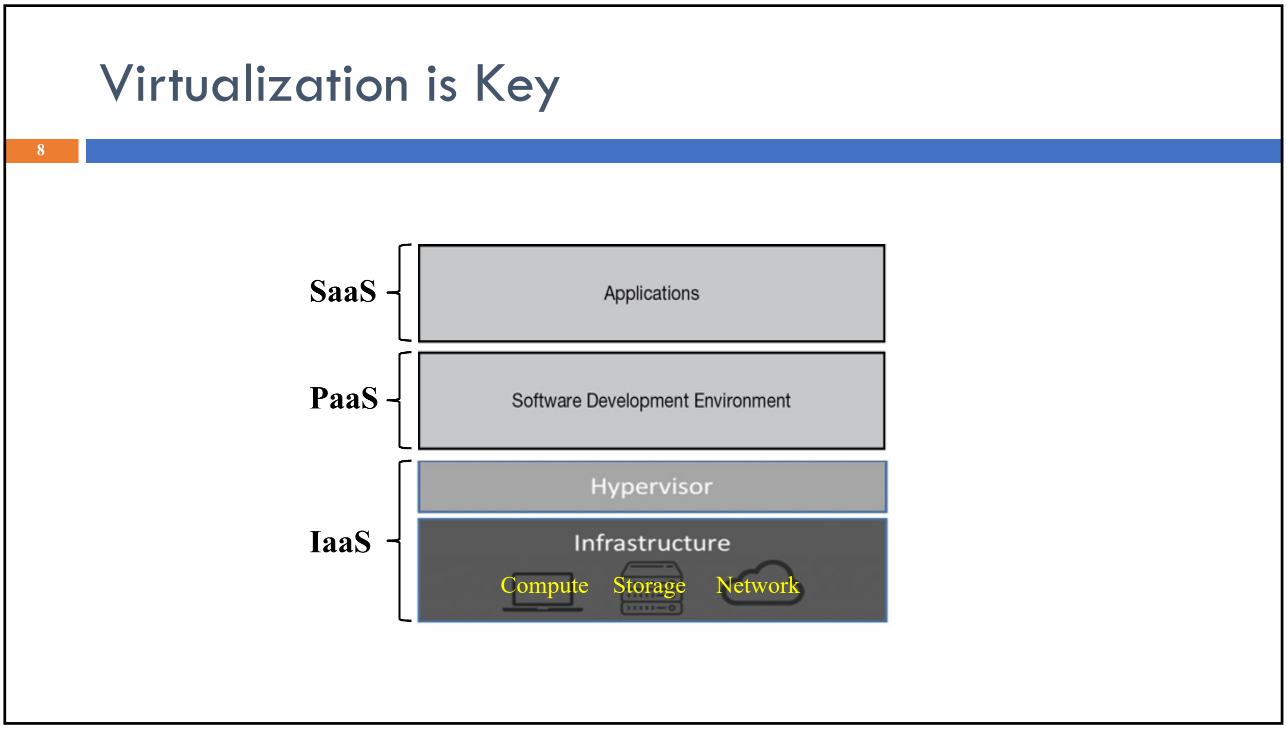
6



7

Recap

7



Virtualization is Key

8

SaaS – Applications

PaaS – Software Development Environment

IaaS – Hypervisor

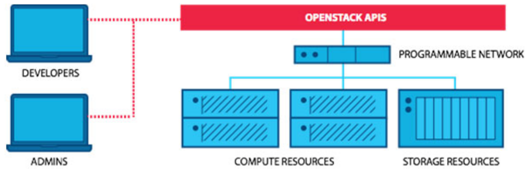
Infrastructure

Compute Storage Network

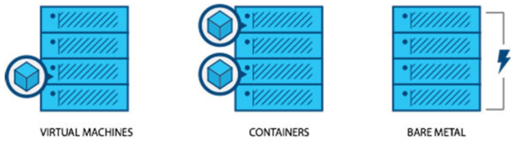
8

openstack.

What is OpenStack



Programmable infrastructure that lays a common set of APIs on top of compute, networking and storage



One platform for virtual machines, containers and bare metal

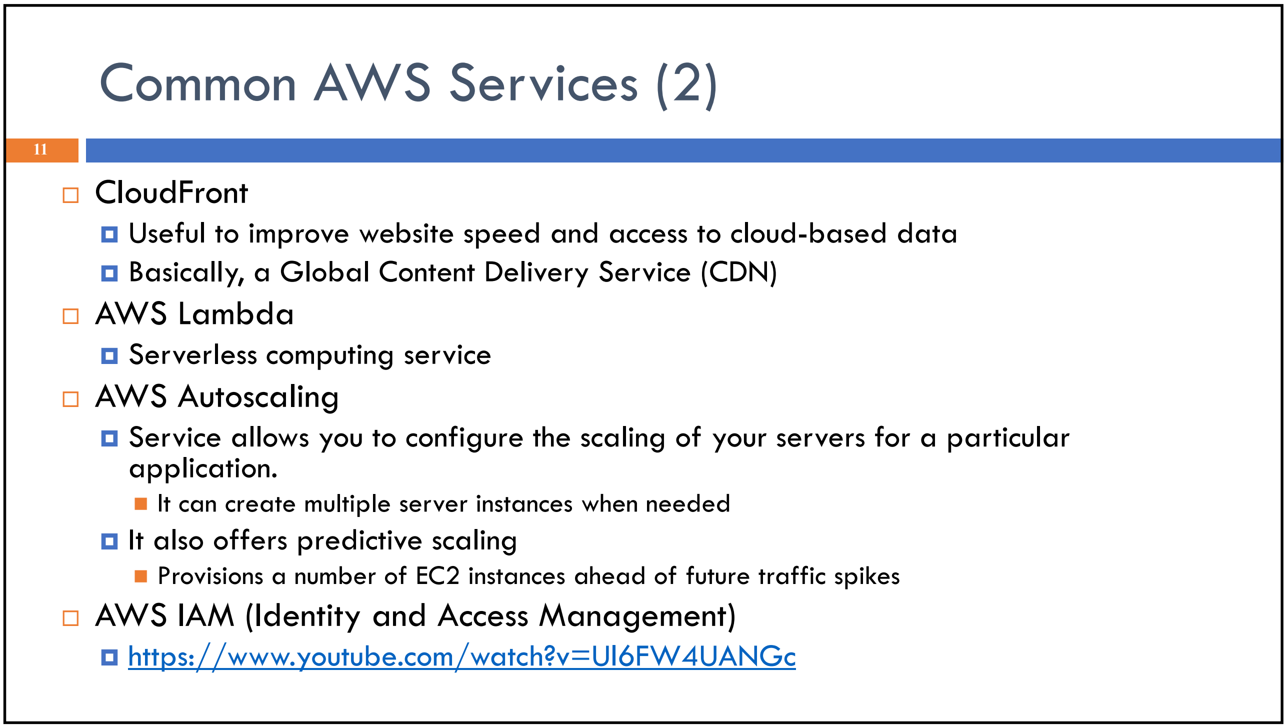
9

Common AWS Services

10

- Amazon EC2
  - Allows you to create virtual machines while managing other server features such as security, and storage.
  - https://www.youtube.com/watch?v=TsRBftzZsQo
- Amazon RDS
  - Relational Database Service (RDS): Allows you to create dedicated instances of databases (in minutes).
  - Instances can support multiple database engines
    - SQL Server, PostgreSQL, etc.
- S3 (Simple Storage Service)
  - Very secure and redundant file storage service
    - By default it stores data in three data centers within a specific region
    - Supports other security and high availability options
- Amazon VPC
  - This service creates a private virtual network that can only be accessed by the people and systems you authorize

10

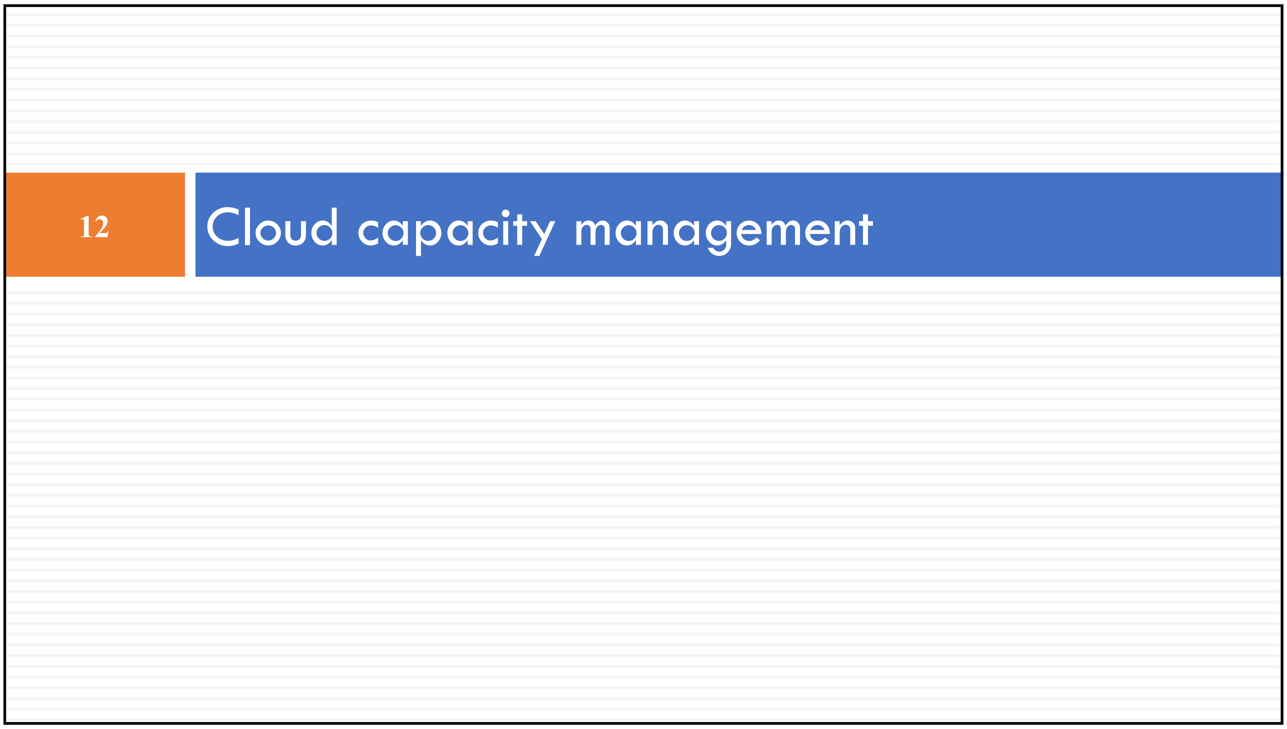


Common AWS Services (2)

11

- CloudFront
  - Useful to improve website speed and access to cloud-based data
  - Basically, a Global Content Delivery Service (CDN)
- AWS Lambda
  - Serverless computing service
- AWS Autoscaling
  - Service allows you to configure the scaling of your servers for a particular application.
    - It can create multiple server instances when needed
  - It also offers predictive scaling
    - Provisions a number of EC2 instances ahead of future traffic spikes
- AWS IAM (Identity and Access Management)
  - https://www.youtube.com/watch?v=Ul6FW4UANGc

