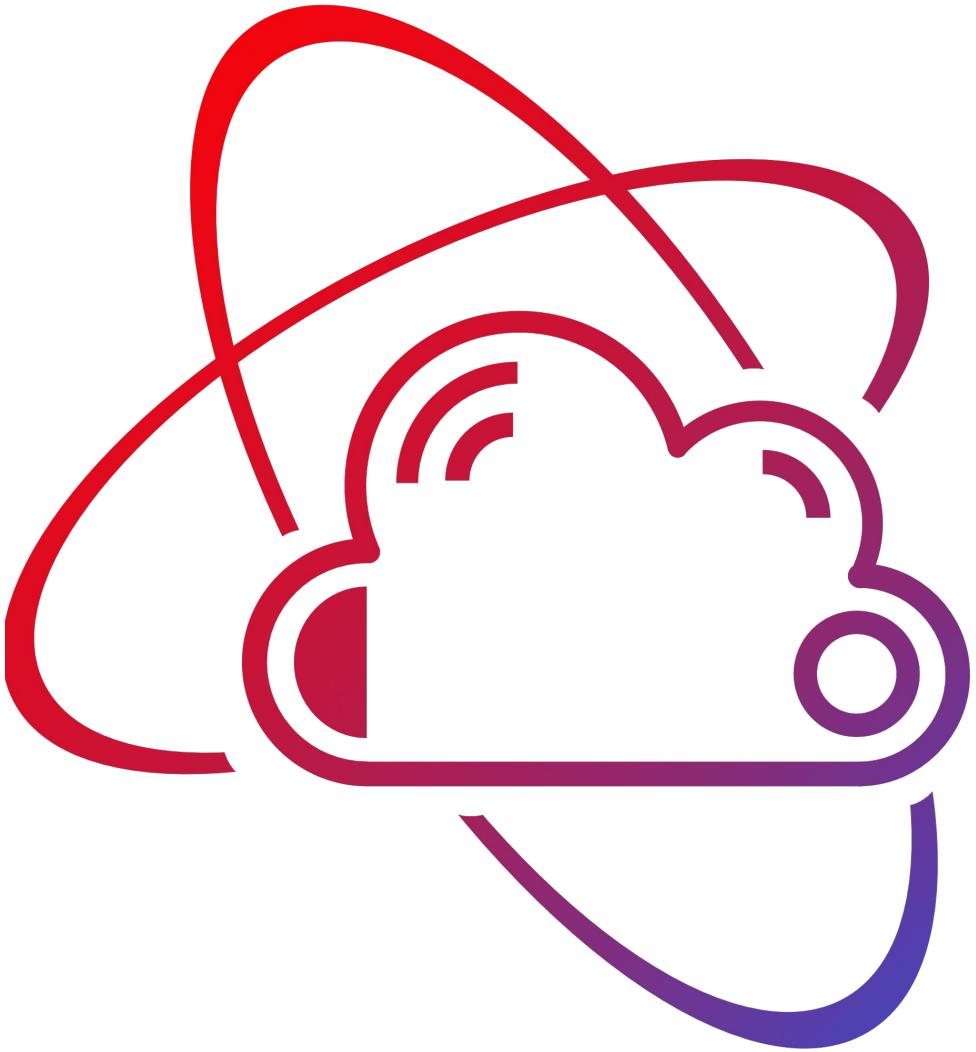
Cloud Computing Architecture

Week 1 – Introduction



SWINBURNE UNIVERSITY OF TECHNOLOGY

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer-east, and pay our respect to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

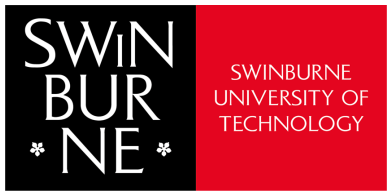
We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

THIS PRESENTATION

- AWS Section of Course
- Layout of Modules
- Week 1 Tasks
- Lab 1 Preview

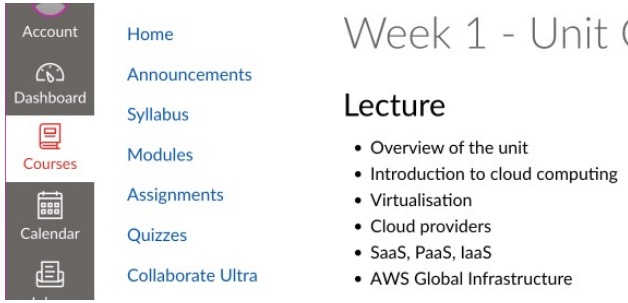
AWS Section of Course



COURSE DELIVERY

Lectures:

- *Lecture recordings and links to course videos can be found on Canvas in the 'Modules' Section.*

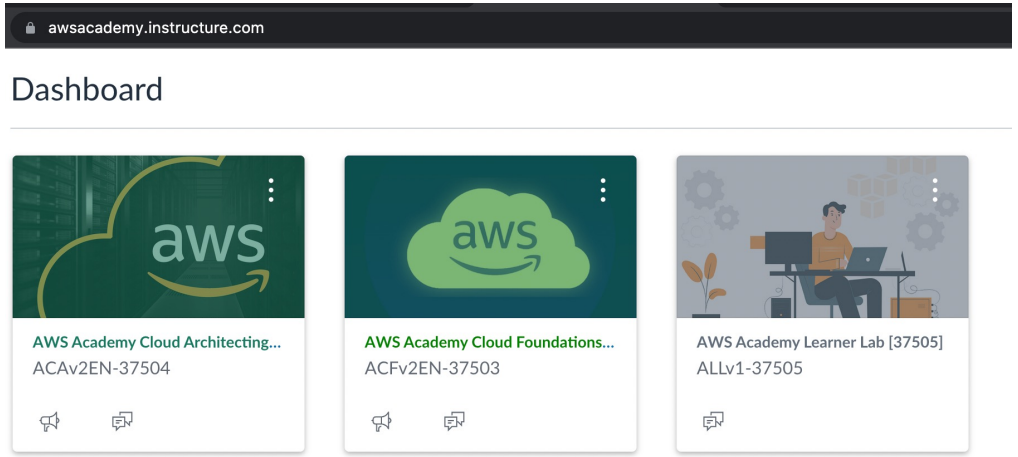


Labs:

- *Done during tutorial sections each week with Tutor.*

AWS ACADEMY CANVAS

- *Invitation emails have been sent. Join classes before your first tutorial.*
- *Extra learning materials, complimentary quizzes and labs can be accessed via Canvas shells set up by AWS Academy.*



AWS ASSIGNMENTS AND WORKING ENVIRONMENT

Assignments are to be implemented on AWS platform. Two ways to access the AWS platform:

1. ***Vocareum environment in AWS Academy (recommended):***

- *Invite emails will be sent out soon if not already sent.*
- *Accessible through Vocareum (inside AWS Canvas) – a third party site.*
- *No cost to you, you'll be given US$100 credit. Be careful how you use your AWS resources so you don't runout of credit!*
- *It is a controlled environment, some restrictions apply, however all resources for labs and assignments are available.*
- *Teaching staff will also be able to access this environment to mark or troubleshoot your assignments.*

2. ***Personal account environment (not recommended):***

- *You can use your personal AWS account to complete the assignments under free tier.*

Note: *you may have to spend money out of your pocket if you do not manage your usage properly!*

AWS CERTIFICATION

aws CERTIFIED

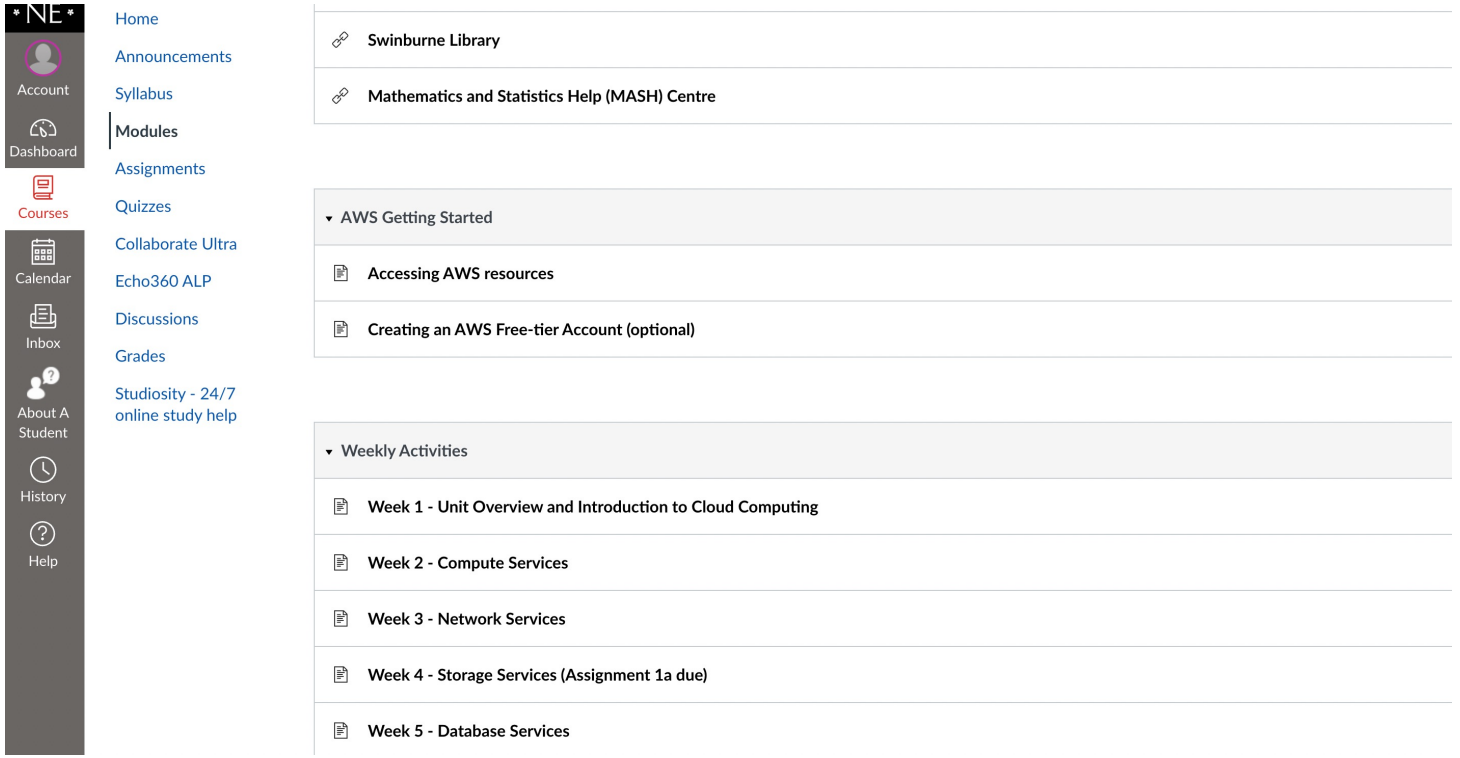
- *This unit embeds material prepared by AWS Academy covering 2 courses:*
  - *Academy Cloud Foundations (ACF) -> **AWS Cloud Practitioner Certification***
  - *Academy Cloud Architecting (ACA) -> **AWS Solutions Architect – Associate***
- *Material includes:*
  - ***Lecture notes***
  - ***Multiple-choice assessments***
  - ***Lab exercises***
- *You should explore all the materials in both of these courses to prepare for certification if you choose to do so.*

Layout of Modules

SWINBURNE UNIVERSITY OF TECHNOLOGY

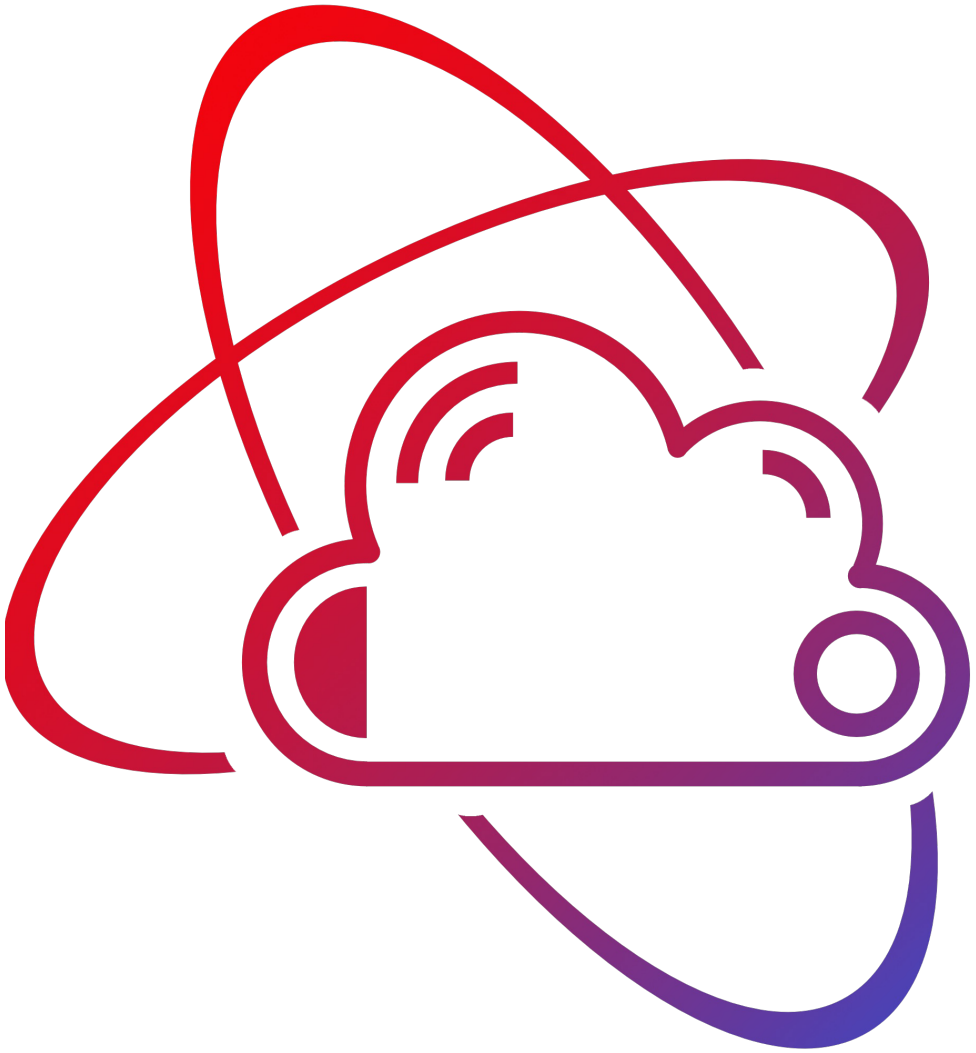
NAVIGATING THE MODULES SECTION

- *Weekly Lab Assignments and Course Materials are collated under the Modules Section.*



Cloud Computing Architecture

Introduction to Cloud Computing



SWINBURNE UNIVERSITY OF TECHNOLOGY

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer-east, and pay our respect to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

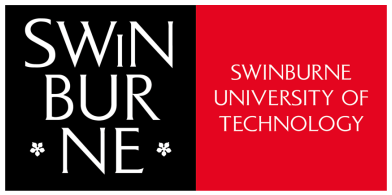
We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

Introduction to Cloud Computing

In this Presentation:

- Virtualisation and Cloud Computing
- 3 Service Models for Cloud Computing
- Advantages of Cloud Computing
- Cloud Providers

Virtualisation and Cloud Computing



What is Cloud Computing?

Definitions:

- "…the *on-demand* delivery of compute power, database storage, applications, and other IT resources through a cloud services platform *via the internet* with *pay-as-you-go* pricing."
- "*Computing as a Service* delivered over the internet."
- "…the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer *faster innovation, flexible resources, and economies of scale*."
- "…a paradigm shift away from the traditional model of purchasing, provisioning and maintaining expensive IT infrastructure, to one in which the *infrastructure is owned and maintained by a third party and delivered to the customer on demand over the internet*."