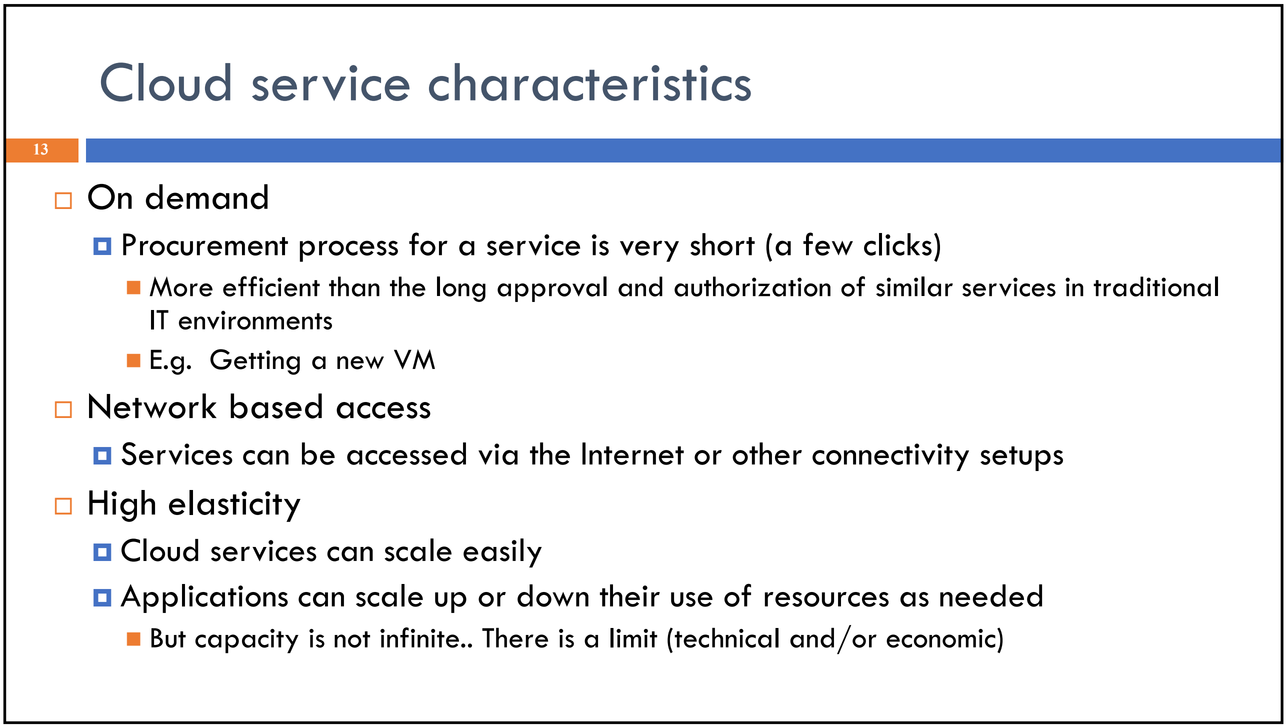
11



12

Cloud capacity management

12

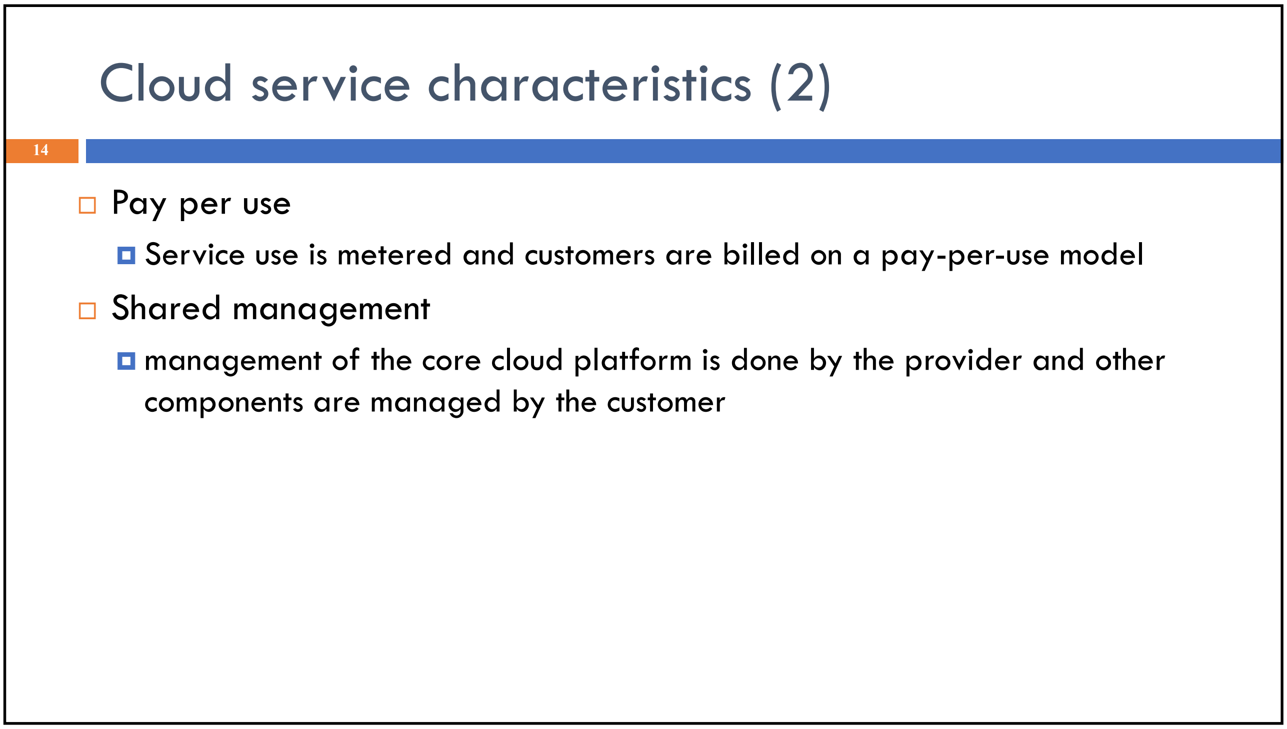


Cloud service characteristics

13

- □ On demand
  - Procurement process for a service is very short (a few clicks)
    - More efficient than the long approval and authorization of similar services in traditional IT environments
    - E.g. Getting a new VM
- □ Network based access
  - Services can be accessed via the Internet or other connectivity setups
- □ High elasticity
  - Cloud services can scale easily
  - Applications can scale up or down their use of resources as needed
    - But capacity is not infinite.. There is a limit (technical and/or economic)

13

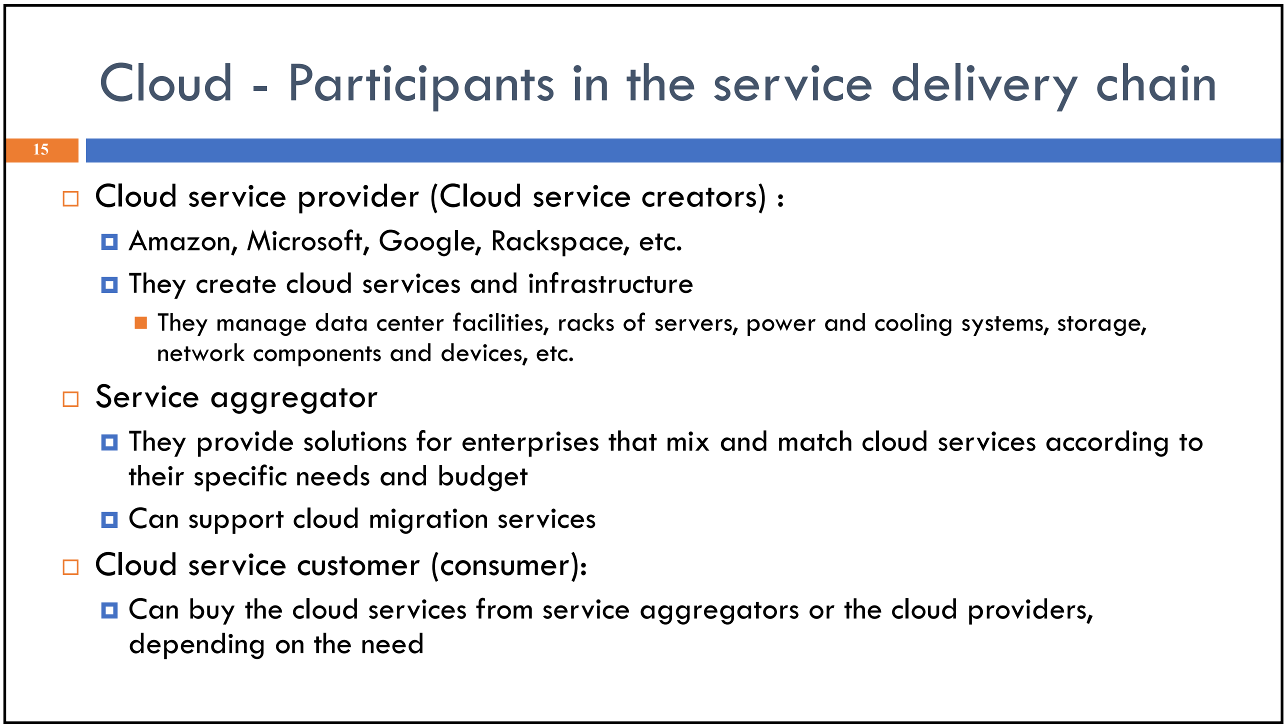


Cloud service characteristics (2)

14

- □ Pay per use
  - Service use is metered and customers are billed on a pay-per-use model
- □ Shared management
  - management of the core cloud platform is done by the provider and other components are managed by the customer

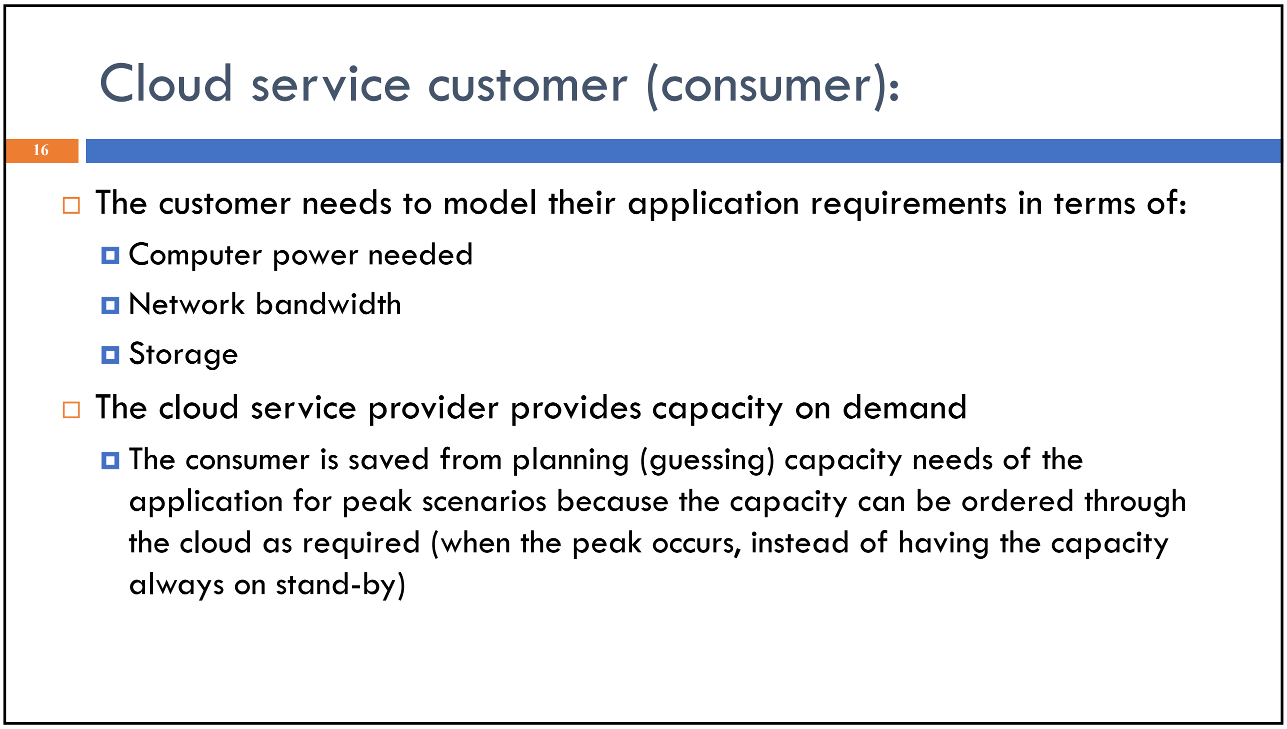
14



Cloud - Participants in the service delivery chain

- Cloud service provider (Cloud service creators) :
  - Amazon, Microsoft, Google, Rackspace, etc.
  - They create cloud services and infrastructure
    - They manage data center facilities, racks of servers, power and cooling systems, storage, network components and devices, etc.
- Service aggregator
  - They provide solutions for enterprises that mix and match cloud services according to their specific needs and budget
  - Can support cloud migration services
- Cloud service customer (consumer):
  - Can buy the cloud services from service aggregators or the cloud providers, depending on the need