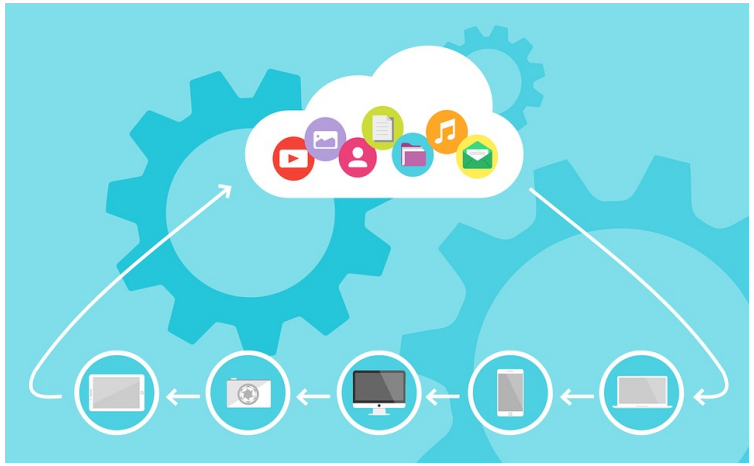


Image from: https://upload.wikimedia.org/wikipedia/commons/d/dc/Cloud-Computing_services.png

Why do we use the term “Cloud”?

History of Cloud Computing timeline:

- 1960s: Computer scientists at MIT begin experimenting with ways to share resources among multiple users.
- 1970s and 1980s: Advances in virtualization technology make it possible to create virtual machines that can run multiple operating systems on a single physical server.
- Late 1990s: The term "cloud computing" is coined as companies like Salesforce.com and Amazon begin offering web-based services that allow customers to access software and data over the internet.
- Early 2000s: Companies like Google and Microsoft begin offering cloud-based services like email and storage.
- Today: Cloud computing is a major force in the technology industry, with companies of all sizes using cloud services to run their applications and store their data.

Image from: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQ-D207CPLJUwNUr2n6agG3nzNhhvd4tfo9ug&usqp=CAU

Cloud Computing: Before and After

Before Cloud Computing

- Hardware solutions are physical. This means they require:
  - Space
  - Staff
  - Physical security
  - Planning
  - Capital expenditure
  - Guessing at theoretical maximum peaks
- Is there enough resource capacity?
- Do we have sufficient storage? What if your needs change?

Before

After

After Cloud Computing

- Hardware solutions are virtual, eliminating the need for physical space, staff, and physical security
- Scalability allows for easy adjustments to meet changing resource and storage needs
- No capital expenditure required for purchasing and maintaining physical hardware
- Pay-as-you-go pricing model allows for cost savings by only paying for resources used
- Easy access to a vast pool of resources and storage capacity
- Automatic updates and maintenance of hardware and software.

What is Virtualisation in Cloud Computing?

Virtualisation in Cloud Computing Definition:

'...the process of creating a virtual version of a physical resource, such as a server, storage device, or network. This virtual resource can then be used in place of the physical resource, allowing for efficient use of resources and increased flexibility in managing them.

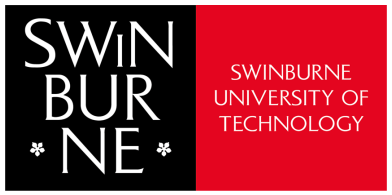
Virtualization enables the abstraction of logical and physical resources, which allows multiple workloads to share the same underlying physical resources. This is a key technology that enables cloud computing and allows for the dynamic allocation of resources on-demand.'

Image from: https://www.educba.com/what-is-virtualization-in-cloud-computing/

Types of Virtualisation in Cloud Computing

- Memory virtualisation
  - Virtual memory
- Storage virtualisation
  - Logical disks and file systems
  - Networked attached storage
  - Cloud storage
- Network virtualisation
  - VLANs (segmentation), VPNs (tunneling), VPCs
- Operating System virtualisation (virtual desktop)
  - Multiple OS on the one computer (host-guest)
- Machine virtualisation
  - Hyper-visors (e.g. Hyper-V, VMWare) allow multiple servers (virtual machines) to be run on a single “metal” computer.

3 Service Models for Cloud Computing



What are the main Cloud Service Models?

Infrastructure as a service (IaaS):

Services in this category are the basic building blocks for cloud IT and typically provide you with access to networking features, computers (virtual or on dedicated hardware), and data storage space. It is the most similar to existing IT resources that many IT departments and developers are familiar with today.

Platform as a service (PaaS):

Services in this category reduce the need for you to manage the underlying infrastructure (usually hardware and operating systems) and enable you to focus on the deployment and management of your applications.

Software as a service (SaaS):

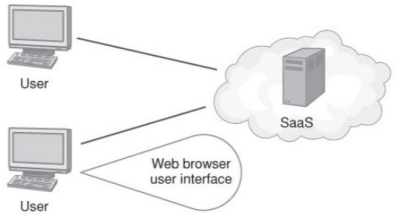
Services in this category provide you with a completed product that the service provider runs and manages. A common example of a SaaS application is web-based email, where you can send and receive email without managing feature additions to the email product or maintaining the servers and operating systems that the email program runs on.

Image from: https://cloudcomputinggate.com/saas-paas-iaas-examples/

How does Software as a Service work?

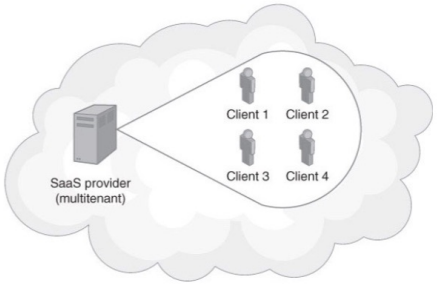
What are Multitenant SaaS Solutions?

- '...cloud-based software applications that are designed to be shared by multiple customers, or tenants. In a multitenant architecture, all tenants share the same instance of the software, but each tenant's data is isolated and kept separate from other tenants' data.'



SaaS applications are often multitenant solutions.

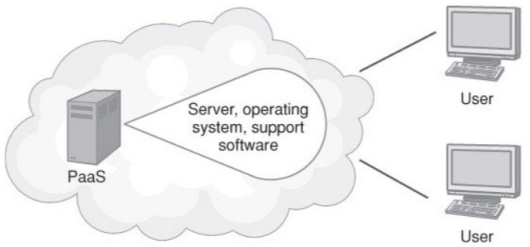
- Within the single SaaS application, two or more tenant companies share the same server resources.
- Each tenant can customize their own version of the software for their clients.



How does Platform as a Service work?

PaaS definition:

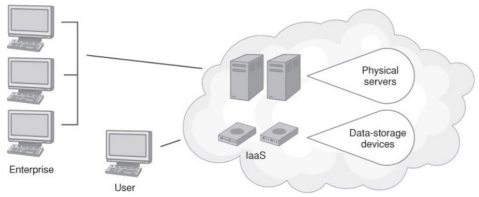
- PaaS provides the underlying hardware technology, such as one or more servers (or virtual servers), operating systems, database solutions, developer tools, and network support, for developers to deploy their own solutions.
- The hardware and software within a PaaS solution is managed by the platform provider.
- Developers need not worry about performing hardware or operating system upgrades. Instead, developers can focus on their own applications.
- Provide a collection of hardware and software resources that developers can use to build and deploy applications within the cloud.
- Depending on their needs, developers may use a Windows-based PaaS solution or a Linux-based PaaS.



How does Infrastructure as a Service work?

IaaS definition:

- An IaaS provider makes all of the computing hardware resources available, and the customers, in turn, are responsible for installing and managing the systems, which they can normally do, for the most part, over the Internet.
- IaaS provides a virtual data center within the cloud.
- IaaS provides servers (physical and virtualized), cloud-based data storage, and more.
- Developers must install their own operating system, database management software, and support software.
- Then the developers (or the company's system administrators) must manage both the hardware and the software.



What are the three Cloud Deployment Models?

"All-In" Public Cloud:

- A cloud-based application is fully deployed in the cloud, and all parts of the application run in the cloud. Applications in the cloud have either been created in the cloud or have been migrated from an existing infrastructure.

Hybrid:

- A hybrid deployment is a way to connect infrastructure and applications between cloud-based resources and existing resources that are not located in the cloud. The most common method of hybrid deployment is between the cloud and existing on-premises infrastructure (sometimes called on-prem).

Private Cloud (On-premises):

- When you run a cloud infrastructure from your own data center, that's called on-premises or private cloud. While this kind of deployment lacks many of the benefits of cloud computing, it does provide dedicated resources and is a popular choice for organizations who need to meet certain compliance standards.

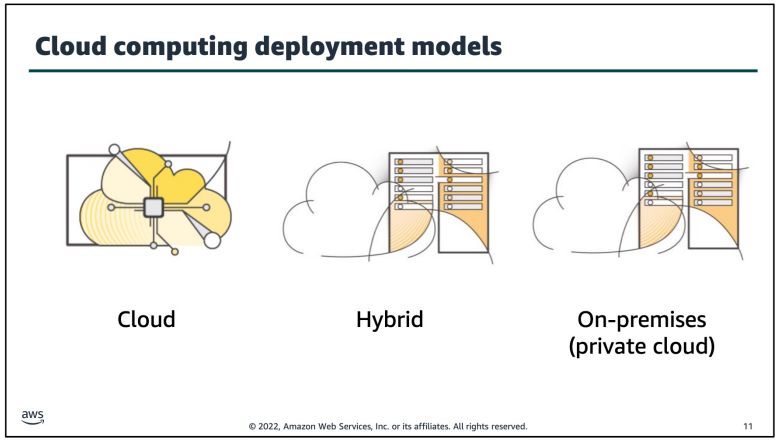


Image from: AWS Academy - ACF Module 1

Advantages of Cloud Computing

What are the advantages of cloud computing? (AWS)

Advantage #1—Trade capital expense for variable expense:

Capital expenses (capex) are funds that a company uses to acquire, upgrade, and maintain physical assets such as property, industrial buildings, or equipment.

Advantage #2—Benefit from massive economies of scale:

By using cloud computing, you can achieve a lower variable cost than you can get on your own.

Advantage #3—Stop guessing capacity:

Eliminate guessing about your infrastructure capacity needs.

Advantage #4—Increase speed and agility:

In a cloud computing environment, new IT resources are only a click away, which means that you reduce the time it takes to make those resources available to your developers from weeks to just minutes.

Advantage #5—Stop spending money on running and maintaining data centers:

Focus on projects that differentiate your business instead of focusing on the infrastructure.

Advantage #6—Go global in minutes:

You can deploy your application around the world with just a few clicks

What are the benefits of cloud-based platforms?

- Scalability.

On demand resource scaling.

- Redundancy.

Servers, storage, and networks.

- Cost benefits from resource pooling.

Shares IT resources across a very large number of companies, which provides cost savings to each.

- Outsourced server management.

Provides an IT staff who maintain operating systems and underlying support software.

- Low cost of entry.

Companies do not need to invest in their own IT data center.

Cloud Providers

Who are some major Cloud Providers?

Amazon Web Services (AWS):

- AWS is a collection of remote computing services (also called web services) that make up a cloud computing platform, offered by Amazon.com. These services operate from 31 geographical regions across the world.

Microsoft Azure:

- Azure is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through a global network of Microsoft-managed data centers. It has over 60 regions around the world.

Google Cloud Platform (GCP):

- GCP is a collection of remote computing services that make up a cloud computing platform, offered by Google. These services operate from over 35 geographical regions across the world.

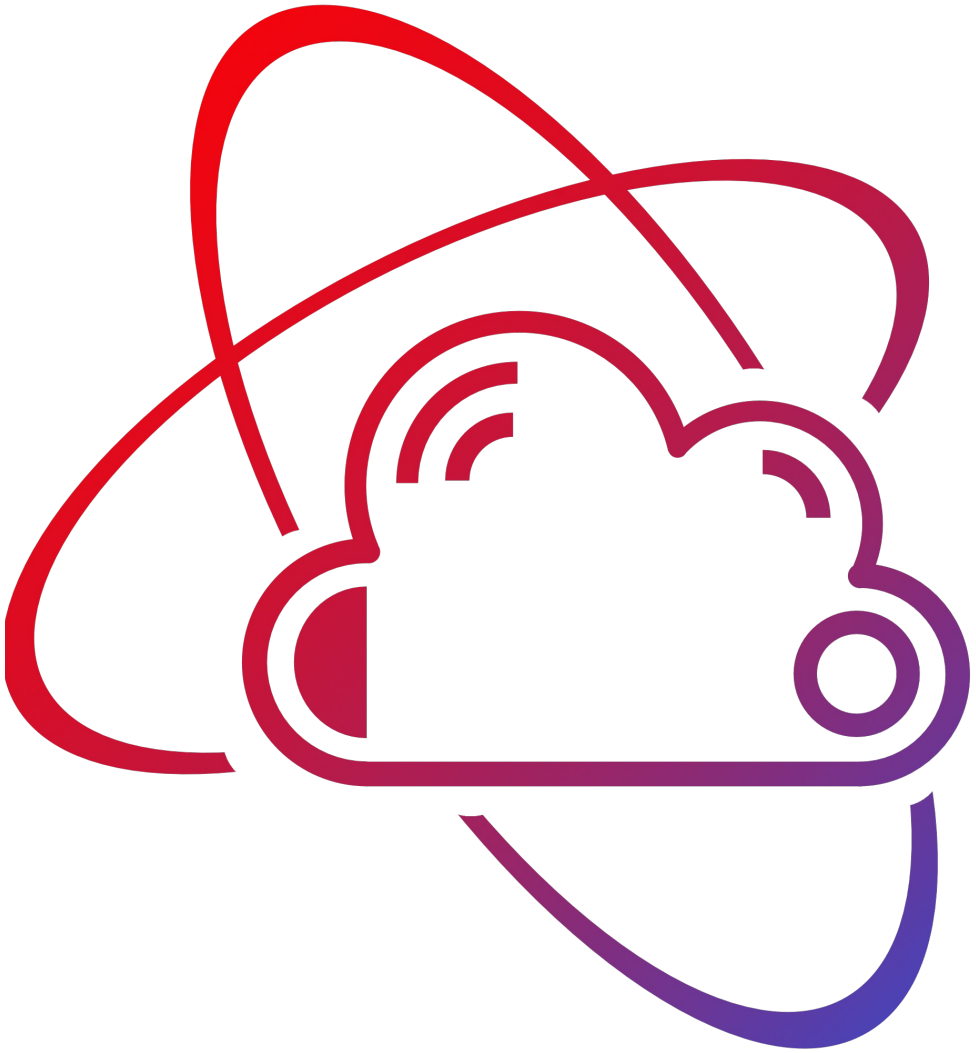
Oracle Cloud:

- Oracle Cloud is a collection of managed services from Oracle Corporation, including data storage, application development, and business analytics services. It offers a range of services including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). It operates from 41 public cloud regions around the world.

Image from: https://blog.clinked.com/top-cloud-providers

Cloud Computing Architecture

Introducing Amazon Web Services



SWINBURNE UNIVERSITY OF TECHNOLOGY

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer-east, and pay our respect to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

Introducing Amazon Web Services

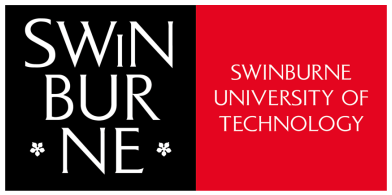
In this Presentation:

– AWS Global Infrastructure

– AWS Core Services

Image from: https://allcode.com/top-aws-services/

AWS Core Services



Introducing Amazon Web Services

What is AWS?

- AWS is a **secure cloud platform** that offers a **broad set of global cloud-based products**.
- AWS provides you with **on-demand access** to compute, storage, network, database, and other IT resources and management tools.
- AWS offers **flexibility**.
- You **pay only for the individual services you need**, for **as long as you use them**.
- AWS services **work together** like building blocks.

aws

24

SWINBURNE UNIVERSITY OF TECHNOLOGY

Introducing Amazon Web Services – On Demand

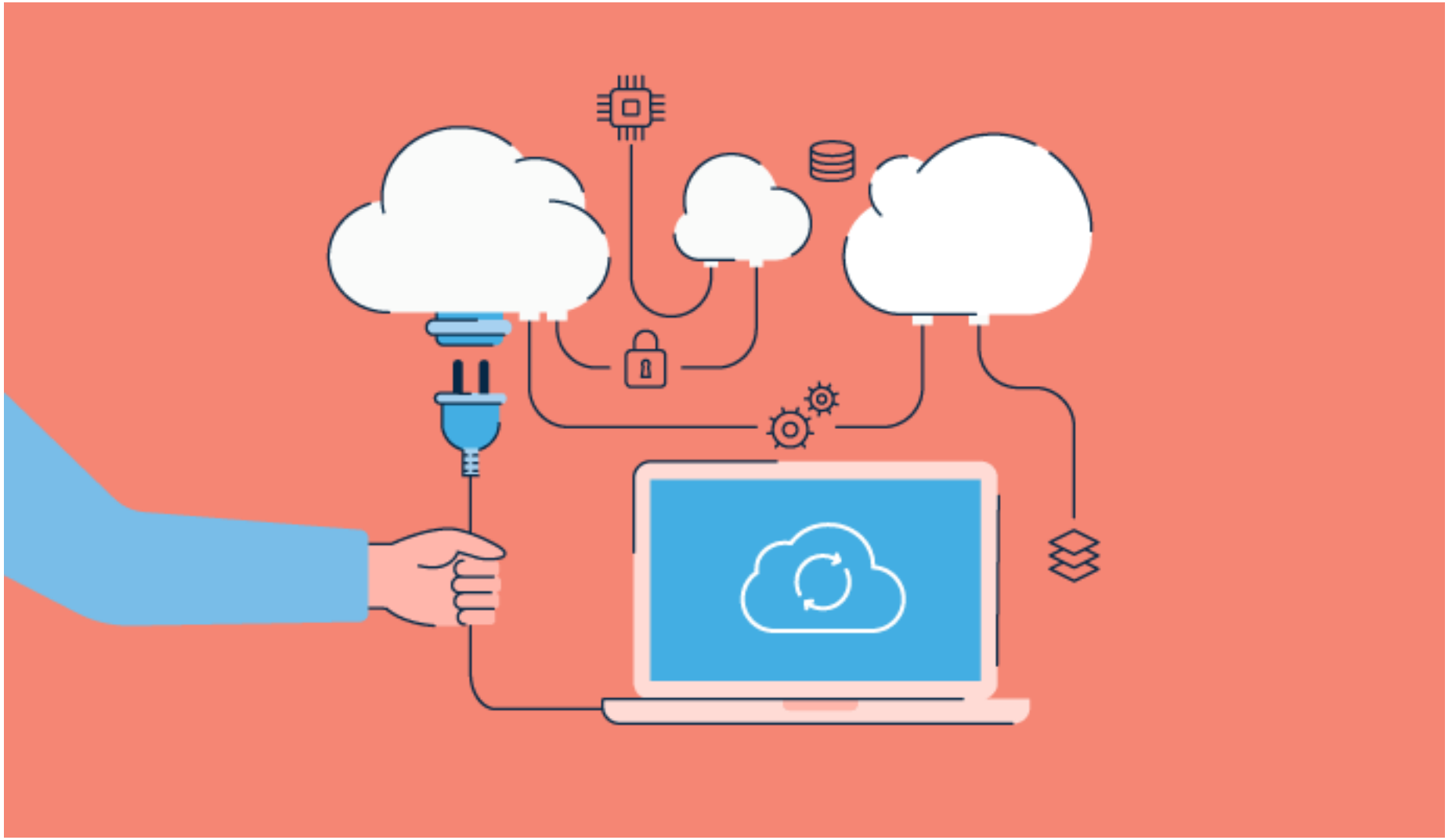
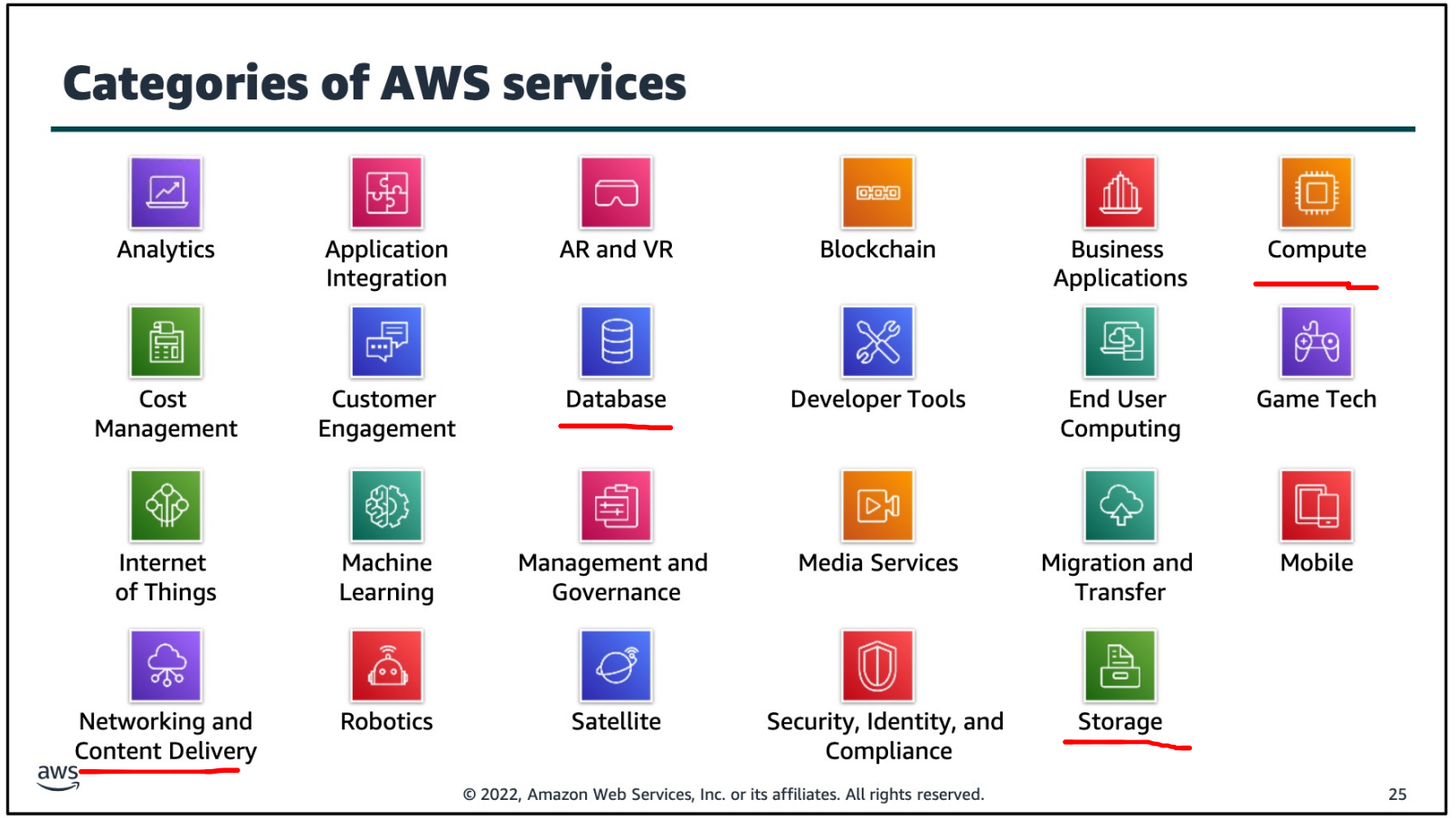


Image from:https://www.g2.com/articles/cloud-computing

Introducing Amazon Web Services – Flexible and Modular

Image from:https://stock.adobe.com/fi/images/cloud-computing-technology-internet-network-data-server-storage-virtualization-scalability-flexibility-reliability-security-software-hardware-infrastructure-platform-service-backup/575386330

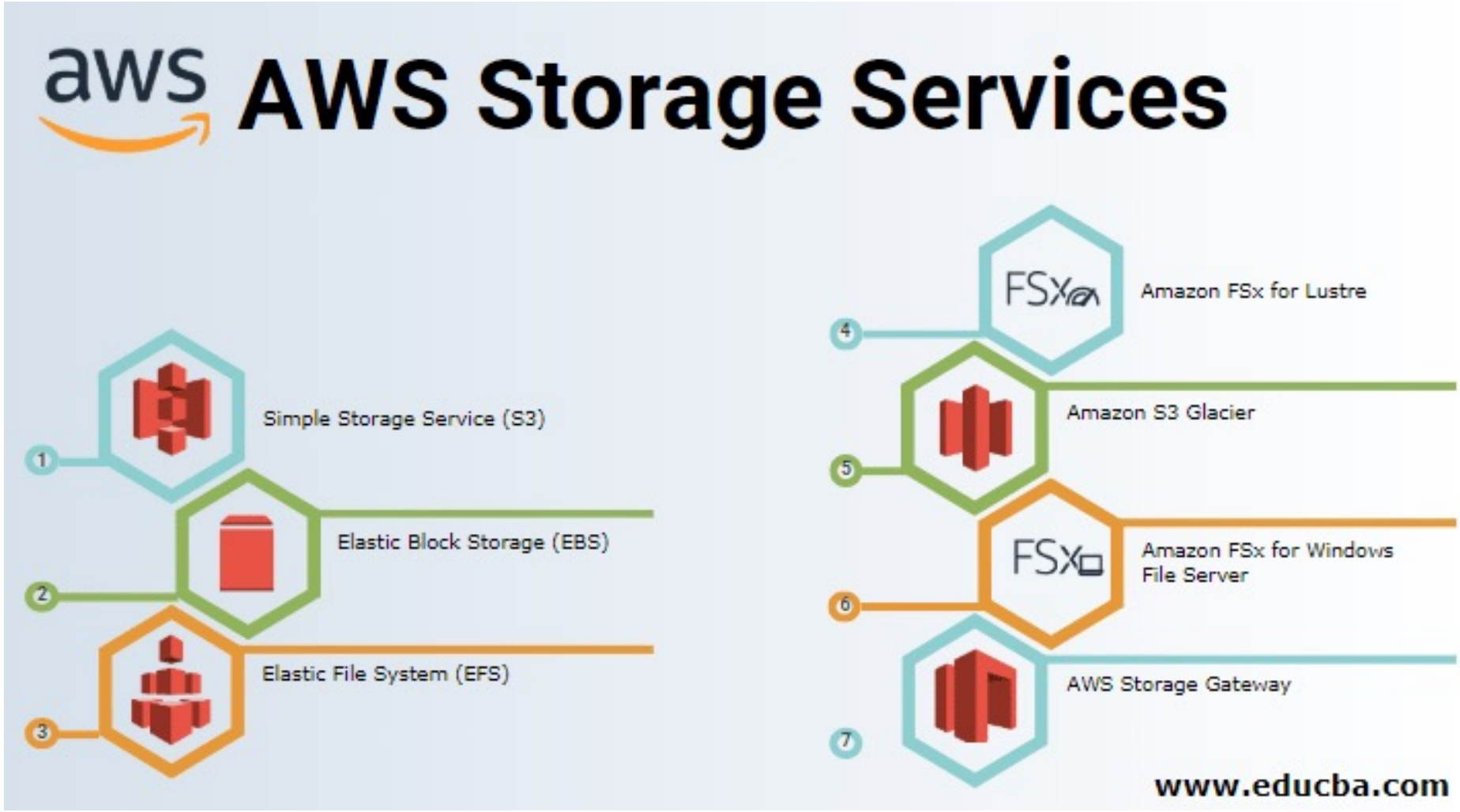
Introducing Amazon Web Services



Introducing Amazon Web Services - Compute

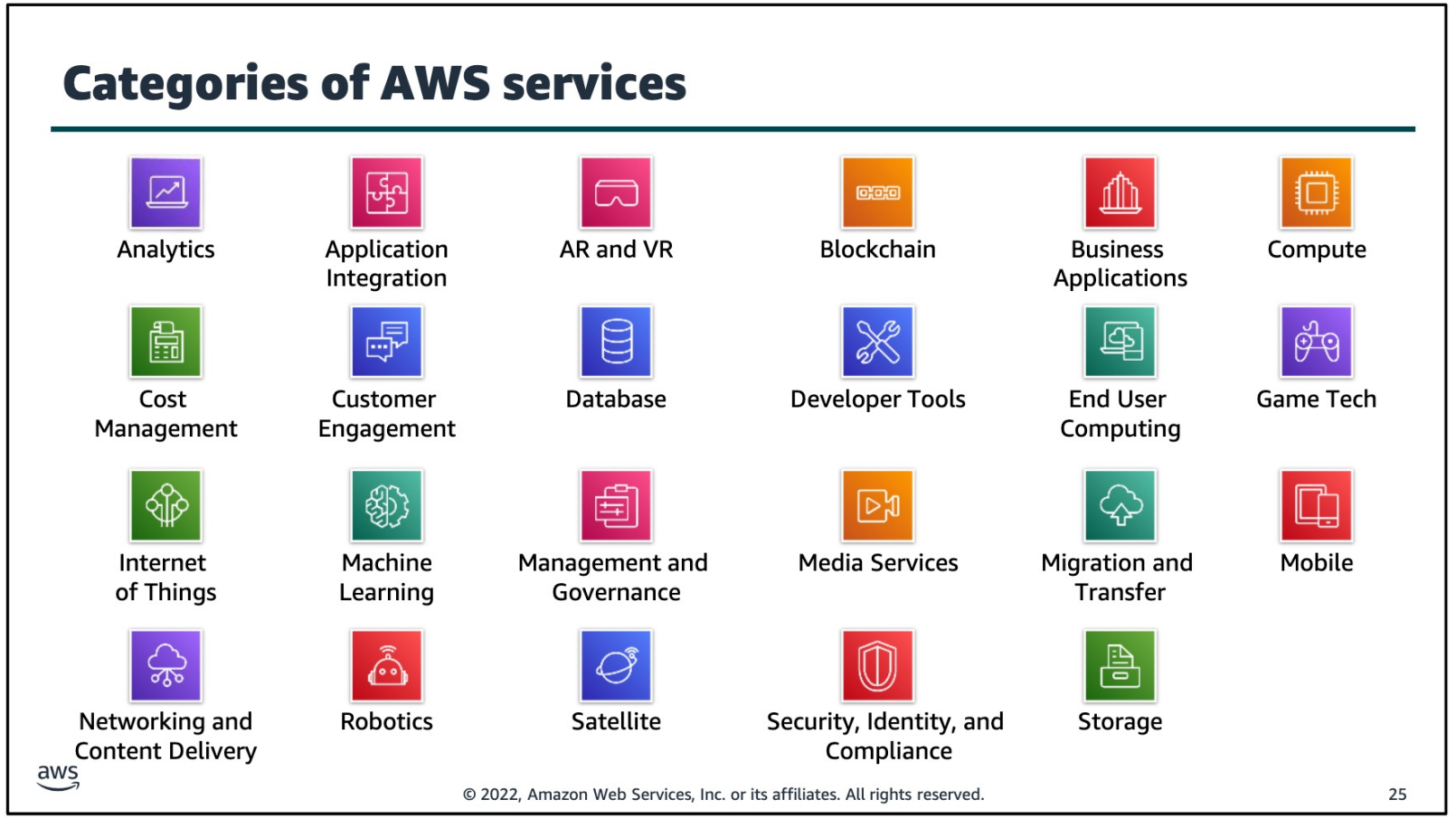
https://dev.to/aws-builders/all-you-need-to-know-about-aws-compute-services-17gf