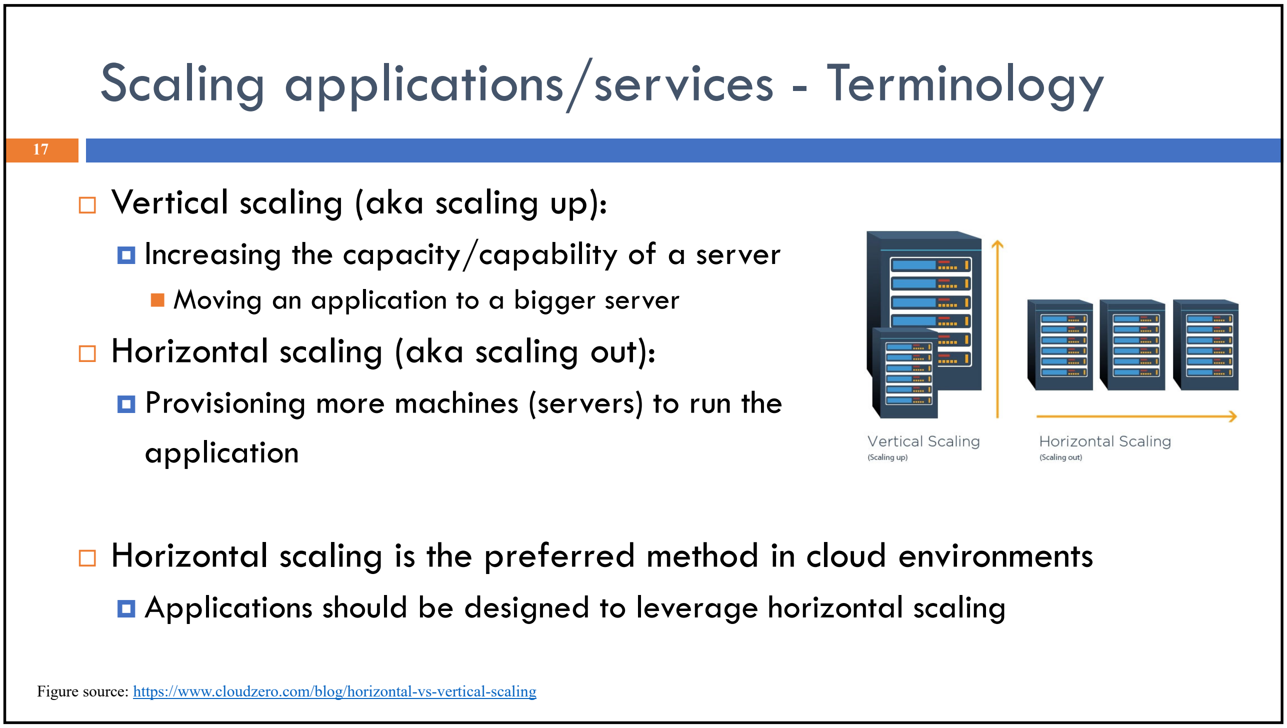
15



Cloud service customer (consumer):

- The customer needs to model their application requirements in terms of:
  - Computer power needed
  - Network bandwidth
  - Storage
- The cloud service provider provides capacity on demand
  - The consumer is saved from planning (guessing) capacity needs of the application for peak scenarios because the capacity can be ordered through the cloud as required (when the peak occurs, instead of having the capacity always on stand-by)

16



Scaling applications/services - Terminology

17

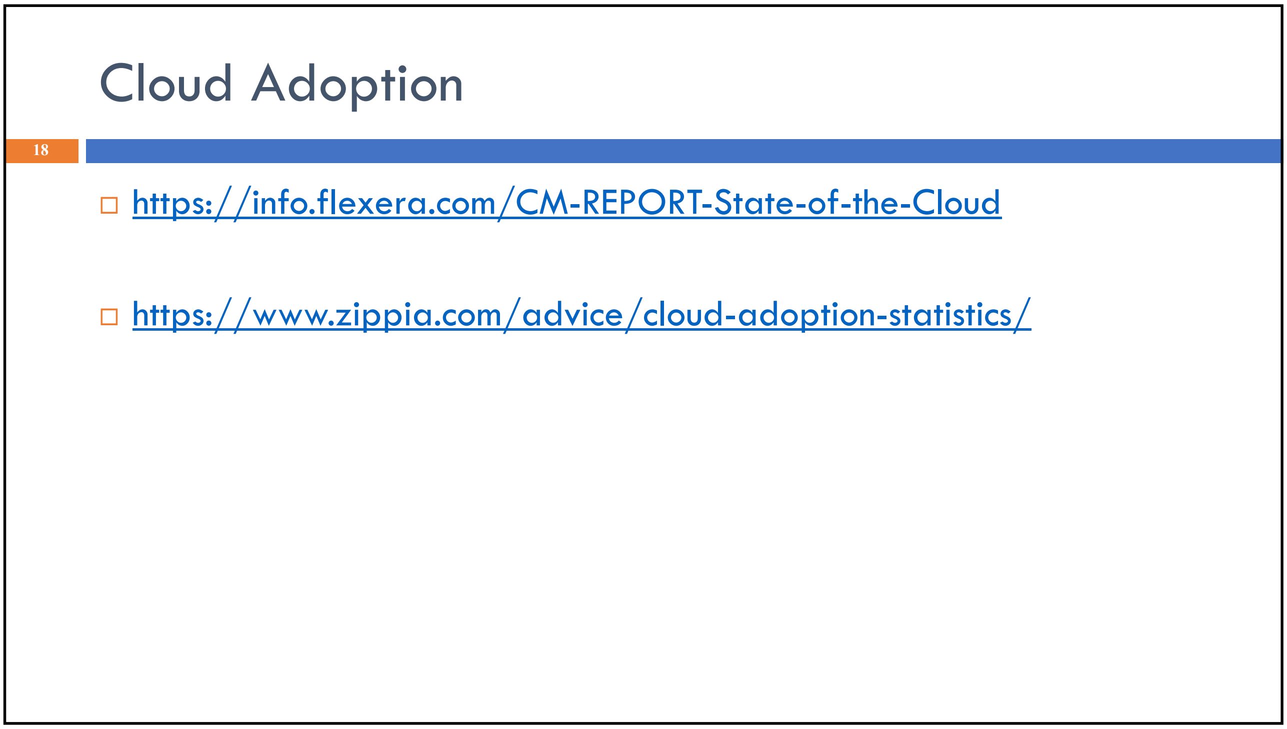
- Vertical scaling (aka scaling up):
  - Increasing the capacity/capability of a server
    - Moving an application to a bigger server
- Horizontal scaling (aka scaling out):
  - Provisioning more machines (servers) to run the application

Vertical Scaling
(Scaling up)
Horizontal Scaling
(Scaling out)

- Horizontal scaling is the preferred method in cloud environments
  - Applications should be designed to leverage horizontal scaling