Introducing Amazon Web Services – Storage

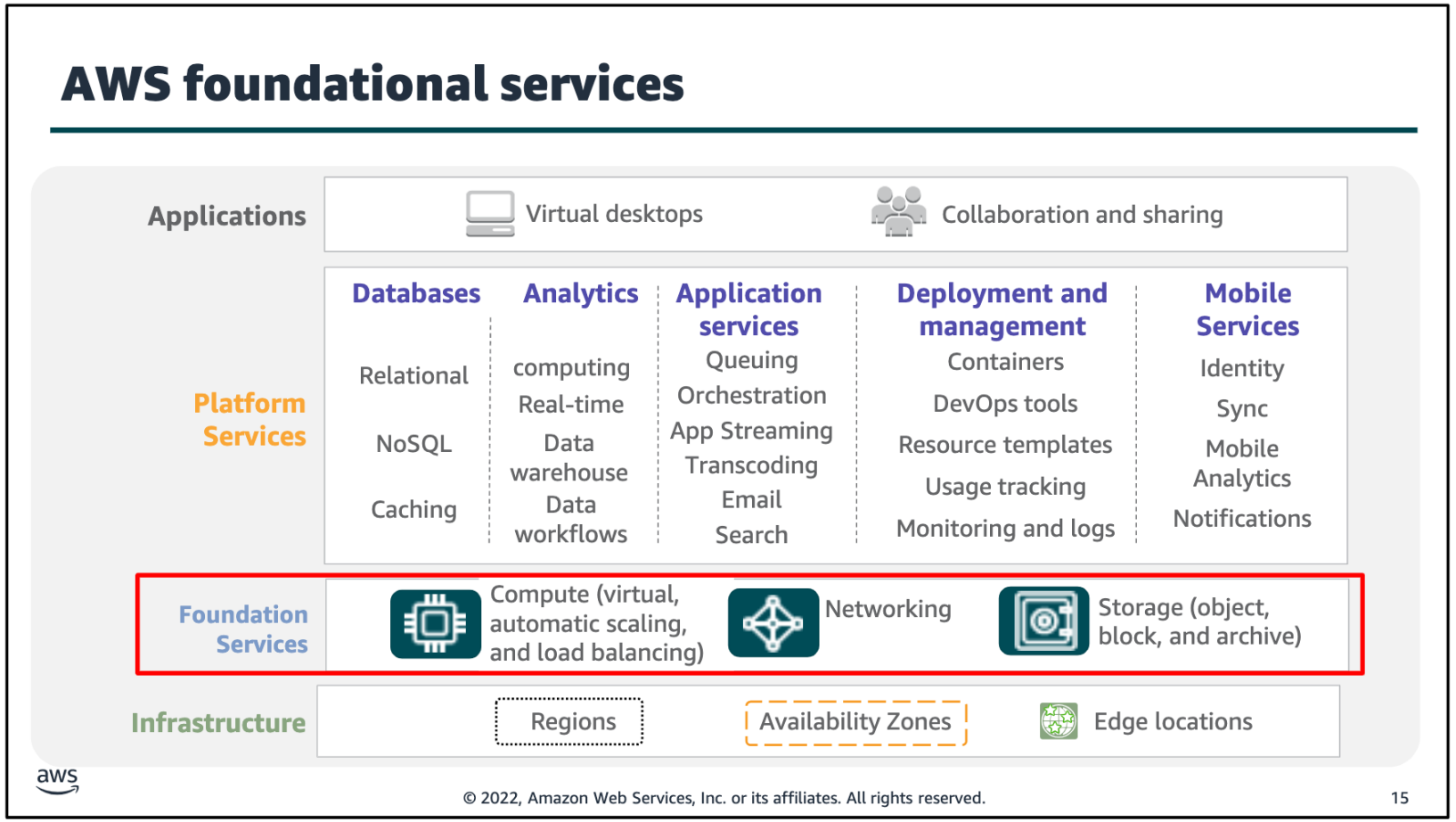


https://www.educba.com/aws-storage-services/

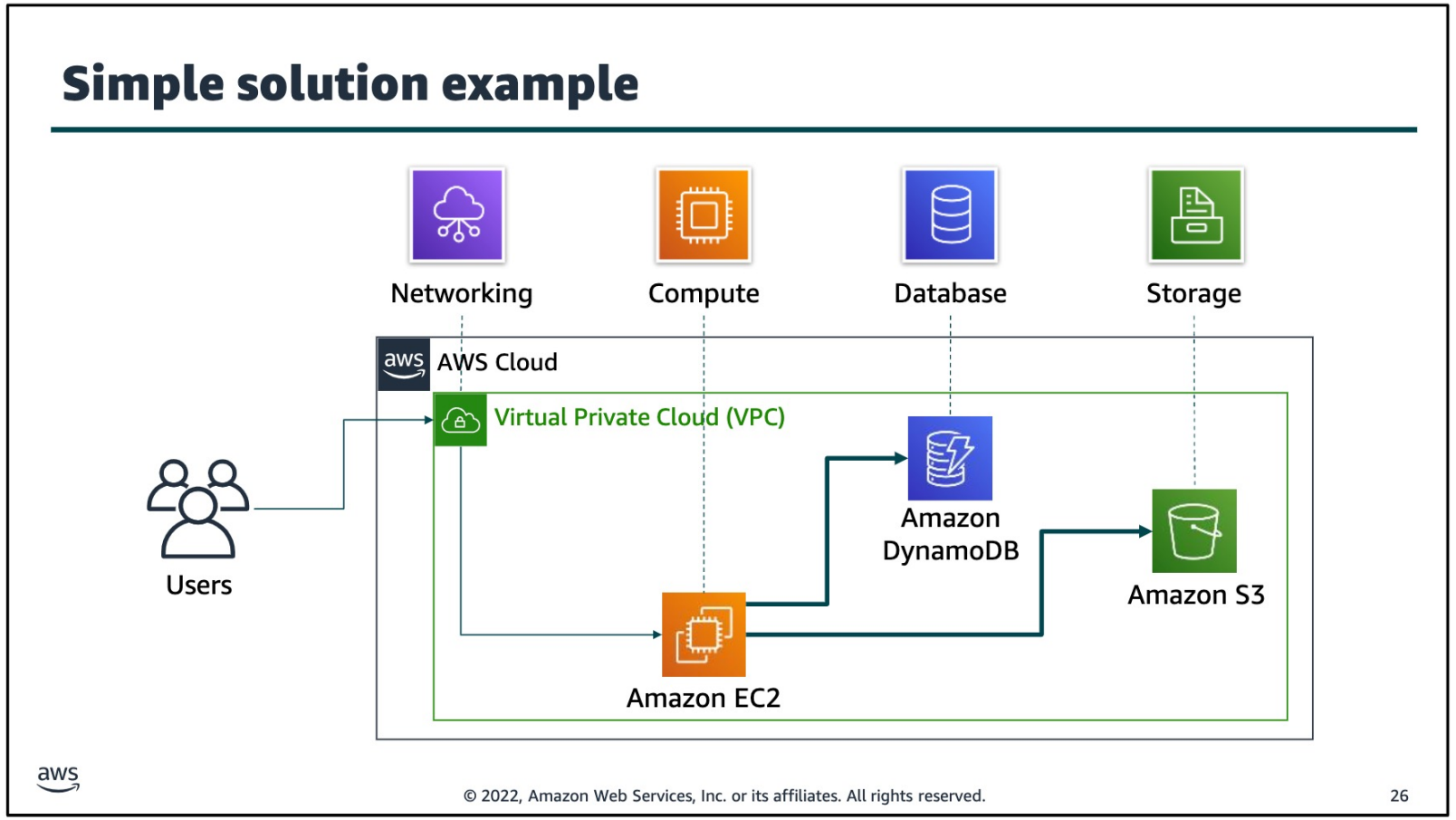
Introducing Amazon Web Services – Databases and Networking



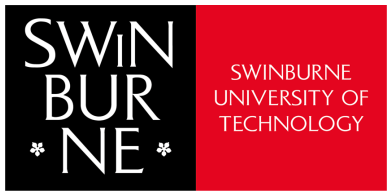
Introducing Amazon Web Services



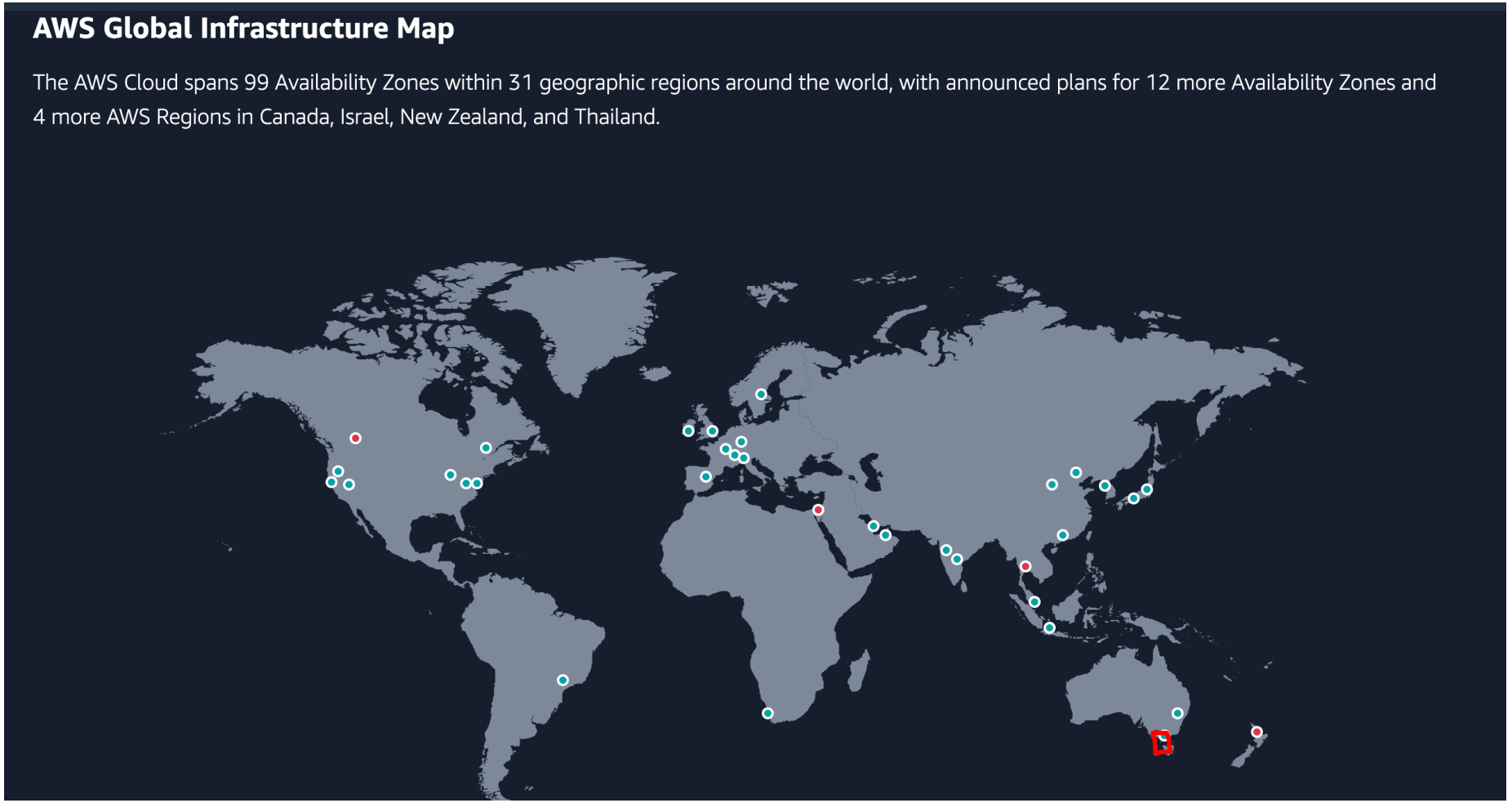
Introducing Amazon Web Services



AWS Global Infrastructure



Introducing Amazon Web Services – Regions and AZs



Introducing Amazon Web Services – Data Centers

Introducing Amazon Web Services

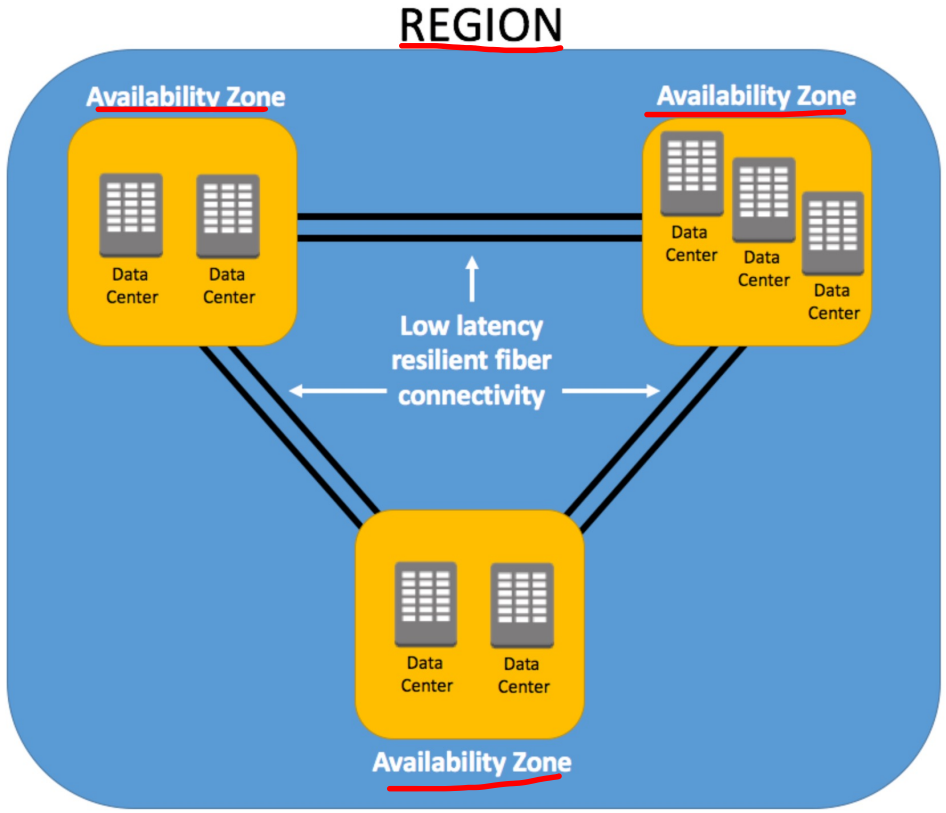


Image from: https://cloudacademy.com/blog/aws-global-infrastructure/

Content Today

- Unit overview and Canvas
  - Unit outline
  - AWS Cloud Practitioner certification
  - Labs
  - Assessment
- Introducing cloud computing
- AWS Global Infrastructure

SWINBURNE UNIVERSITY OF TECHNOLOGY

1

Content Today

- Unit overview and Blackboard
- **Introducing cloud computing**
  - ☐ Virtualisation and Cloud Computing
  - ☐ 3 Models of cloud Computing – IaaS, PaaS, SaaS
  - ☐ Advantages of Cloud Computing
  - ☐ Cloud Providers
- AWS Global Infrastructure

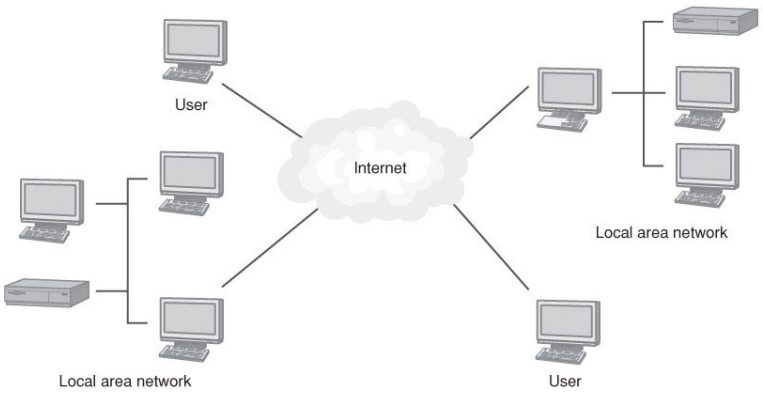
SWINBURNE UNIVERSITY OF TECHNOLOGY

2

Why Use the Term "Cloud"

For years developers and network administrators have represented the Internet as a cloud.

e.g. iCloud



SWINBURNE UNIVERSITY OF TECHNOLOGY

What is Cloud Computing?