Figure source: https://www.cloudzero.com/blog/horizontal-vs-vertical-scaling

17



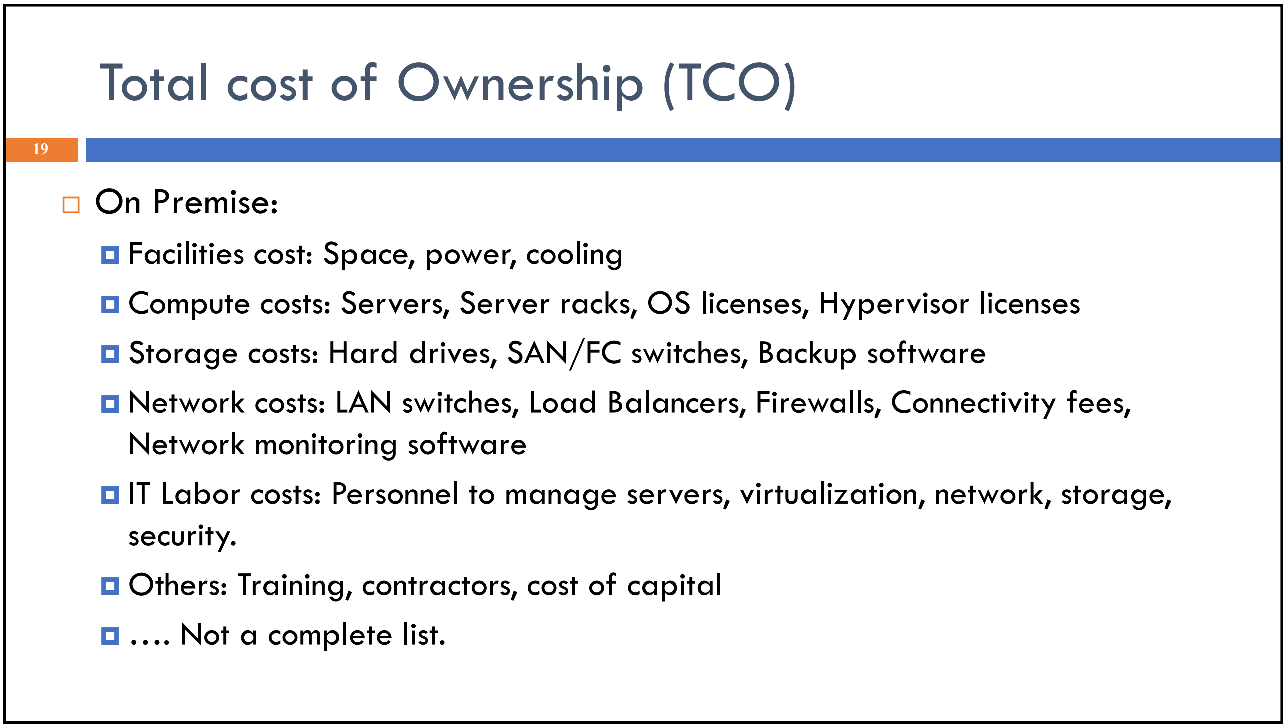
Cloud Adoption

18

- https://info.flexera.com/CM-REPORT-State-of-the-Cloud

- https://www.zippia.com/advice/cloud-adoption-statistics/

18

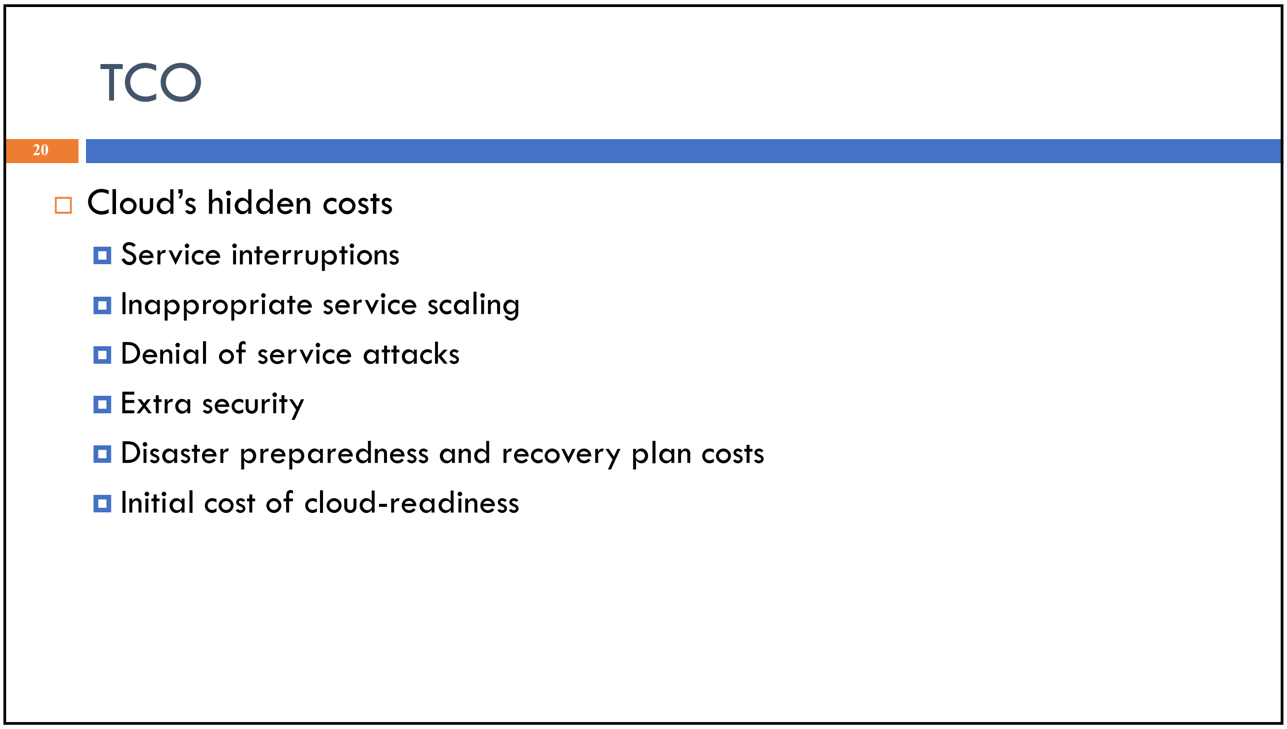


Total cost of Ownership (TCO)

19

- On Premise:
  - Facilities cost: Space, power, cooling
  - Compute costs: Servers, Server racks, OS licenses, Hypervisor licenses
  - Storage costs: Hard drives, SAN/FC switches, Backup software
  - Network costs: LAN switches, Load Balancers, Firewalls, Connectivity fees, Network monitoring software
  - IT Labor costs: Personnel to manage servers, virtualization, network, storage, security.
  - Others: Training, contractors, cost of capital
  - …. Not a complete list.

19

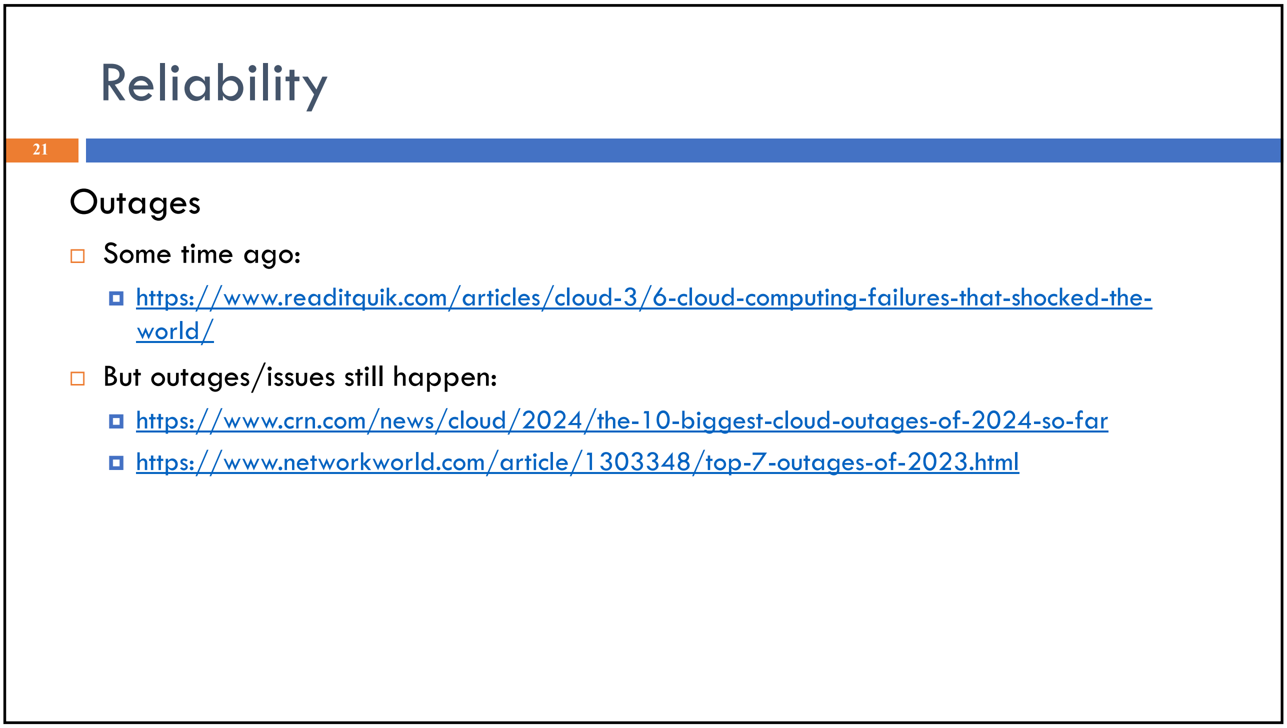


TCO

20

- Cloud’s hidden costs
  - Service interruptions
  - Inappropriate service scaling
  - Denial of service attacks
  - Extra security
  - Disaster preparedness and recovery plan costs
  - Initial cost of cloud-readiness

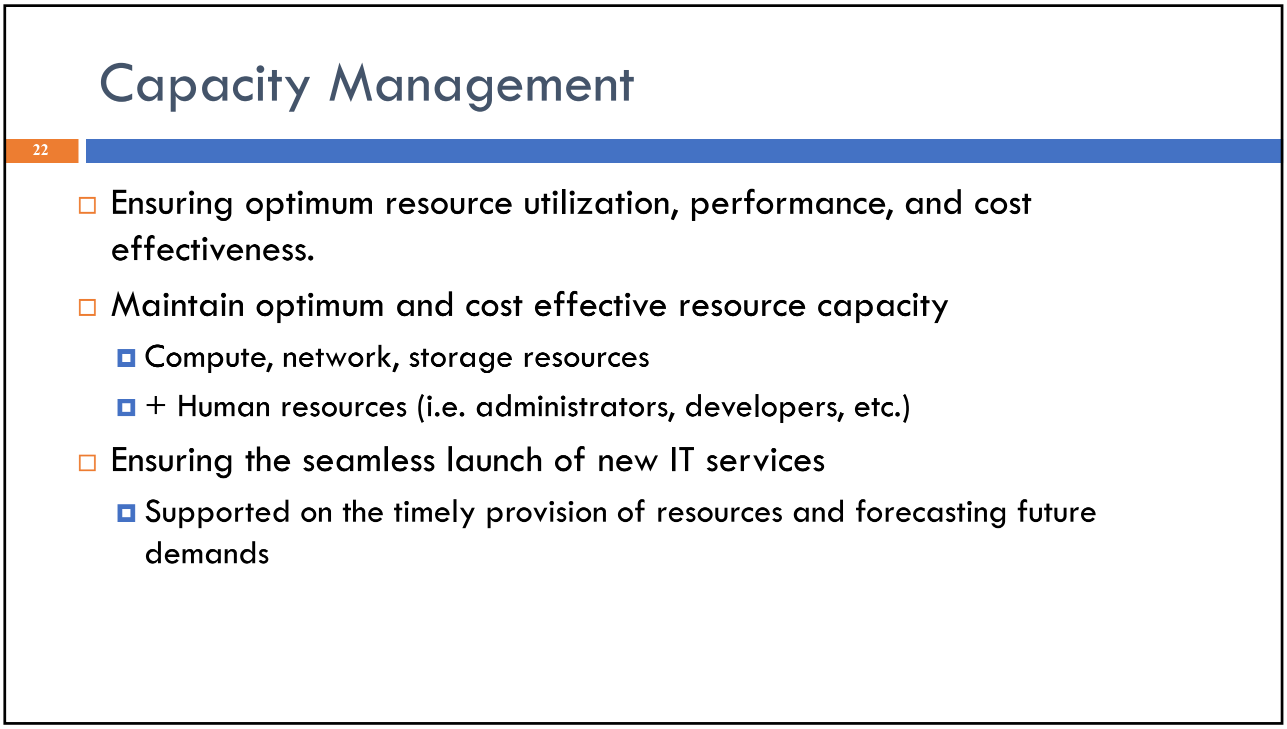
20



Reliability

Outages

- Some time ago:
  - https://www.readitquik.com/articles/cloud-3/6-cloud-computing-failures-that-shocked-the-world/
- But outages/issues still happen:
  - https://www.crn.com/news/cloud/2024/the-10-biggest-cloud-outages-of-2024-so-far
  - https://www.networkworld.com/article/1303348/top-7-outages-of-2023.html



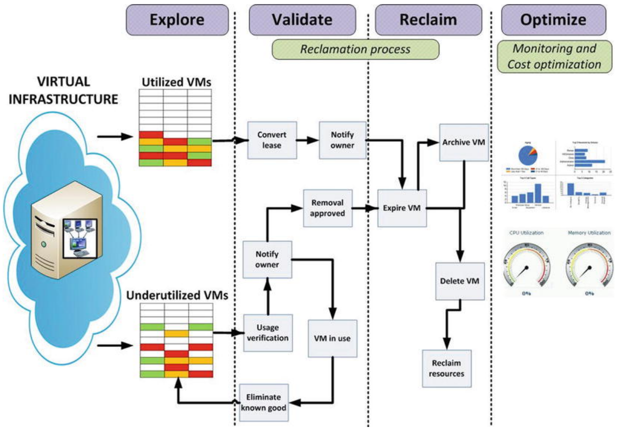
Capacity Management

- Ensuring optimum resource utilization, performance, and cost effectiveness.
- Maintain optimum and cost effective resource capacity
  - Compute, network, storage resources
  - + Human resources (i.e. administrators, developers, etc.)
- Ensuring the seamless launch of new IT services
  - Supported on the timely provision of resources and forecasting future demands

Capacity management goals

23

- □ Efficient resource utilization
  - Manage workloads efficiently
- □ Reduce wasted resources
- □ Enable and support monitoring of service levels
- □ Controlling VM sprawl



Source: Cloud Capacity Management, P. Wali; N. Sabharwal, Apress, 2013

23

Capacity management goals (2)

24

- □ Define rules for handling resources that fail
- □ Define auto-scaling mechanisms to handle unforeseen demand and/or seasonal variations in demand
- □ Ensure the operation of cloud/multi-cloud resources under appropriate economic constraints

24

Capacity management activities

25

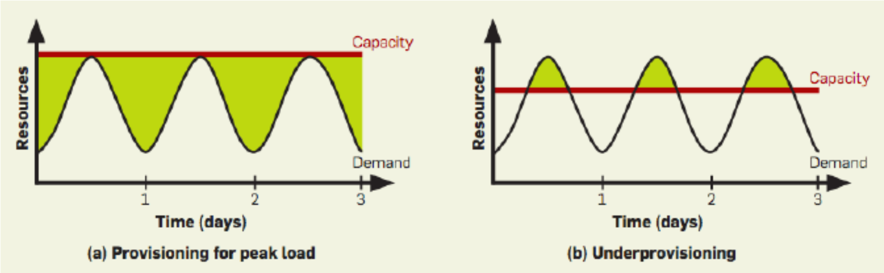
- Proactive activities
  - Taking actions on performance issues before they occur
  - Forecasting future capacity requirements
  - Modeling predicted changes in IT services
  - Managing upgrades so that they don’t disrupt current services
  - Tuning and optimizing the performance of services and components
- Reactive activities
  - Monitoring and adjusting the current performance of both services and components
  - Reacting to and assisting with specific performance issues
  - Responding to all capacity-related events and implementing a corrective action