- Cloud computing is the ***on-demand*** delivery of compute power, database storage, applications, and other IT resources through a cloud services platform ***via the internet*** with ***pay-as-you-go*** pricing.
- Computing as a ***service*** delivered over the internet
- Virtualisation
- Automation

SWINBURNE UNIVERSITY OF TECHNOLOGY

The most basic way to define what the "cloud" is that it is a computer located somewhere else that is accessed via the Internet and utilized in some way. *Web services* is also another name for what people call *the cloud*.

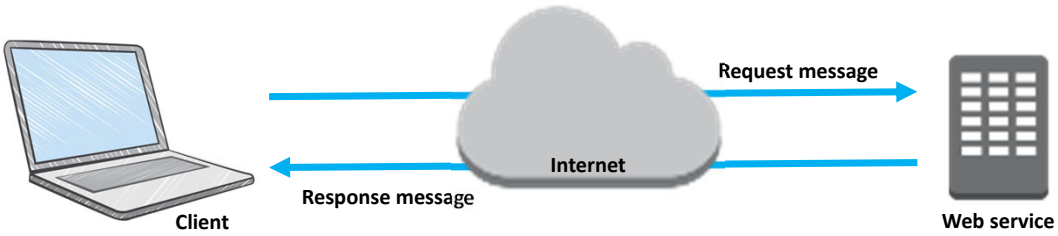
The cloud is comprised of server computers located in large data centers in different locations around the world. When you use a cloud service like Amazon Web Services (AWS), you are utilizing the computers owned by AWS. AWS is a cloud services provider.

The computers contain various technology features and services, like building blocks, that can be used to assemble solutions that help a user meet their business goals and technology requirements. With cloud computing, organizations can consume on-demand computing and storage resources rather than building, operating, and improving infrastructure on their own.

For more information, see https://aws.amazon.com/what-is-cloud-computing/.

What are Web Services?

A **web service** is any piece of software that makes itself available over the internet and uses a **standardized format** (XML or JSON) for the request and the response of an **API interaction**.



SWINBURNE UNIVERSITY OF TECHNOLOGY

A web service is any service that:

- Is available over the Internet or private (intranet) networks
- Uses a standardized data format (XML or JSON)
- Is not tied to any one operating system or programming language
- Is self-describing via an interface definition file
- Is discoverable

Before Cloud Computing

Cloud computing enables you to **stop thinking of your infrastructure as hardware**, and instead **think of it (and use it) as software.**

SWINBURNE UNIVERSITY OF TECHNOLOGY

Before cloud computing, you would have to provision capacity based on guessing theoretical maximum peaks. If you didn't meet your projected maximum peaks, or you exceeded them, you would be paying for expensive resources that would stay idle or have insufficient capacity to meet your needs.

Before Cloud Computing

- Hardware solutions are **physical**. This means they require:
  - Space
  - Staff
  - Physical security
  - Planning
  - Capital expenditure
- Guess at theoretical maximum peaks
  - Is there enough resource capacity?
  - Do we have sufficient storage?

What if your needs change?

*You have to go through the **time, effort, and cost** required to change all these.*

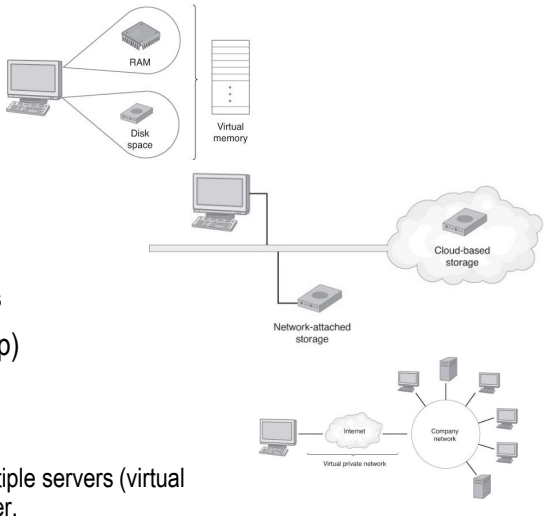
SWINBURNE UNIVERSITY OF TECHNOLOGY

Managing hardware takes away time and resources you could be using to improve your architecture and your application.

For example, before cloud computing, if you wanted to provision a new web site, you would have to go out and buy the hardware, rack and stack it, put it in a data center, and then manage it or have someone else manage it. This approach is very expensive.

Virtualisation – is *everywhere* in computing

- Memory virtualisation
  - Virtual memory
- Storage virtualisation
  - Logical disks and file systems
  - Networked attached storage → **Cloud storage**
- Network virtualisation
  - VLANs (segmentation), VPNs (tunneling), **VPCs**
- Operating System virtualisation (virtual desktop)
  - Multiple OS on the one computer (host-guest)
- Machine virtualisation
  - Hyper-visors (e.g. Hyper-V, VMWare) allow multiple servers (virtual machines) to be run on a single "metal" computer.



SWINBURNE UNIVERSITY OF TECHNOLOGY

https://www.youtube.com/watch?v=KXkBZCe699A

8

Motivation to Virtualize

- Increased device utilization (particularly CPU utilization)
- Decreased device footprint
- Decreased power consumption
- Simplified operating system and application administration
- Ease of software provisioning and patch releases
- Device and storage scalability
- Increased user access to key resources
- Increased flexibility in supporting multiple operating system environments
- Improved use and management of software licenses
- Improved utilization reporting, which leads to improved capacity planning
- Improved disaster recovery and business continuity

SWINBURNE UNIVERSITY OF TECHNOLOGY

9

Virtualisation and Automation

- Virtual environments are defined in software
- Software enables the *creation* and *modification* of virtual environments to be automated.
- Example: Extra web servers are automatically ‘stood up’ when demand increases.

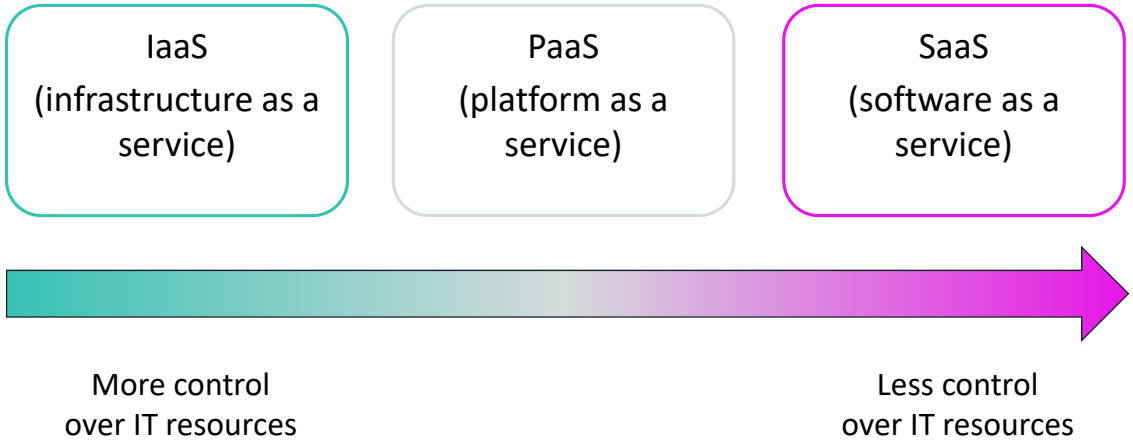
SWINBURNE UNIVERSITY OF TECHNOLOGY

Not All Applications suit Virtualization

- **Applications with unique hardware requirements**: If an application requires a unique device or hardware device driver, the virtualization software may be unable to support the device.
- **Graphics-intensive applications**: If an application is graphics intensive, such as a 3-D modeling program, the virtual device drivers may slow down the I/O processing to an unacceptable level.

SWINBURNE UNIVERSITY OF TECHNOLOGY

Cloud service models



There are three main cloud service models. Each model represents a different part of the cloud computing stack and gives you a different level of control over your IT resources:

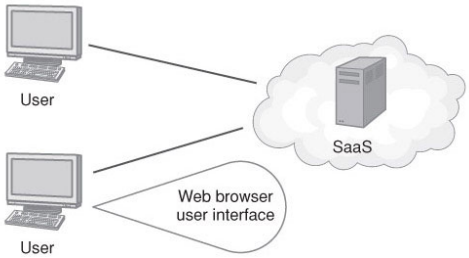
- **Infrastructure as a service (IaaS)**: Services in this category are the basic building blocks for cloud IT and typically provide you with access to networking features, computers (virtual or on dedicated hardware), and data storage space. IaaS provides you with the highest level of flexibility and management control over your IT resources. It is the most similar to existing IT resources that many IT departments and developers are familiar with today.

- **Platform as a service (PaaS)**: Services in this category reduce the need for you to manage the underlying infrastructure (usually hardware and operating systems) and enable you to focus on the deployment and management of your applications.

- **Software as a service (SaaS)**: Services in this category provide you with a completed product that the service provider runs and manages. In most cases, *software as a service* refers to end-user applications. With a SaaS offering, you do not have to think about how the service is maintained or how the underlying infrastructure is managed. You need to think only about how you plan to use that particular piece of software. A common example of a SaaS application is web-based email, where you can send and receive email without managing feature additions to the email product or maintaining the servers and operating systems that the email program runs on.

Software as a Service (SaaS)

- SaaS provides a cloud-based foundation for ***software applications*** on demand.
- Web-delivered content that users access via a web browser.
- The software can reside within any of the deployment-model clouds.



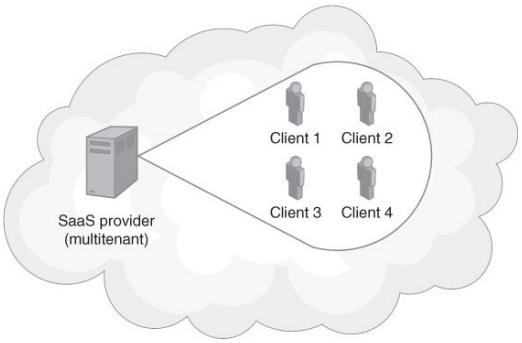
SWINBURNE UNIVERSITY OF TECHNOLOGY

13

Multitenant SaaS Solutions

SaaS applications are often **multitenant solutions**

- Within the single SaaS application, two or more tenant companies share the same server resources.
- Each tenant can customize their own version of the software for their clients.



SWINBURNE UNIVERSITY OF TECHNOLOGY

14

SaaS Advantages

- Eliminate the need for an on-site data center.
- Eliminate the need for application administration.
- Allow customers to pay on demand for software use, normally on a per-user basis.
- Offer application, processor, and data storage scalability.
- Offer device-independent access to applications.
- Increase disaster recovery and business continuity.

SWINBURNE UNIVERSITY OF TECHNOLOGY

15

SaaS Disadvantages

- The biggest concern, or potential disadvantage, is that the data, like the applications, reside in the cloud. Many companies are concerned about letting go of their data.
- Also, because the company does not own the solution, it can be challenging or expensive to customize the application.
- (government regulations)

SWINBURNE UNIVERSITY OF TECHNOLOGY

16

Real World: Salesforce.com

- Cloud-based **customer relationship management (CRM)** solutions.
- Companies accomplish the following:
  - ☐ Manage sales contacts and leads
  - ☐ Centralize contact information, presentations, and project details
  - ☐ Access sales information and reports from anyplace, at any time, with any device
  - ☐ Manage project quotes and project work flow
  - ☐ Sync sales contacts and meetings with existing tools, such as Microsoft Outlook

SWINBURNE UNIVERSITY OF TECHNOLOGY

17

Platform as a Service (PaaS)

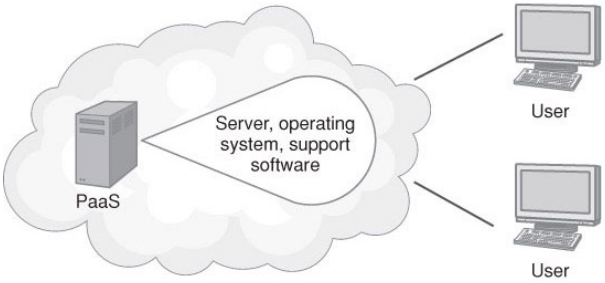
- PaaS provides the underlying hardware technology, such as one or more servers (or virtual servers), operating systems, database solutions, developer tools, and network support, for developers to deploy their own solutions.
- The hardware and software within a PaaS solution is managed by the platform provider.
- Developers need not worry about performing hardware or operating system upgrades. Instead, developers can focus on their own applications.

SWINBURNE UNIVERSITY OF TECHNOLOGY

18

Platform as a Service (PaaS)

- Provide a collection of hardware and software resources that developers can use to build and deploy applications within the cloud.
- Depending on their needs, developers may use a Windows-based PaaS solution or a Linux-based PaaS.



SWINBURNE UNIVERSITY OF TECHNOLOGY

19

PaaS Advantages

- Developers eliminate the need to buy and maintain hardware, and the need to install and manage operating system and database software.
- Because the computing resources no longer reside in the data center, but rather in the cloud, the resources can scale on demand and the company can pay for only resources it consumes.
- Further, because PaaS eliminates the developers' need to worry about servers, they can more quickly deploy their web-based solutions.

SWINBURNE UNIVERSITY OF TECHNOLOGY

20

PaaS Disadvantages

- Some developers and administrators want finer control over the underlying systems (versions, patch releases/applications, …)

SWINBURNE UNIVERSITY OF TECHNOLOGY

Real World: Windows Azure as a PaaS

- Microsoft .NET has driven the development of many dynamic web solutions and web services.
- Windows Azure is a PaaS running within Microsoft data centers.
- Users pay only for the scalable processor resources that they consume.
- SQL Azure provides a cloud-based database solution for applications running within Windows Azure.
- Windows Azure goes beyond .NET and includes support for Java, PHP, and Ruby. Developers can build and deploy their solutions to Azure using an IDE such as Visual Studio or Eclipse.
- Developers can interface to SQL Azure using much of the same code they would use to access a local database.

SWINBURNE UNIVERSITY OF TECHNOLOGY

22

Infrastructure as a Service (IaaS)

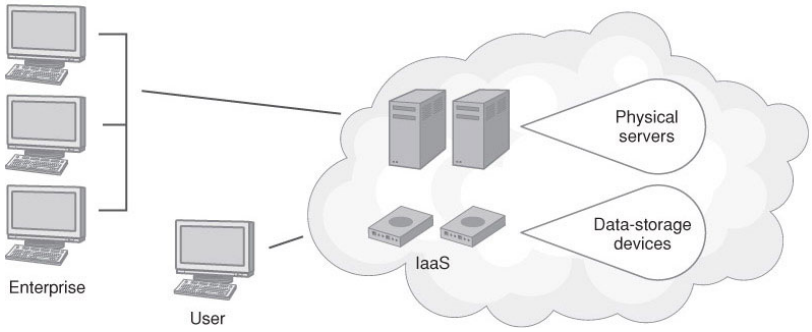
- IaaS provides a virtual data center within the cloud.
- IaaS provides servers (physical and virtualized), cloud-based data storage, and more.
- Developers must install their own operating system, database management software, and support software.
- Then the developers (or the company's system administrators) must manage both the hardware and the software.