25

Planning computing capacity

26

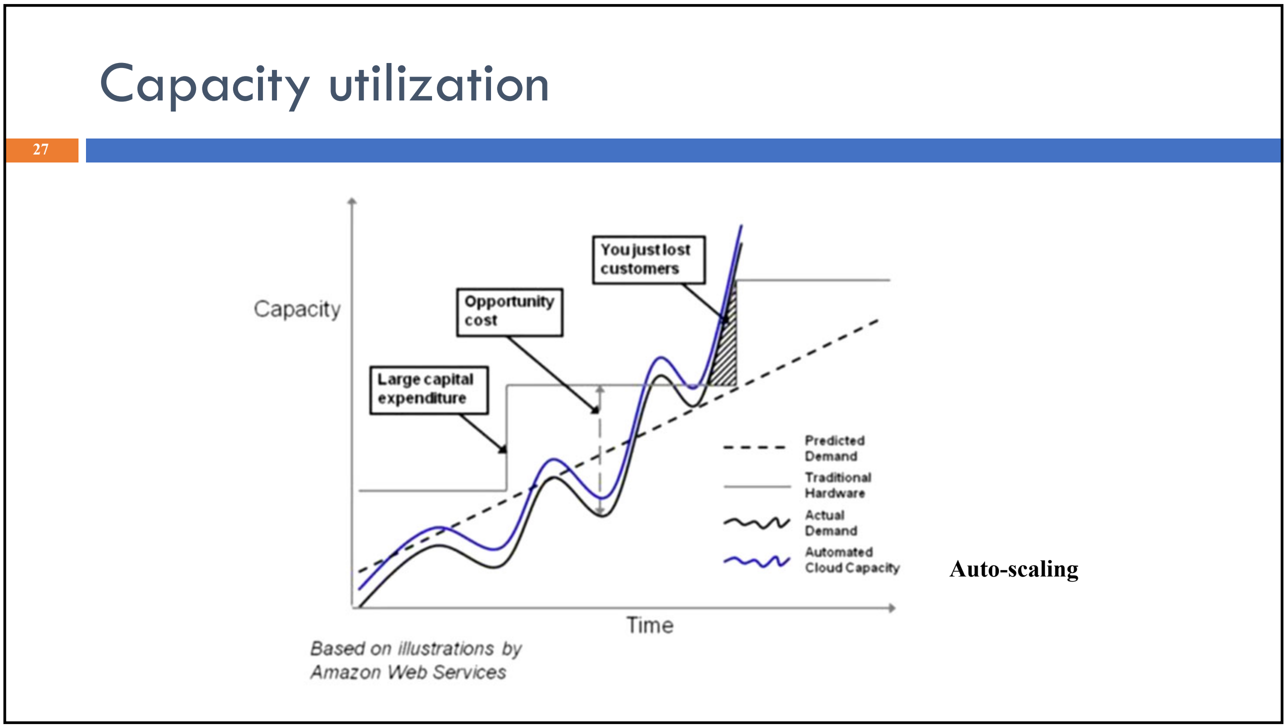
Without elasticity, capacity can be wasted or workloads don’t get enough resources



a) Even with peak load correctly predicted, without elasticity we waste resources (shared area) during non-peak times.
b) Under provisioning. Cloud provider loses revenue from workloads that are not served (shaded area). Customer, not happy.

Figure – source: Armbrust, Michael, et. al. “A view of cloud computing”, Communications of the ACM 53.4 (2010)

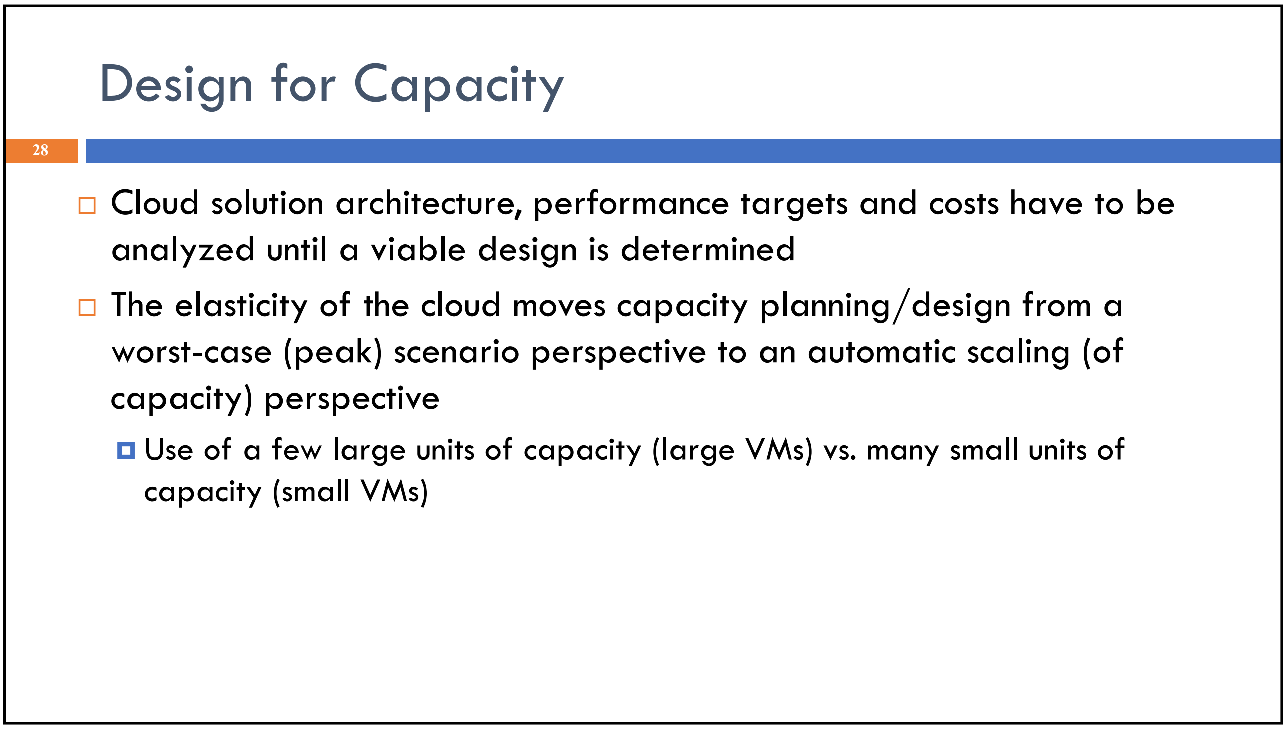
26



Capacity utilization

Auto-scaling

Based on illustrations by Amazon Web Services



Design for Capacity

- Cloud solution architecture, performance targets and costs have to be analyzed until a viable design is determined
- The elasticity of the cloud moves capacity planning/design from a worst-case (peak) scenario perspective to an automatic scaling (of capacity) perspective
  - Use of a few large units of capacity (large VMs) vs. many small units of capacity (small VMs)

29

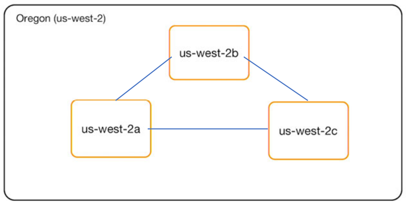
AWS / Azure – Part 2

29

AWS Regions

30

- □ https://aws.amazon.com/about-aws/global-infrastructure/regions_az/
- □ Each AWS Region consists of multiple, isolated, and physically separate Availability Zones within a geographic area.
  - ◘ Each AZ has independent power, cooling, and physical security and is connected via redundant, ultra-low-latency networks
  - ◘ Two to five AZs may be present in an AWS region
- □ Choosing between regions
  - ◘ Cost
  - ◘ Regulation (data sovereignty)
  - ◘ Latency



30

AWS Availability Zones

31

- □ An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region.
  - Useful for the deployment of applications that need fault tolerance, high availability and scalability.
- □ All traffic between AZs is encrypted.
- □ AZs are physically separated by a meaningful distance, many kilometers, from any other AZ, although all are within 100 km (60 miles) of each other.

31

Service Level Agreements (SLA)

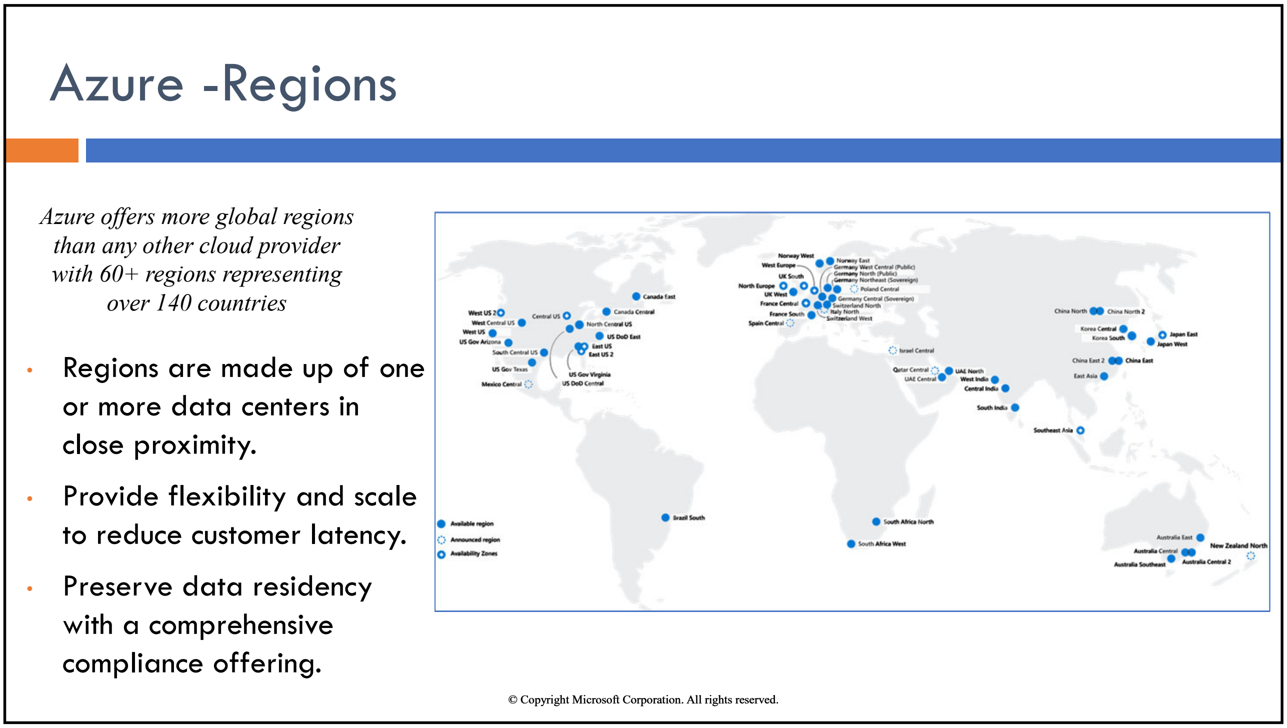
32

- □ SLA uptime

Uptime Percentage	Average Annual Downtime
99%	87 hours, 40 minutes
99.9%	8 hours, 46 minutes
99.99%	52 minutes, 36 seconds
99.999%	5 minutes, 16 seconds
99.9999%	31.6 seconds

Amazon compute SLA: https://aws.amazon.com/compute/sla/

32

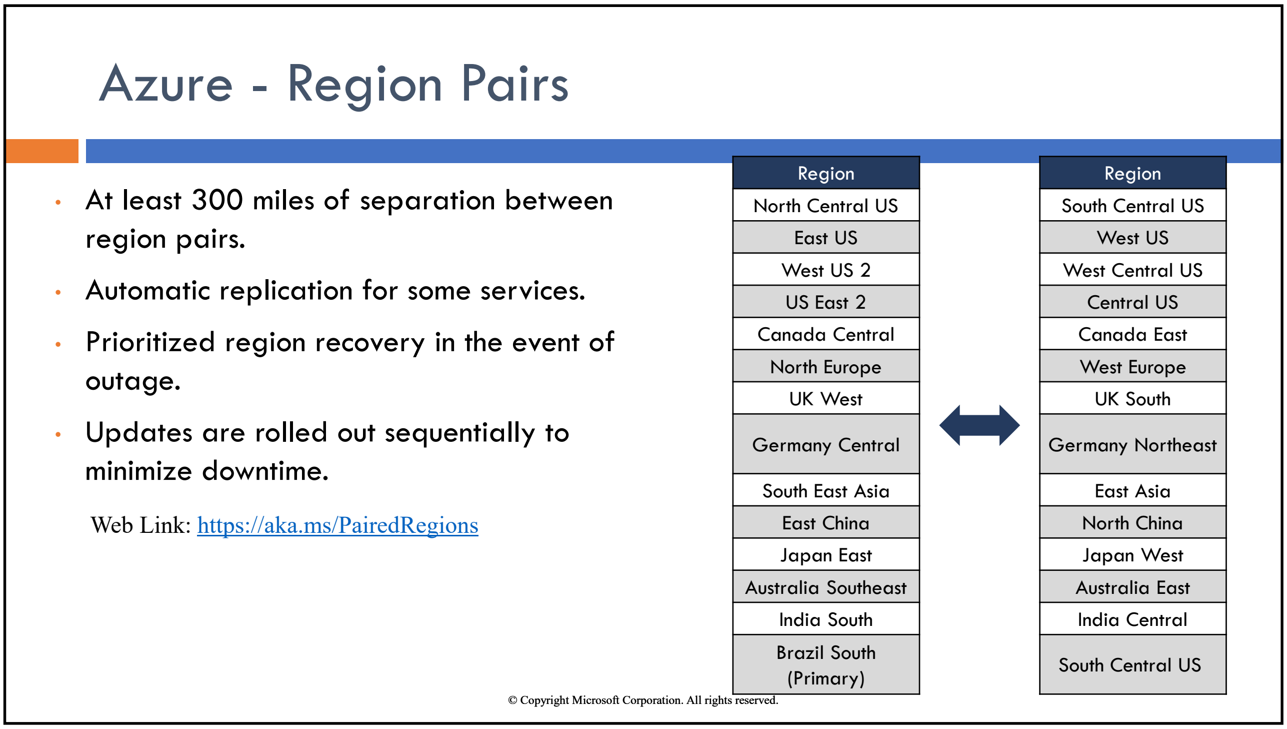


Azure -Regions

Azure offers more global regions than any other cloud provider with 60+ regions representing over 140 countries

- Regions are made up of one or more data centers in close proximity.
- Provide flexibility and scale to reduce customer latency.
- Preserve data residency with a comprehensive compliance offering.

33



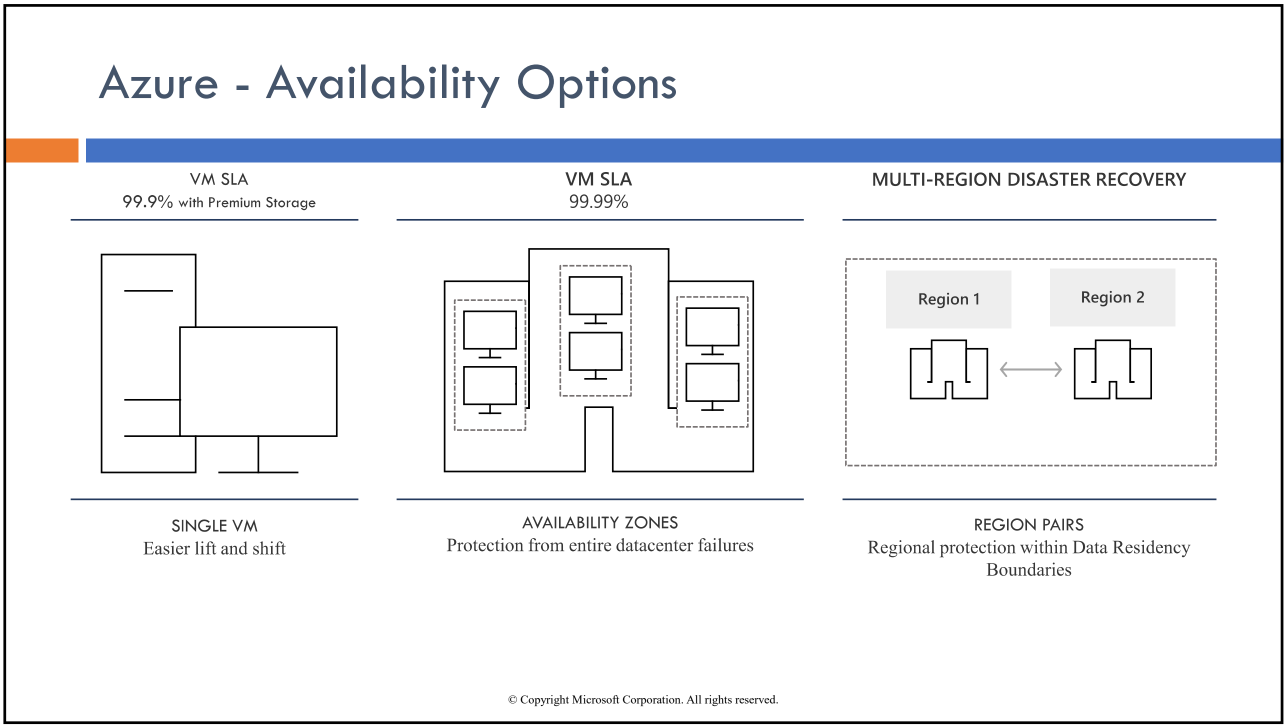
Azure - Region Pairs

- At least 300 miles of separation between region pairs.
- Automatic replication for some services.
- Prioritized region recovery in the event of outage.
- Updates are rolled out sequentially to minimize downtime.

Web Link: https://aka.ms/PairedRegions

Region	Region
North Central US	South Central US
East US	West US
West US 2	West Central US
US East 2	Central US
Canada Central	Canada East
North Europe	West Europe
UK West	UK South
Germany Central	Germany Northeast
South East Asia	East Asia
East China	North China
Japan East	Japan West
Australia Southeast	Australia East
India South	India Central
Brazil South (Primary)	South Central US

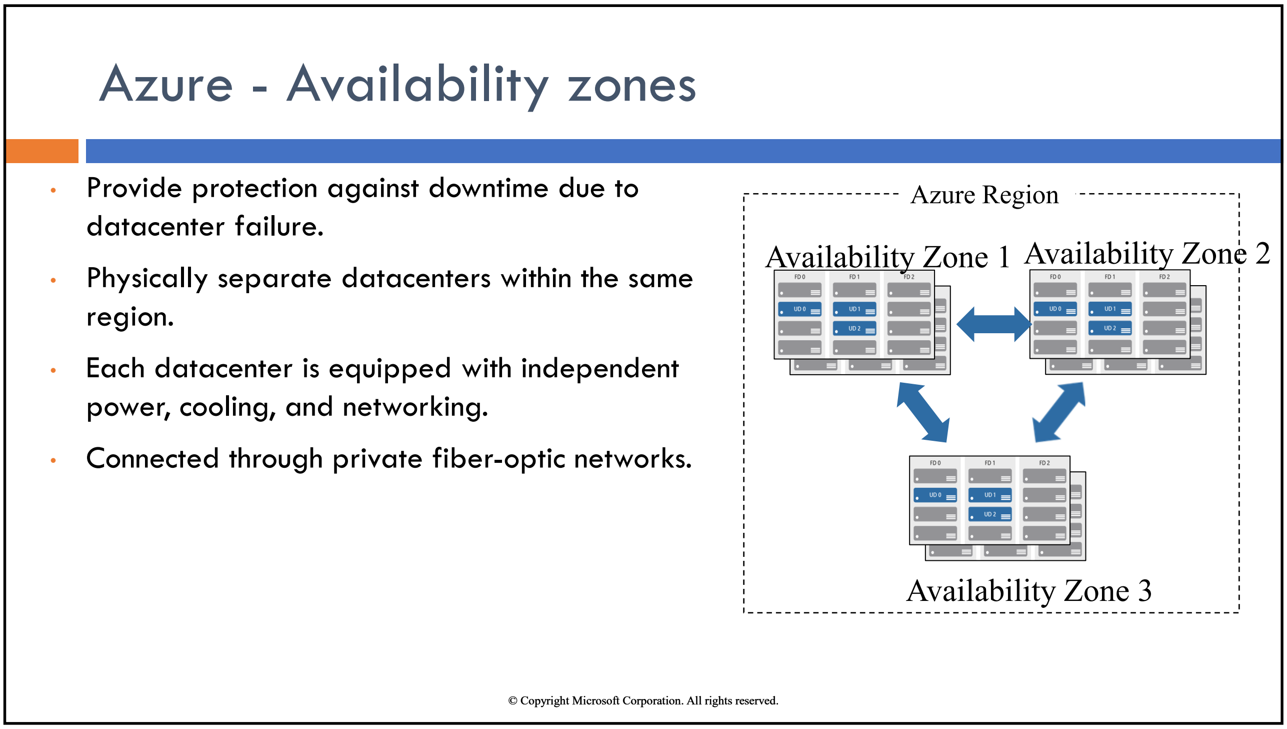
34



Azure - Availability Options

VM SLA
99.9% with Premium Storage
VM SLA
99.99%
MULTI-REGION DISASTER RECOVERY
Region 1
Region 2
SINGLE VM
Easier lift and shift
AVAILABILITY ZONES
Protection from entire datacenter failures
REGION PAIRS
Regional protection within Data Residency Boundaries

35



Azure - Availability zones

- Provide protection against downtime due to datacenter failure.
- Physically separate datacenters within the same region.
- Each datacenter is equipped with independent power, cooling, and networking.
- Connected through private fiber-optic networks.