SWINBURNE UNIVERSITY OF TECHNOLOGY

23

IaaS Defined

An IaaS provider makes all of the computing hardware resources available, and the customers, in turn, are responsible for installing and managing the systems, which they can normally do, for the most part, over the Internet.



SWINBURNE UNIVERSITY OF TECHNOLOGY

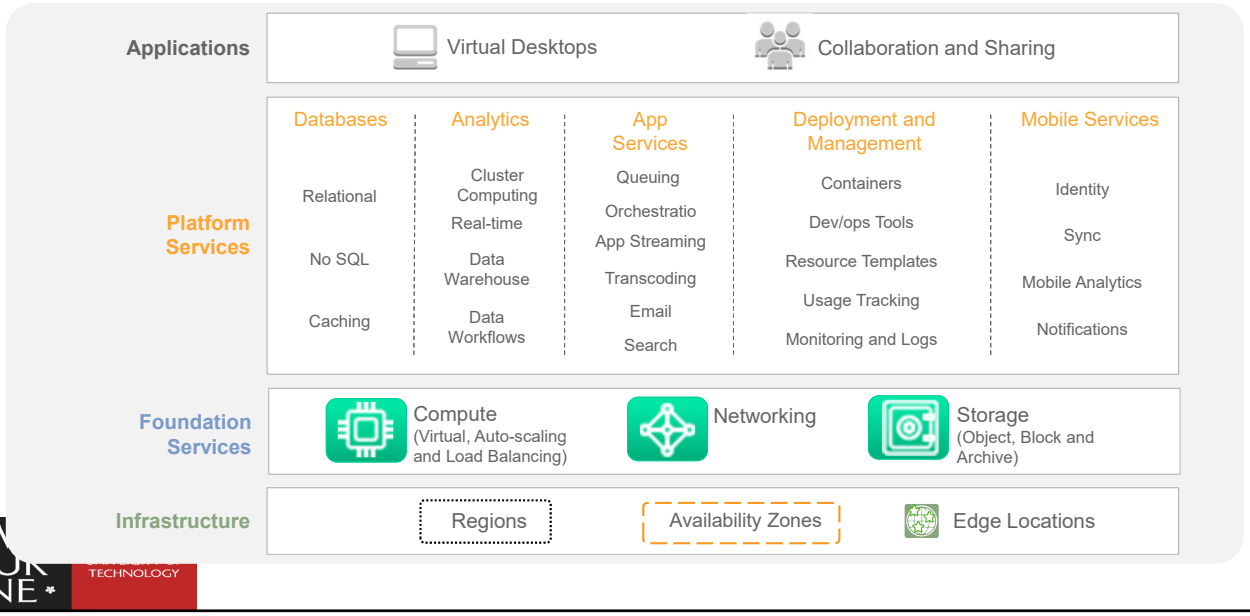
What Data Centers Must Provide

- Access to high-speed and redundant Internet service
- Sufficient air conditioning to eliminate the heat generated by servers and disk storage devices
- Conditioned power with the potential for uninterrupted power supply in the short term and long term through the use of on-site diesel powered generators
- Fire suppression systems
- Administrative staffing to support hardware, networks, and operating systems

SWINBURNE UNIVERSITY OF TECHNOLOGY

25

Combining IaaS, PaaS, SaaS. Example AWS

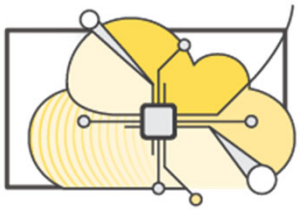


As we discussed earlier, AWS provides a broad set of services—such as compute power, storage options, networking, and databases—delivered as an on-demand utility that is available in seconds, with pay-as-you-go pricing. All of these services sit on AWS global infrastructure.

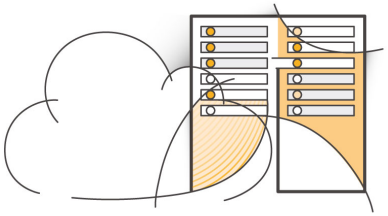
AWS's global infrastructure can be broken down into three elements: Regions, Availability Zones, and edge locations.

Let's take a more in depth look at the AWS infrastructure and see what these are.

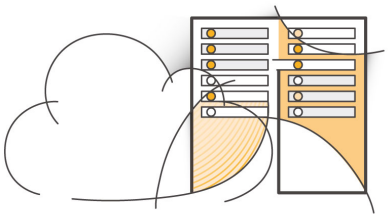
Three Cloud Deployment Models



Public Cloud



Hybrid



Private Cloud (On-premises)

https://aws.amazon.com/enterprise/hybrid/.

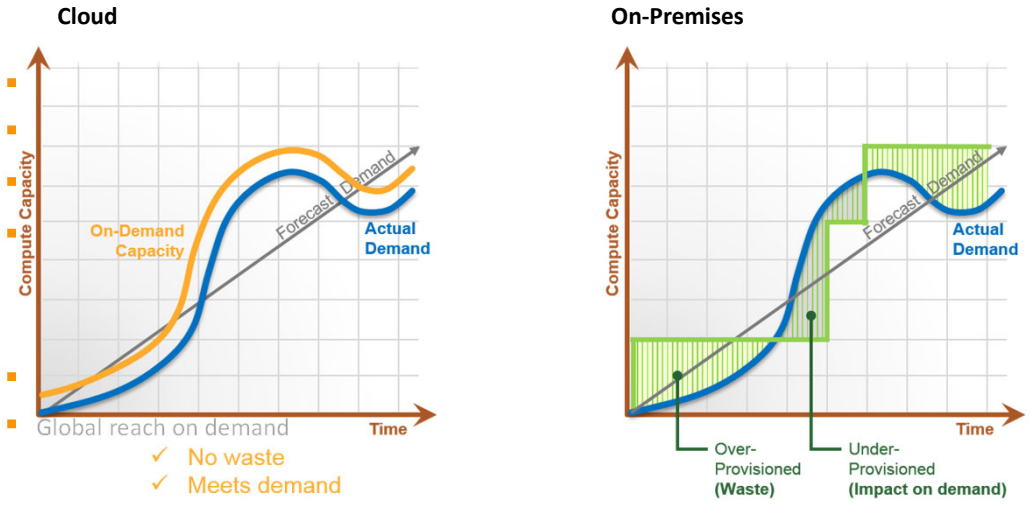
"All-In" Public Cloud:
A cloud-based application is fully deployed in the cloud, and all parts of the application run in the cloud. Applications in the cloud have either been created in the cloud or have been migrated from an existing infrastructure. Cloud-based applications can be built on low-level infrastructure pieces (for example, networking, compute or storage) or can use higher-level services that provide abstraction from the management, architecting, and scaling requirements of core infrastructure.

Hybrid:
A hybrid deployment is a way to connect infrastructure and applications between cloud-based resources and existing resources that are not located in the cloud. The most common method of hybrid deployment is between the cloud and existing on-premises infrastructure (sometimes called *on-prem*). On-premises infrastructure is located within the physical confines of an enterprise, often in the company's data center. A hybrid deployment model is used to extend an organization's infrastructure into the cloud while connecting cloud resources to an internal system. For more information on how AWS can help you with your hybrid deployment, please visit our hybrid page at https://aws.amazon.com/enterprise/hybrid/.

Private Cloud (On-premises):
When you run a cloud infrastructure from your own data center, that's called on-premises or private cloud. While this kind of deployment lacks many of the benefits of cloud computing, it does provide dedicated resources and is a popular choice for

organizations who need to meet certain compliance standards. In most cases, this deployment model is the same as legacy IT infrastructure while using application management and virtualization to increase resource utilization.

Cloud versus On-Premises

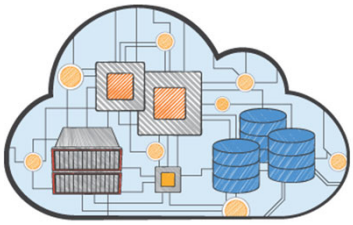


Let's take a closer look at *capacity* in the All-In Cloud and On-Premises solutions.

In the "All In" solution, capacity is in sync with demand. Resources are provisioned and decommissioned in response to demand with only a couple clicks.

In contrast, in "On-Premises" deployments, because you rely on physical hardware, you have to forecast your capacity needs well in advance of the actual demand. Instead of resources that expand and contract with demand, the on-premises solution results in idle, wasted resources waiting for demand to catch up. If demand suddenly outpaces capacity, the shortfall may result in unhappy customers. Your ability to respond quickly to this situation can be limited by long procurement cycles or by constraints on where you house your IT resources. Furthermore, building an on-premises infrastructure can be slow and expensive.

All-In Cloud versus On-Premises



All-In Cloud

- No upfront investment
- Low ongoing costs
- Focus on innovation
- Flexible capacity
- Speed and agility
- Global reach on demand

On-Premises

- Large initial purchase
- Labor, patches, and upgrade cycles
- Systems administration
- Fixed capacity
- Long procurement cycle and setup
- Limited geographic regions

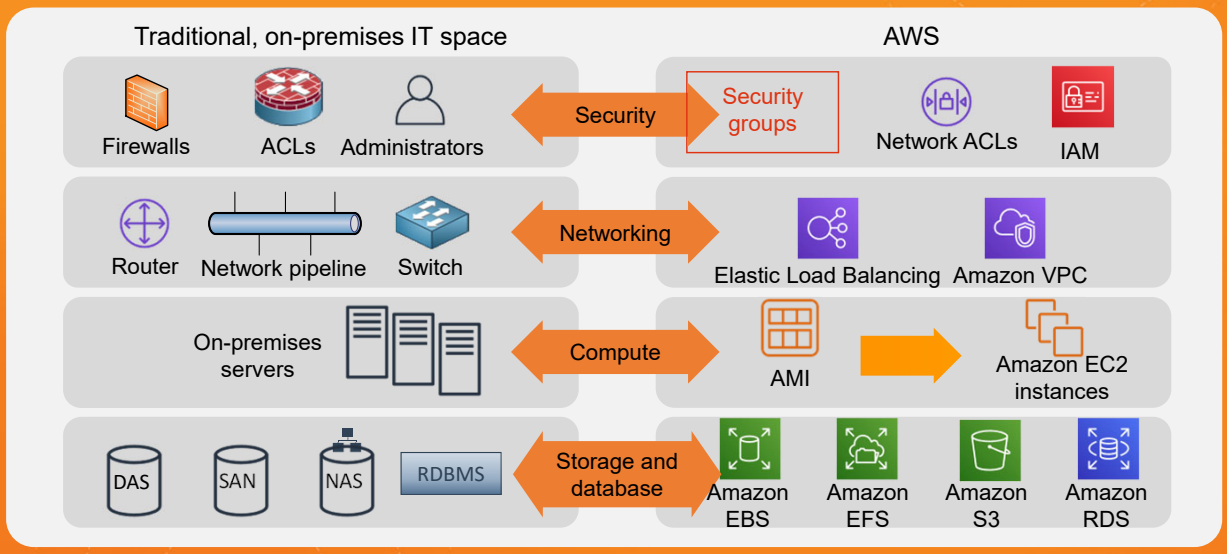
To summarize:

- With the All-In solution, you avoid the large capital purchases required for an On-Premises solution. You have immediate access to resources without having to procure, install, and configure cabling, racks, servers, and storage in a physical location with appropriate facilities like cooling and power. Instead, you just click to order and pay for the resources you need, which are available almost immediately.
- Cloud computing helps you reduce ongoing IT costs in multiple ways. AWS continually lowers prices due to massive economies of scale and continual improvements. Multiple pricing options also help you optimize costs based on your unique workloads. You **pay only for what you use** on a **variable**, monthly basis. On-premises solutions typically require upgrades on 1-year, 3-year, or 5-year cycles.
- Cloud gives you managed IT resources on demand, at a fraction of the cost of traditional infrastructure. This cost savings empowers organizations to shift resources toward **innovative** new projects that grow their business by focusing on “apps not ops.”
- Predicting how customers are going to adopt your new application is complex, making it difficult to estimate your infrastructure capacity needs. **Flexible capacity** means that your resources are dynamic. You can quickly provision resources as demand goes up and turn off what you don’t need as demand declines.
- Cloud computing’s **speed and agility** makes it possible for you to respond to changing market conditions. With AWS, resources can be provisioned as needed. This self-service environment

changes how you develop and deploy applications, allowing your team to experiment more quickly and more frequently. The amount of time it takes to get a server procured, delivered, and running limits this in a traditional infrastructure.

- With on-premises, it is hard to deliver great performance to a distributed user base. So, companies focus on one geographic region at a time to save costs and time. Without geographical limitations, you can deploy your application in any of the AWS regions around the world with lower latency and at minimal cost.

Similarities between AWS and traditional IT



There are many similarities between AWS and the traditional, on-premises IT space:

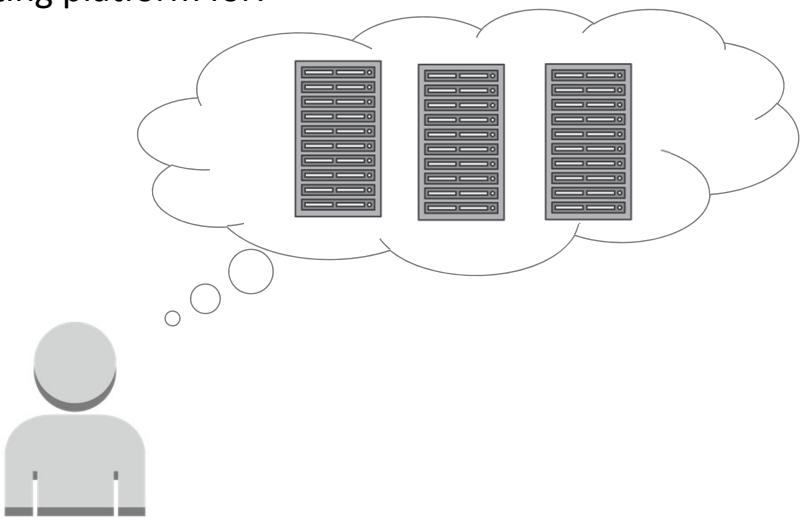
- AWS security groups, network access control lists (network ACLs), and AWS Identity and Access Management (IAM) are similar to firewalls, access control lists (ACLs), and administrators.
- Elastic Load Balancing and Amazon Virtual Private Cloud (Amazon VPC) are similar to routers, network pipelines, and switches.
- Amazon Machine Images (AMIs) and Amazon Elastic Compute Cloud (Amazon EC2) instances are similar to on-premises servers.
- Amazon Elastic Block Store (Amazon EBS), Amazon Elastic File System (Amazon EFS), Amazon Simple Storage Service (Amazon S3), and Amazon Relational Database Service (Amazon RDS) are similar to direct attached storage (DAS), storage area networks (SAN), network attached storage (NAS), and a relational database management service (RDBMS).

With AWS services and features, you can do almost everything that you would want to do with a traditional data center.

What can you do in the cloud?

You can use a cloud computing platform for:

- Application Hosting
- Backup and Storage
- Content Delivery
- Websites
- Enterprise IT
- Databases



You can use the cloud computing platform for the following:

- **Application Hosting** – use on-demand infrastructure to host internal or SaaS applications
- **Backup and Storage** - store data and build dependable backup solutions
- **Content Delivery** - distribute content worldwide, with high data transfer speeds
- **Websites** – host static and dynamic websites
- **Enterprise IT** - host internal- or external-facing IT applications in AWS's secure environment
- **Databases** – use a variety of scalable database solutions, from hosted enterprise database software to non-relational database solutions

Important Cloud Terminology

- **High Availability (Highly Available)**
  - Accessible when you need it
- **Fault Tolerance (Fault Tolerant)**
  - Ability to withstand a certain amount of failure and still remain functional
- **Scalability (Scalable)**
  - Ability to easily grow in size, capacity, and/or scope when required
  - Growth is (usually) based on demand
- **Elasticity (Elastic)**
  - Ability to grow (scale) when required and to reduce in size when resources are no longer needed

High availability, fault tolerance, scalability, and elasticity are four terms that often used when discussing the cloud. These concepts are the fundamental building blocks of AWS and will be referred to throughout the course.

High availability refers to a resource that is accessible when you attempt to access it. For example, if every time you go to the ATM to make a withdrawal it works as expected the ATM is highly available; however, if you go to use it and there is a sign on the front that says “Out of Order”, it is not highly available.

Fault tolerance is the ability to withstand a certain amount of failure and still remain functional. It also refers to the ability of a system to be self-healing and return to full capacity despite a failure. It is the ability of a system to fail in some way but still remain functional.

Scalability is the ability to easily grow in size, capacity, and/or scope when required particularly in response to demand. If something cannot quickly grow in an easy manner it is not scalable.

Elasticity is the ability to not only grow (scale) when required, but also to reduce or contract in size as needed. A system that is elastic can scale to grow as needed usually based on demand and contractnas demand decreases.

Advantages of cloud computing

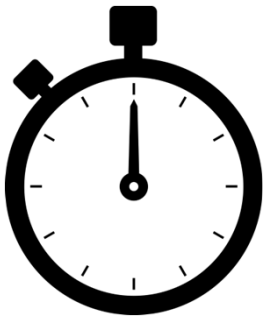
aws academy

Section 2: Advantages of cloud computing

Why are so many companies interested in moving to the cloud? This section presents six advantages of cloud computing.

Trade capital expense for variable expense

Data center investment based on forecast



Pay only for the amount you consume

Advantage #1—Trade capital expense for variable expense: *Capital expenses (capex)* are funds that a company uses to acquire, upgrade, and maintain physical assets such as property, industrial buildings, or equipment. Do you remember the data center example in the traditional computing model where you needed to rack and stack the hardware, and then manage it all? You must pay for everything in the data center whether you use it or not.

By contrast, a *variable expense* is an expense that the person who bears the cost can easily alter or avoid. Instead of investing heavily in data centers and servers before you know how you will use them, you can pay only when you consume resources and pay only for the amount you consume. Thus, you save money on technology. It also enables you to adapt to new applications with as much space as you need in minutes, instead of weeks or days. Maintenance is reduced, so you can spend focus more on the core goals of your business.

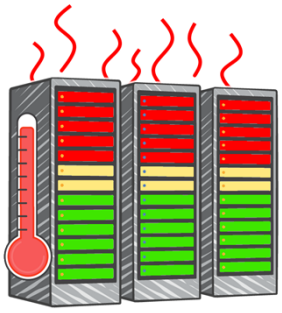
Massive economies of scale

Because of aggregate usage from all customers, AWS can achieve higher economies of scale and pass savings on to customers.

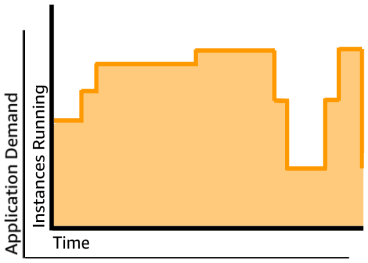
Advantage #2—Benefit from massive economies of scale: By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers is aggregated in the cloud, providers such as AWS can achieve higher economies of scale, which translates into lower pay-as-you-go prices.

Stop guessing capacity

Overestimated server capacity



Underestimated server capacity

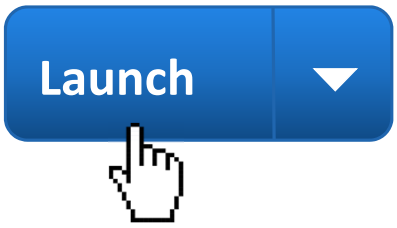


Scaling on demand

Advantage #3—Stop guessing capacity: Eliminate guessing about your infrastructure capacity needs. When you make a capacity decision before you deploy an application, you often either have expensive idle resources or deal with limited capacity. With cloud computing, these problems go away. You can access as much or as little as you need, and scale up and down as required with only a few minutes' notice.

Increase speed and agility

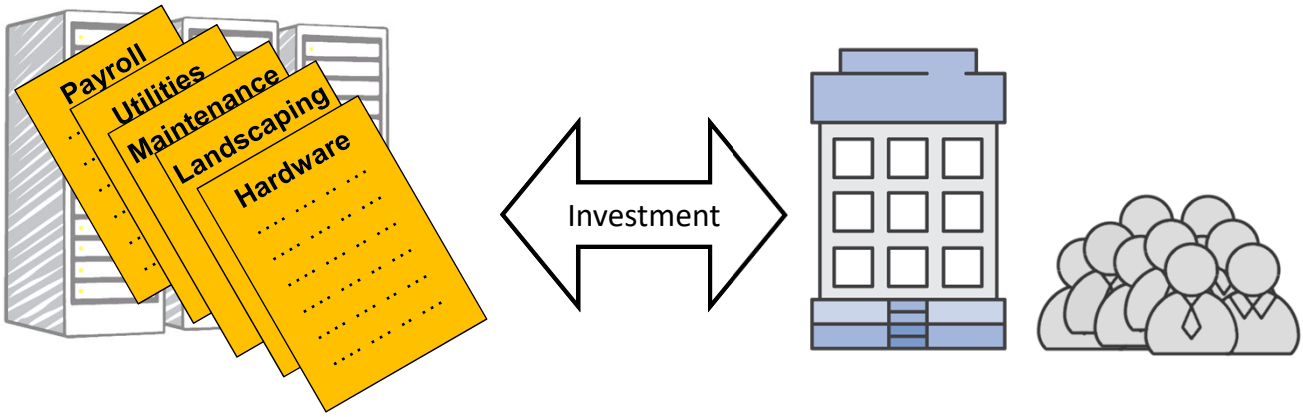
Weeks between wanting resources and having resources



Minutes between wanting resources and having resources

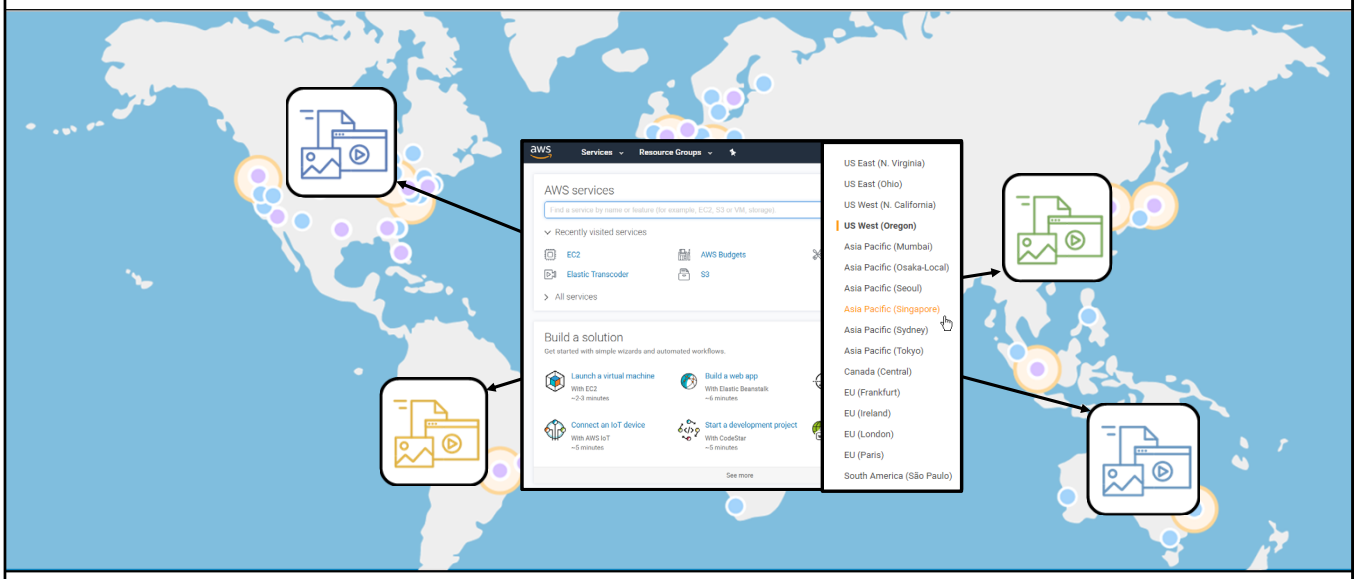
Advantage #4—Increase speed and agility: In a cloud computing environment, new IT resources are only a click away, which means that you reduce the time it takes to make those resources available to your developers from weeks to just minutes. The result is a dramatic increase in agility for the organization because the cost and time that it takes to experiment and develop are significantly lower.

Stop spending money on running and maintaining data centers



Advantage #5—Stop spending money on running and maintaining data centers: Focus on projects that differentiate your business instead of focusing on the infrastructure. Cloud computing enables you to focus on your own customers instead of the heavy lifting of racking, stacking, and powering servers.

Go global in minutes



Advantage #6—Go global in minutes: You can deploy your application in multiple AWS Regions around the world with just a few clicks. As a result, you can provide a lower latency and better experience for your customers simply and at minimal cost.

Key takeaways

- Trade capital expense for variable expense
- Benefit from massive economies of scale
- Stop guessing capacity
- Increase speed and agility
- Stop spending money on running and maintaining data centers
- Go global in minutes

The key takeaways from this section of the module include the six advantages of cloud computing:

- Trade capital expense for variable expense
- Massive economies of scale
- Stop guessing capacity
- Increase speed and agility
- Stop spending money on running and maintaining data centers
- Go global in minutes

Disadvantages?

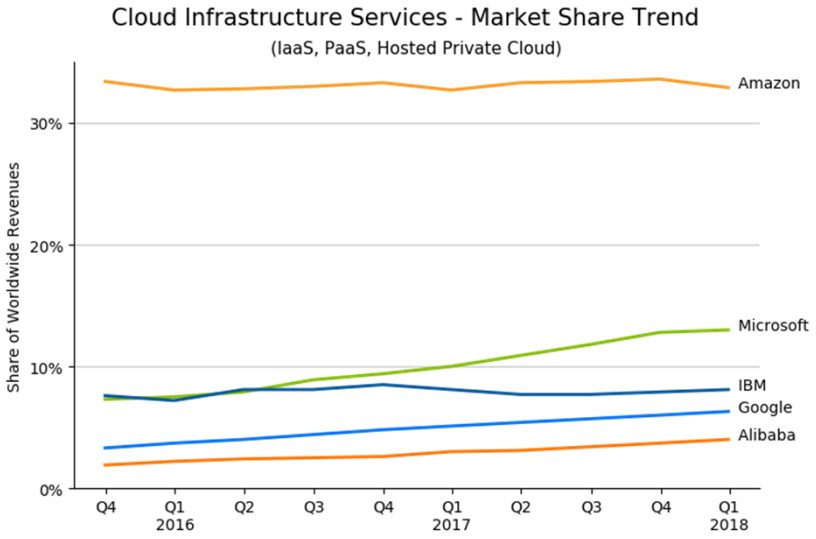
- Cost?
- Security?
- Control?
- Legislative?

Features of Cloud-based Platforms

- **Scalability.** On demand resource scaling.
- **Redundancy.** Servers, storage, and networks.
- **Cost benefits from resource pooling.** Shares IT resources across a very large number of companies, which provides cost savings to each.
- **Outsourced server management.** Provides an IT staff who maintain operating systems and underlying support software.
- **Low cost of entry.** Companies do not need to invest in their own IT data center.

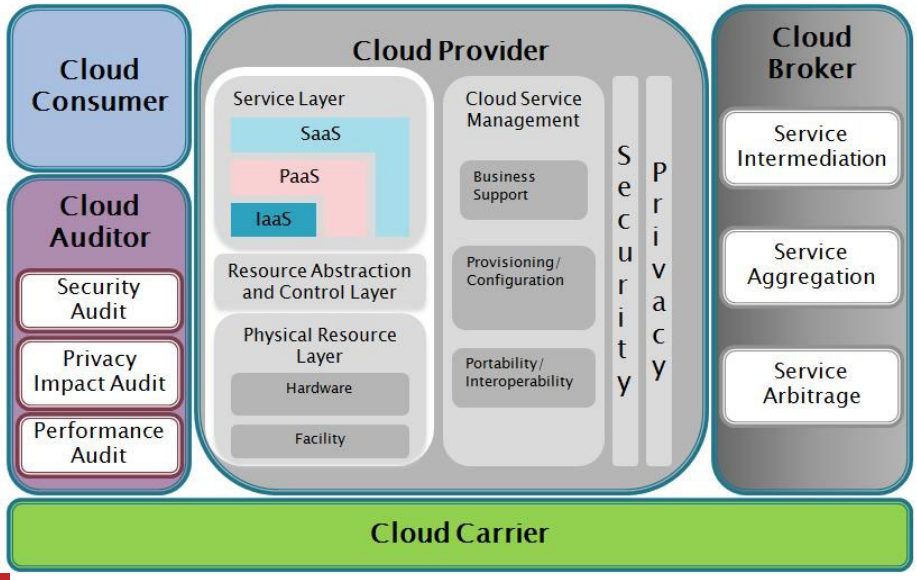
SWINBURNE UNIVERSITY OF TECHNOLOGY

Cloud Providers



https://www.srgresearch.com/

NIST Reference Architecture



SWINBURNE UNIVERSITY OF TECHNOLOGY

44

Microsoft Azure

- https://mva.microsoft.com/en-US/training-courses/cloud-application-development-17172?l=5YgyxecZD_2811316548
- https://mva.microsoft.com/en-US/training-courses/cloud-application-development-17172?l=zQlAmhcZD_8411316548
- https://azure.microsoft.com/en-us/resources/videos/tour-of-microsoft-azure/
- https://www.redpixie.com/blog/microsoft-azure-aws-guide

SWINBURNE UNIVERSITY OF TECHNOLOGY

45

Content Today

- Unit overview and Blackboard
  - Unit outline
  - AWS Cloud Practitioner certification
  - Labs
  - Assessment
- Introducing cloud computing
  - Virtualisation and Cloud Computing
  - Advantages of Cloud Computing
  - Cloud Providers
- AWS Global Infrastructure and Services

SWINBURNE UNIVERSITY OF TECHNOLOGY

46

Welcome to Module 3: AWS Global Infrastructure Overview.

Module 3: AWS Global Infrastructure Overview

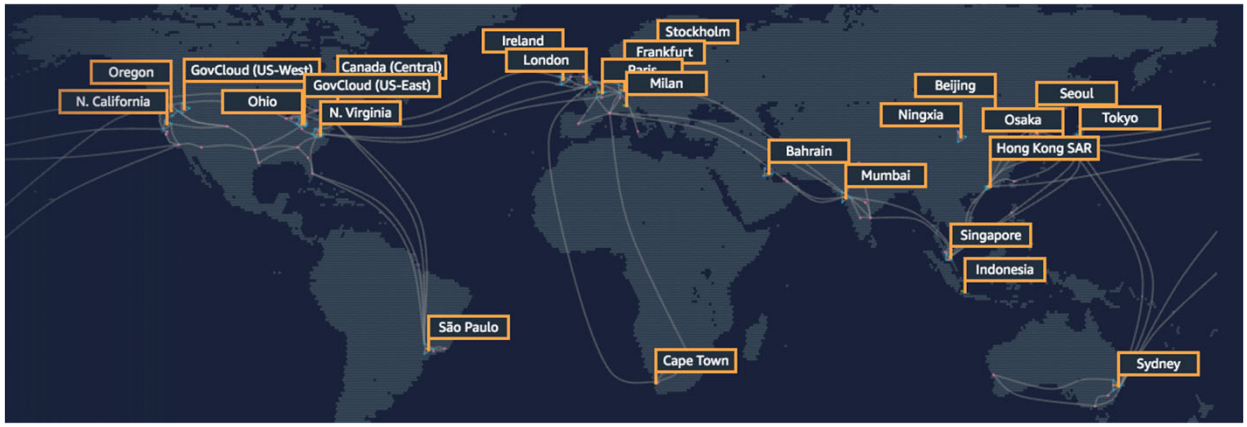
Section 1: AWS Global Infrastructure

aws academy

Introducing Section 1: AWS Global Infrastructure.