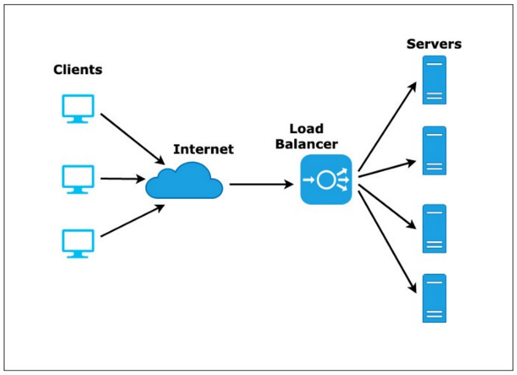
36

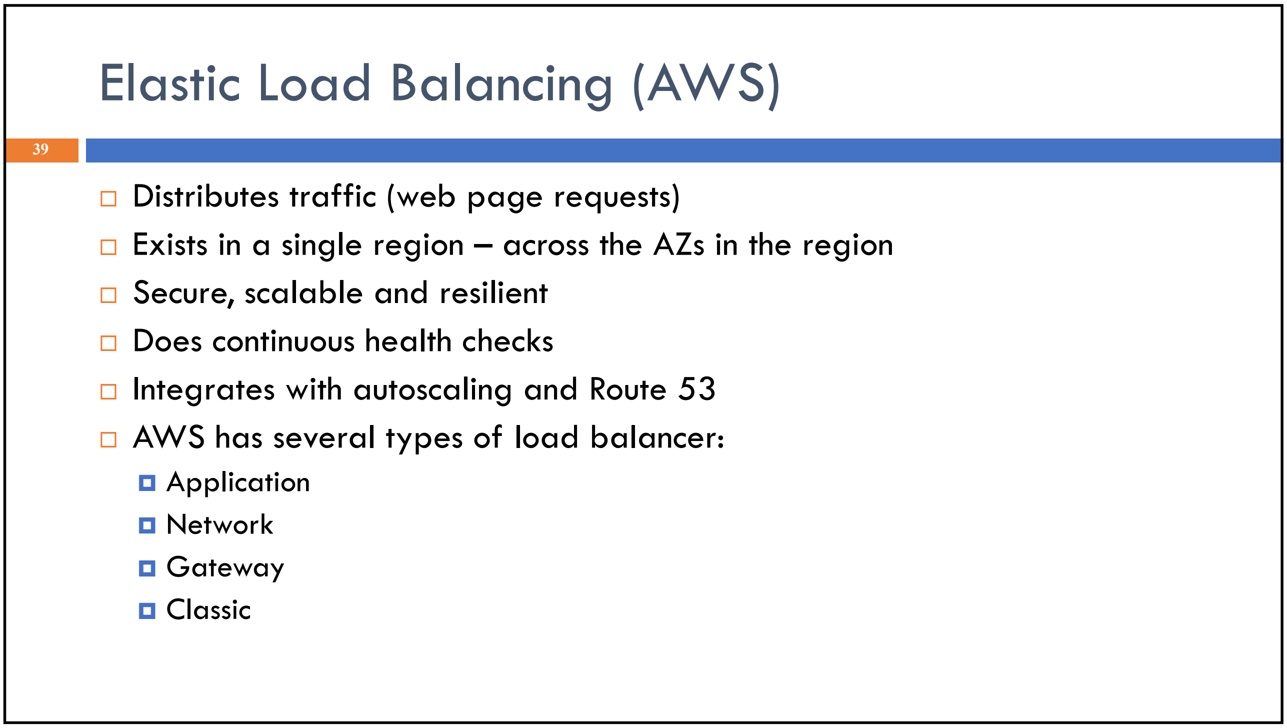
Storage

- Storage can impact the performance of an application running on a VM instance
- Types of storage (Amazon):
  - Elastic Block Storage: Persistent block storage for EC2 instance
  - Simple Storage Service (S3): Object storage. Highly reliable.
  - Elastic File System (EFS): Provides massively parallel shared access to thousands of Amazon EC2 instances. Great for applications with high levels of aggregate throughput and IOPS.
  - S3 Glacier: secure, durable, and extremely low-cost storage service for data archiving and long-term backup
  - More at https://docs.aws.amazon.com/whitepapers/latest/aws-overview/storage-services.html
- Types of storage (Azure)
  - https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction

Load balancing

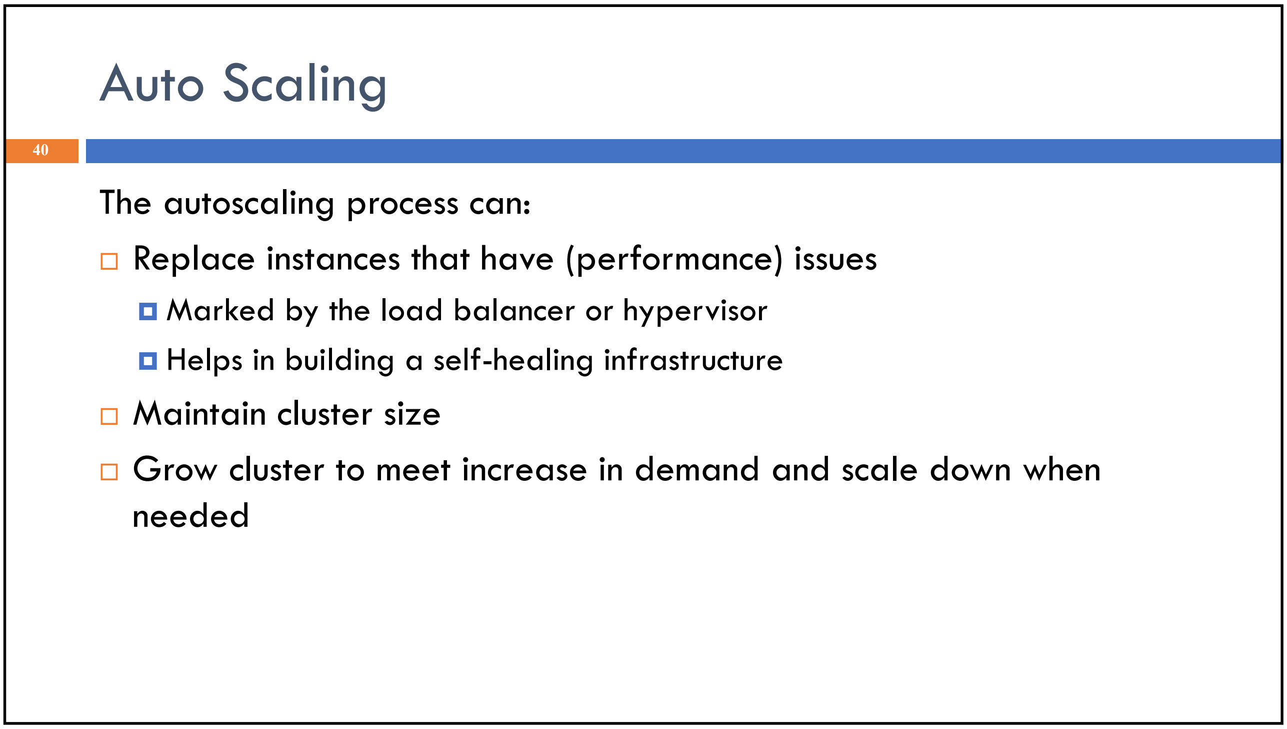
- A load balancer distributes incoming traffic among a set of distributed servers (a cluster)
- Load balancers also monitor the health of the servers in a cluster
  - If a server is down, it will not get any more requests forwarded to it
- Load balancers can adapt to the scaling (up or down) of the number of servers in the cluster





Elastic Load Balancing (AWS)

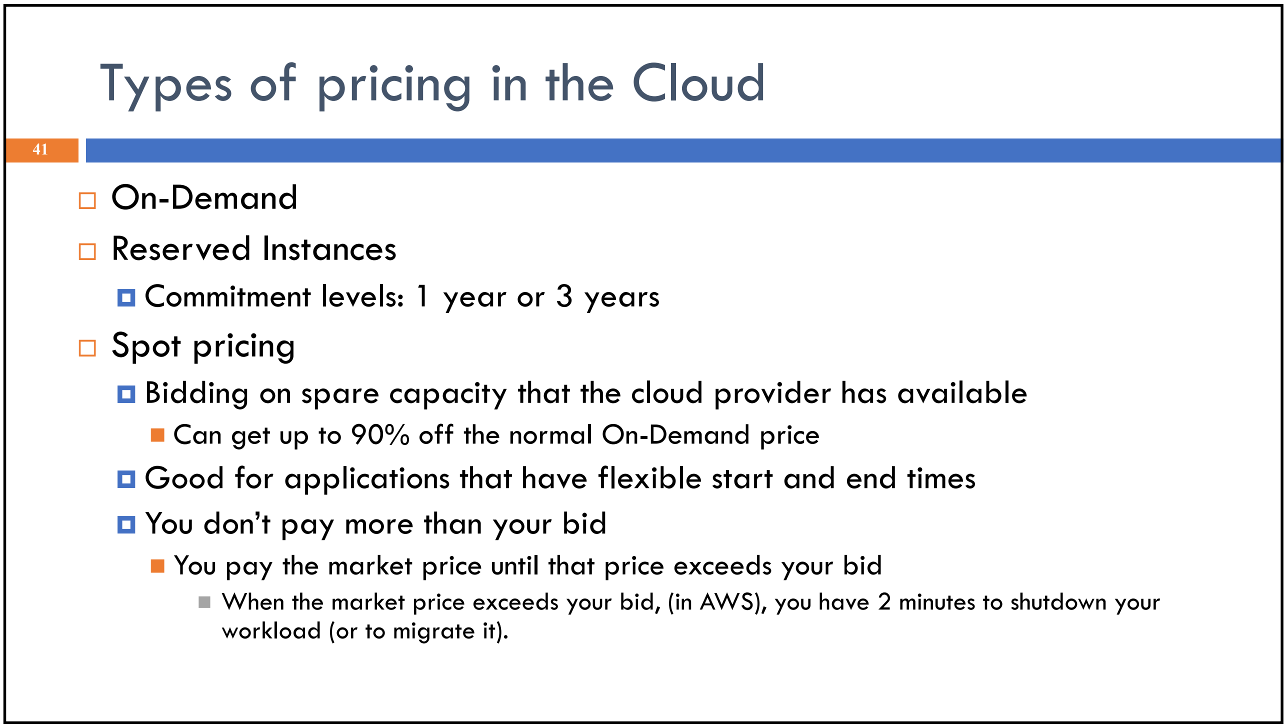
- □ Distributes traffic (web page requests)
- □ Exists in a single region – across the AZs in the region
- □ Secure, scalable and resilient
- □ Does continuous health checks
- □ Integrates with autoscaling and Route 53
- □ AWS has several types of load balancer:
  - ▣ Application
  - ▣ Network
  - ▣ Gateway
  - ▣ Classic



Auto Scaling

The autoscaling process can:

- □ Replace instances that have (performance) issues
  - ▣ Marked by the load balancer or hypervisor
  - ▣ Helps in building a self-healing infrastructure
- □ Maintain cluster size
- □ Grow cluster to meet increase in demand and scale down when needed

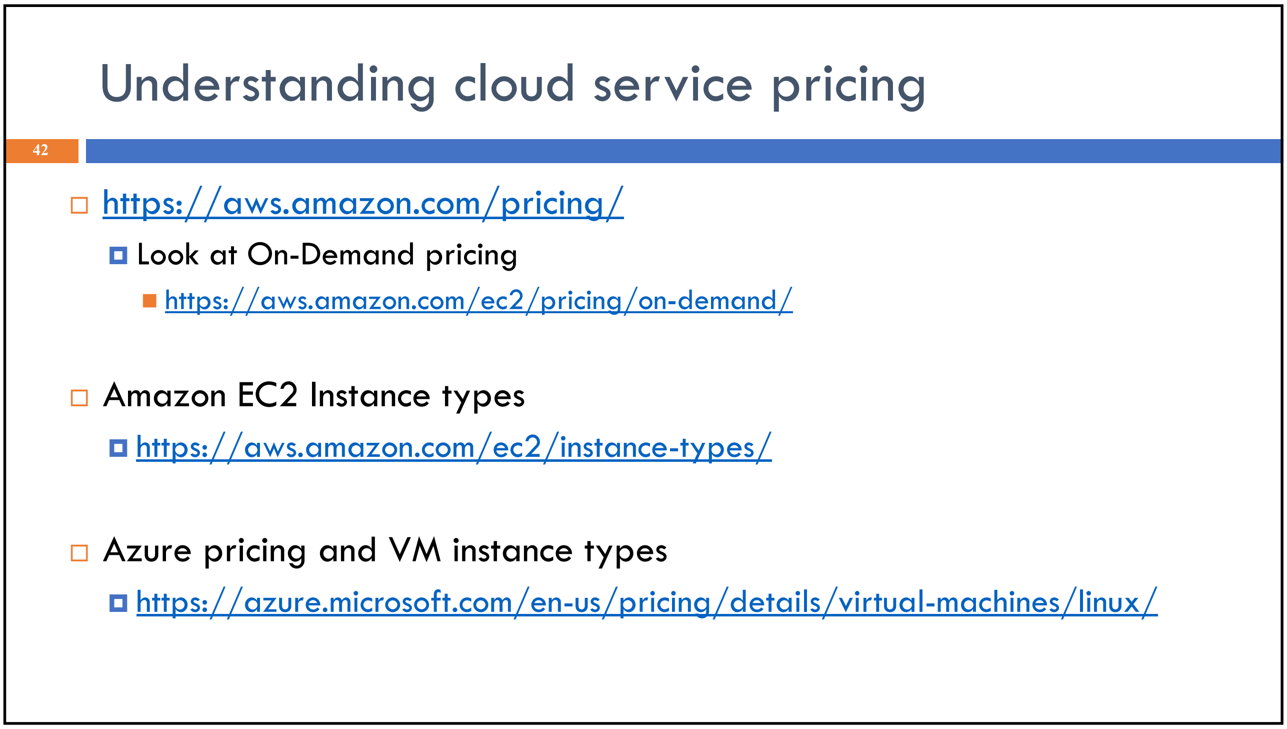


Types of pricing in the Cloud

41

- On-Demand
- Reserved Instances
  - Commitment levels: 1 year or 3 years
- Spot pricing
  - Bidding on spare capacity that the cloud provider has available
    - Can get up to 90% off the normal On-Demand price
  - Good for applications that have flexible start and end times
  - You don't pay more than your bid
    - You pay the market price until that price exceeds your bid
      - When the market price exceeds your bid, (in AWS), you have 2 minutes to shutdown your workload (or to migrate it).

41

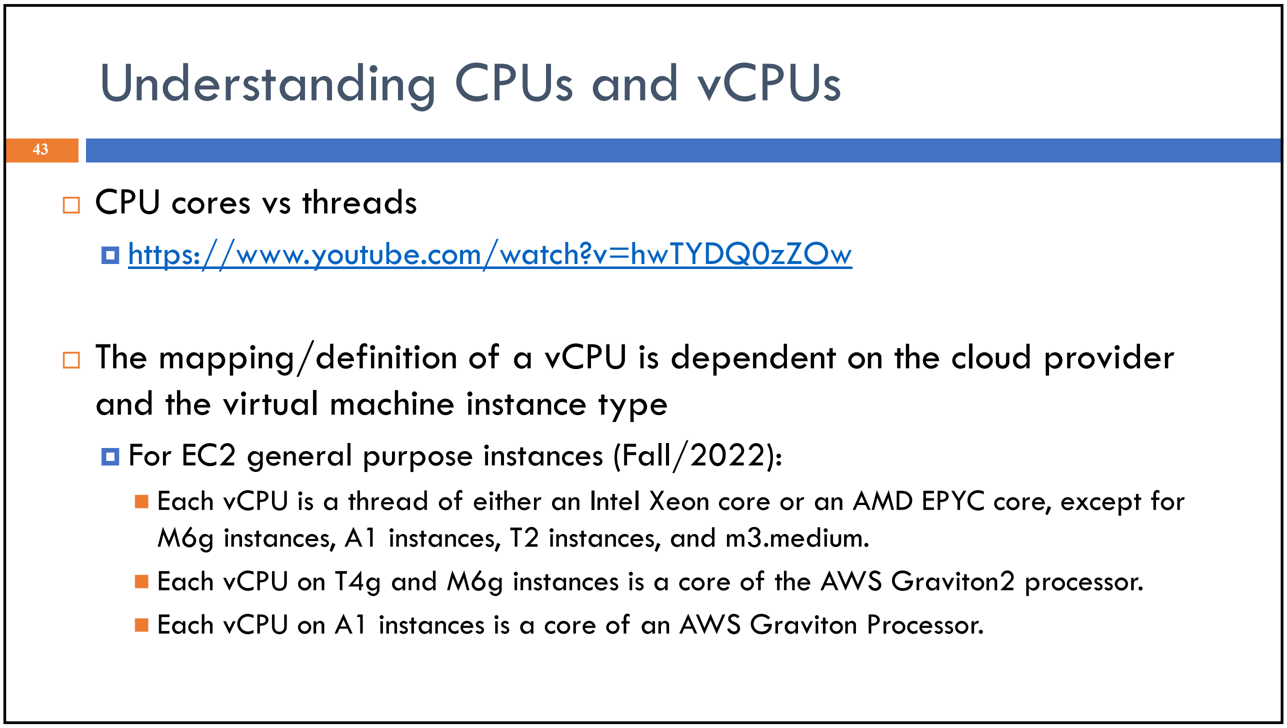


Understanding cloud service pricing

42

- https://aws.amazon.com/pricing/
  - Look at On-Demand pricing
    - https://aws.amazon.com/ec2/pricing/on-demand/
- Amazon EC2 Instance types
  - https://aws.amazon.com/ec2/instance-types/
- Azure pricing and VM instance types
  - https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/

42

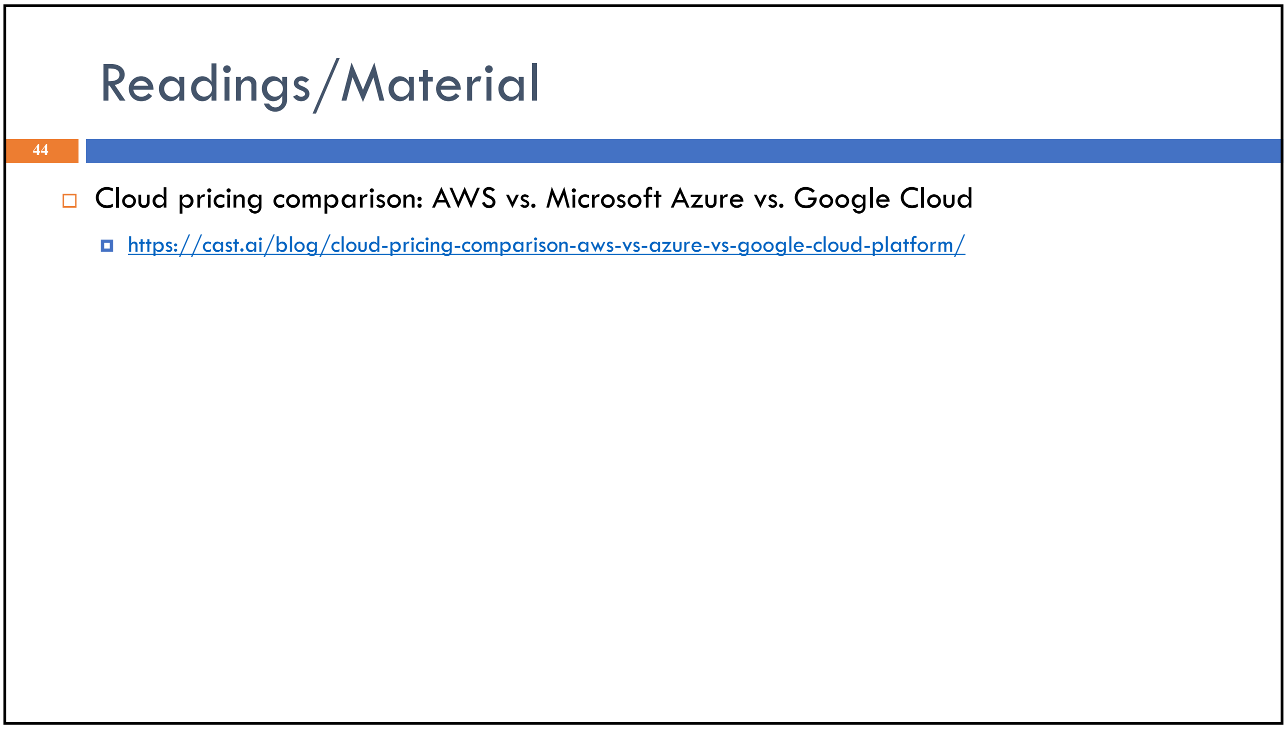


Understanding CPUs and vCPUs

43

- CPU cores vs threads
  - https://www.youtube.com/watch?v=hwTYDQ0zZOw
- The mapping/definition of a vCPU is dependent on the cloud provider and the virtual machine instance type
  - For EC2 general purpose instances (Fall/2022):
    - Each vCPU is a thread of either an Intel Xeon core or an AMD EPYC core, except for M6g instances, A1 instances, T2 instances, and m3.medium.
    - Each vCPU on T4g and M6g instances is a core of the AWS Graviton2 processor.
    - Each vCPU on A1 instances is a core of an AWS Graviton Processor.

43

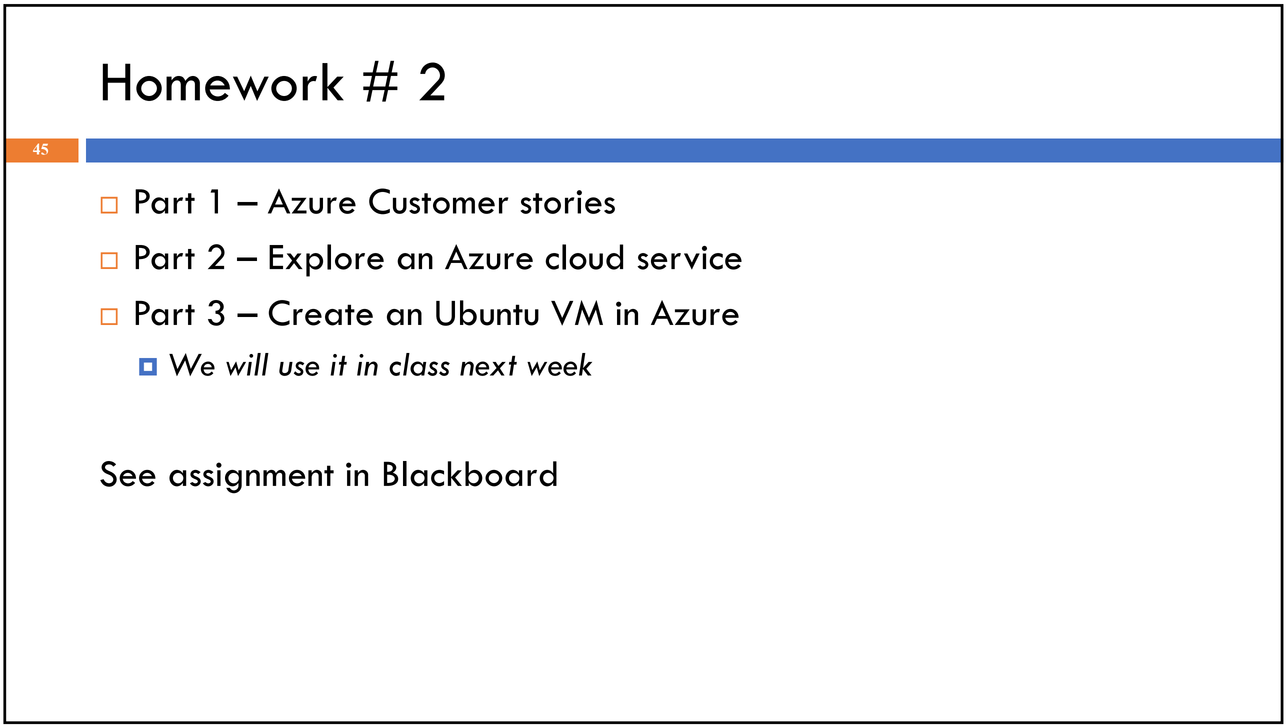


Readings/Material

44

- Cloud pricing comparison: AWS vs. Microsoft Azure vs. Google Cloud
  - https://cast.ai/blog/cloud-pricing-comparison-aws-vs-azure-vs-google-cloud-platform/

44



Homework # 2

- □ Part 1 – Azure Customer stories
- □ Part 2 – Explore an Azure cloud service
- □ Part 3 – Create an Ubuntu VM in Azure
  - *We will use it in class next week*

See assignment in Blackboard