AWS Global Infrastructure

- The **AWS Global Infrastructure** is designed and built to deliver a **flexible**, **reliable**, **scalable**, and **secure** cloud computing environment with high-quality **global network performance**.
- This map from https://infrastructure.aws shows the current **AWS Regions** and more that are coming soon.



To learn more about the AWS Regions that are currently available, see: https://aws.amazon.com/about-aws/global-infrastructure/.

The diagram shows the 22 current AWS Regions, as well as a few Regions that will become available soon, including Milan, Cape Town, and Indonesia (as of October 2019).

AWS Regions

- An **AWS Region** is a geographical area.
  - Data replication across Regions is controlled by you.
  - Communication between Regions uses AWS backbone network infrastructure.
- Each Region provides full redundancy and connectivity to the network.
- A Region typically consists of two or more **Availability Zones**.

Example: London Region

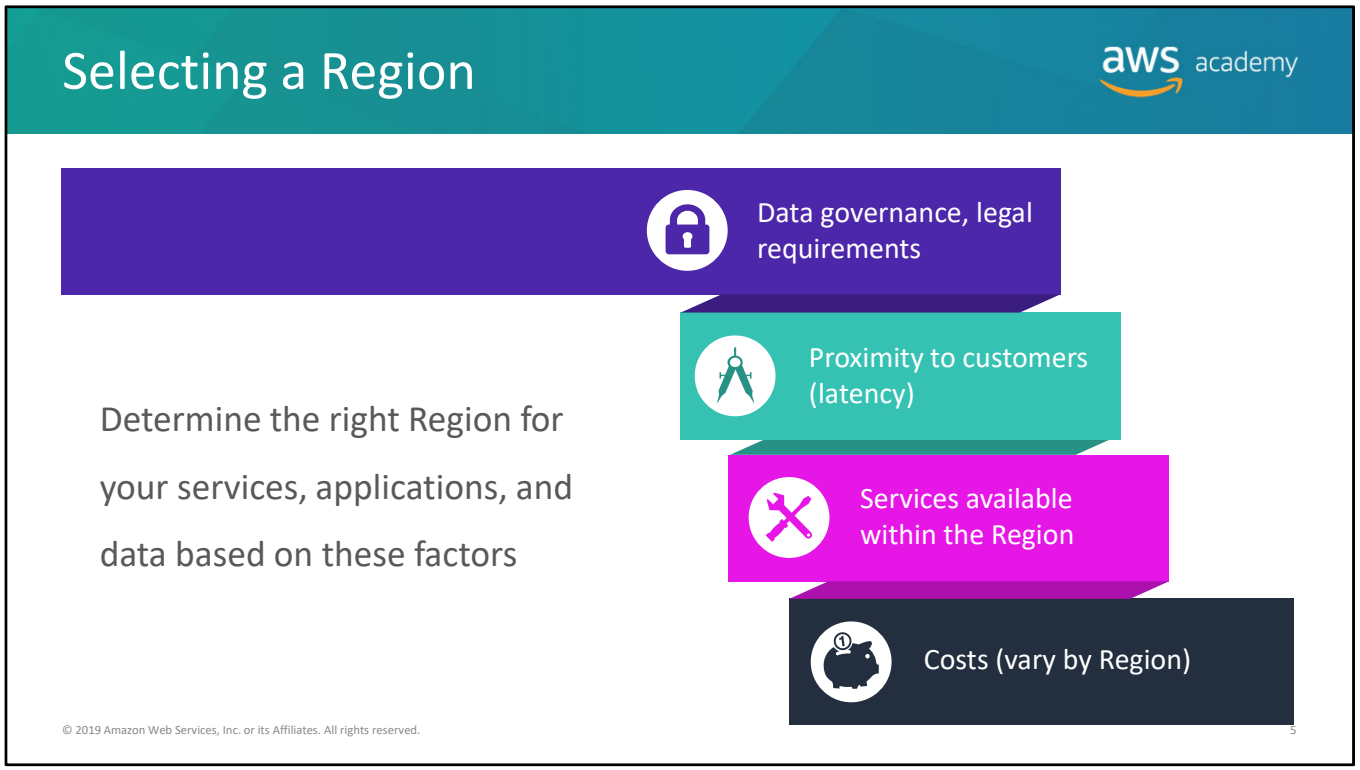
The AWS Cloud infrastructure is built around Regions. AWS has 22 Regions worldwide. An **AWS Region** is a physical geographical location with one or more **Availability Zones**. Availability Zones in turn consist of one or more **data centers**.

To achieve fault tolerance and stability, Regions are isolated from one another. Resources in one Region are not automatically replicated to other Regions. When you store data in a specific Region, it is not replicated outside that Region.

It is your responsibility to replicate data across Regions, if your business needs require it.

AWS Regions that were introduced before March 20, 2019 are *enabled* by default. Regions that were introduced after March 20, 2019—such as Asia Pacific (Hong Kong) and Middle East (Bahrain)—are *disabled* by default. You must enable these Regions before you can use them. You can use the AWS Management Console to enable or disable a Region.

Some Regions have restricted access. An Amazon AWS (**China**) account provides access to the Beijing and Ningxia Regions only. To learn more about AWS in China, see: https://www.amazonaws.cn/en/about-aws/china/. The isolated **AWS GovCloud (US)** Region is designed to allow US government agencies and customers to move sensitive workloads into the cloud by addressing their specific regulatory and compliance requirements.



There are a few factors that you should consider when you select the optimal Region or Regions where you store data and use AWS services.

One essential consideration is **data governance and legal requirements**. Local laws might require that certain information be kept within geographical boundaries. Such laws might restrict the Regions where you can offer content or services. For example, consider the European Union (EU) Data Protection Directive.

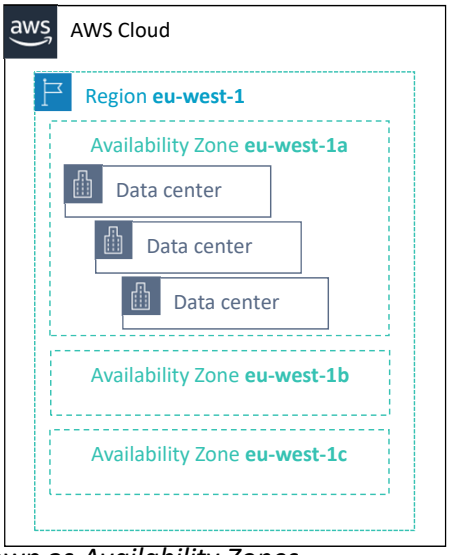
All else being equal, it is generally desirable to run your applications and store your data in a Region that is as close as possible to the user and systems that will access them. This will help you **reduce latency**. CloudPing is one website that you can use to test latency between your location and all AWS Regions. To learn more about CloudPing, see: http://www.cloudping.info/

Keep in mind that not all services are available in all Regions. To learn more, see: https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/?p=tgi&loc=4.

Finally, there is some variation in the **cost** of running services, which can depend on which Region you choose. For example, as of this writing, running an On-Demand t3.medium size Amazon Elastic Compute Cloud (Amazon EC2) Linux instance in the US East (Ohio) Region costs $0.0416 per hour, but running the same instance in the Asia Pacific (Tokyo) Region costs $0.0544 per hour.

Availability Zones

- Each **Region** has multiple Availability Zones.
- Each **Availability Zone** is a fully isolated partition of the AWS infrastructure.
  - There are currently 69 Availability Zones worldwide
  - Availability Zones consist of discrete **data centers**
  - They are designed for fault isolation
  - They are interconnected with other Availability Zones by using high-speed private networking
  - You choose your Availability Zones.
  - **AWS recommends replicating data and resources across Availability Zones** for resiliency.



Each AWS Region has multiple, isolated locations that are known as *Availability Zones*.

Each Availability Zone provides the ability to operate applications and databases that are more highly available, fault-tolerant, and scalable than would be possible with a single data center. Each Availability Zone can include multiple data centers (typically three), and at full-scale, they can include hundreds of thousands of servers. They are fully isolated partitions of the AWS Global Infrastructure. Availability Zones have their own power infrastructure, and they are physically separated by many kilometers from other Availability Zones—though all Availability Zones are within 100 km of each other.

All Availability Zones are interconnected with high-bandwidth, low-latency networking over fully redundant, dedicated fiber that provides high-throughput between Availability Zones. The network accomplishes synchronous replication between Availability Zones.

Availability Zones help build highly available applications. When an application is partitioned across Availability Zones, companies are better isolated and protected from issues such as lightning, tornadoes, earthquakes, and more.

You are responsible for selecting the Availability Zones where your systems will reside. Systems can span multiple Availability Zones. AWS recommends replicating across Availability Zones for resiliency. You should design your systems to survive the temporary or prolonged failure of an Availability Zone if a disaster occurs.

AWS data centers

- AWS data centers are designed for security.
- Data centers are where the data resides and data processing occurs.
- Each data center has redundant power, networking, and connectivity, and is housed in a separate facility.
- A data center typically has 50,000 to 80,000 physical servers.

The foundation for the AWS infrastructure is the data centers. Customers do not specify a data center for the deployment of resources. Instead, an Availability Zone is the most granular level of specification that a customer can make. However, a data center is the location where the actual data resides. Amazon operates state-of-the-art, highly available data centers. Although rare, failures can occur that affect the availability of instances in the same location. If you host all your instances in a single location that is affected by such a failure, none of your instances will be available.

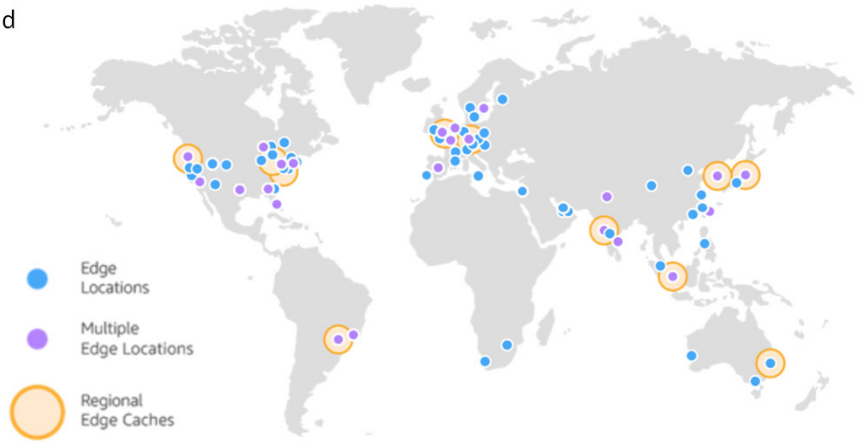
Data centers are securely designed with several factors in mind:

- Each location is carefully evaluated to **mitigate environmental risk**.
- Data centers have a **redundant design** that anticipates and tolerates failure while maintaining service levels.
- To ensure availability, **critical system components are backed up** across multiple Availability Zones.
- To ensure capacity, AWS continuously monitors service usage to deploy infrastructure to support availability commitments and requirements.
- Data center **locations are not disclosed** and all access to them is restricted.
- In case of failure, automated processes move data traffic away from the affected area.

AWS uses **custom network equipment** sourced from **multiple original device manufacturers (ODMs)**. ODMs design and manufacture products based on specifications from a second company. The second company then rebrands the products for sale.

Points of Presence

- AWS provides a global network of 187 Points of Presence locations
- Consists of 176 **edge locations** and 11 **Regional edge caches**
- Used with Amazon CloudFront
  - A global Content Delivery Network (CDN), that delivers content to end users with reduced latency
- Regional edge caches used for content with infrequent access.



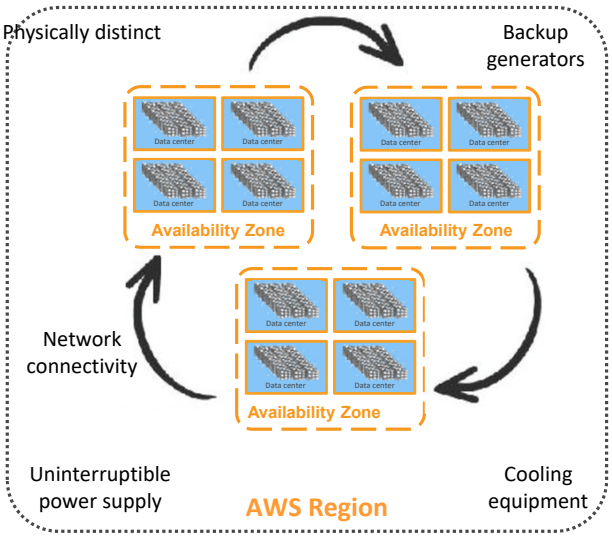
Amazon CloudFront is **a content delivery network** (CDN) used to distribute content to end users to reduce latency**. Amazon Route 53** is a Domain Name System (DNS) service. Requests going to either one of these services will be routed to the nearest **edge location** automatically in order to lower latency.

AWS **Points of Presence** are located in most of the major cities (69 cities in total) across 30 countries around the world. By **continuously measuring internet connectivity, performance and computing to find the best way to route requests**, the Points of Presence deliver a better near real-time user experience. They are used by many AWS services, including Amazon CloudFront, Amazon Route 53, AWS Shield, and AWS Web Application Firewall (AWS WAF) services.

Regional edge caches are used by default with Amazon CloudFront. Regional edge caches are used when you have content that is not accessed frequently enough to remain in an **edge location**. Regional edge caches absorb this content and provide an alternative to that content having to be fetched from the origin server.

AWS infrastructure features

- Elasticity and scalability
  - Elastic infrastructure; dynamic adaption of capacity
  - Scalable infrastructure; adapts to accommodate growth
- Fault-tolerance
  - Continues operating properly in the presence of a failure
  - Built-in redundancy of components
- High availability
  - High level of operational performance
  - Minimized downtime
  - No human intervention



Now that you have a good understanding of the major components that comprise the AWS Global Infrastructure, let's consider the benefits provided by this infrastructure.

The AWS Global Infrastructure has several valuable features:

- First, it is **elastic** and **scalable**. This means resources can dynamically adjust to increases or decreases in capacity requirements. It can also rapidly adjust to accommodate growth.
- Second, this infrastructure is **fault tolerant**, which means it has built-in component redundancy which enables it to continue operations despite a failed component.
- Finally, it requires minimal to no human intervention, while providing **high availability** with minimal down time.

Key takeaways

- The **AWS Global Infrastructure** consists of Regions and Availability Zones.
- Your choice of a **Region** is typically based on compliance requirements or to reduce latency.
- Each **Availability Zone** is physically separate from other Availability Zones and has redundant power, networking, and connectivity.
- **Edge locations**, and **Regional edge caches** improve performance by caching content closer to users.

Some key takeaways from this section of the module include:

- The AWS Global Infrastructure consists of Regions and Availability Zones.
- Your choice of a Region is typically based on compliance requirements or to reduce latency.
- Each Availability Zone is physically separate from other Availability Zones and has redundant power, networking, and connectivity.
- Edge locations, and Regional edge caches improve performance by caching content closer to users.

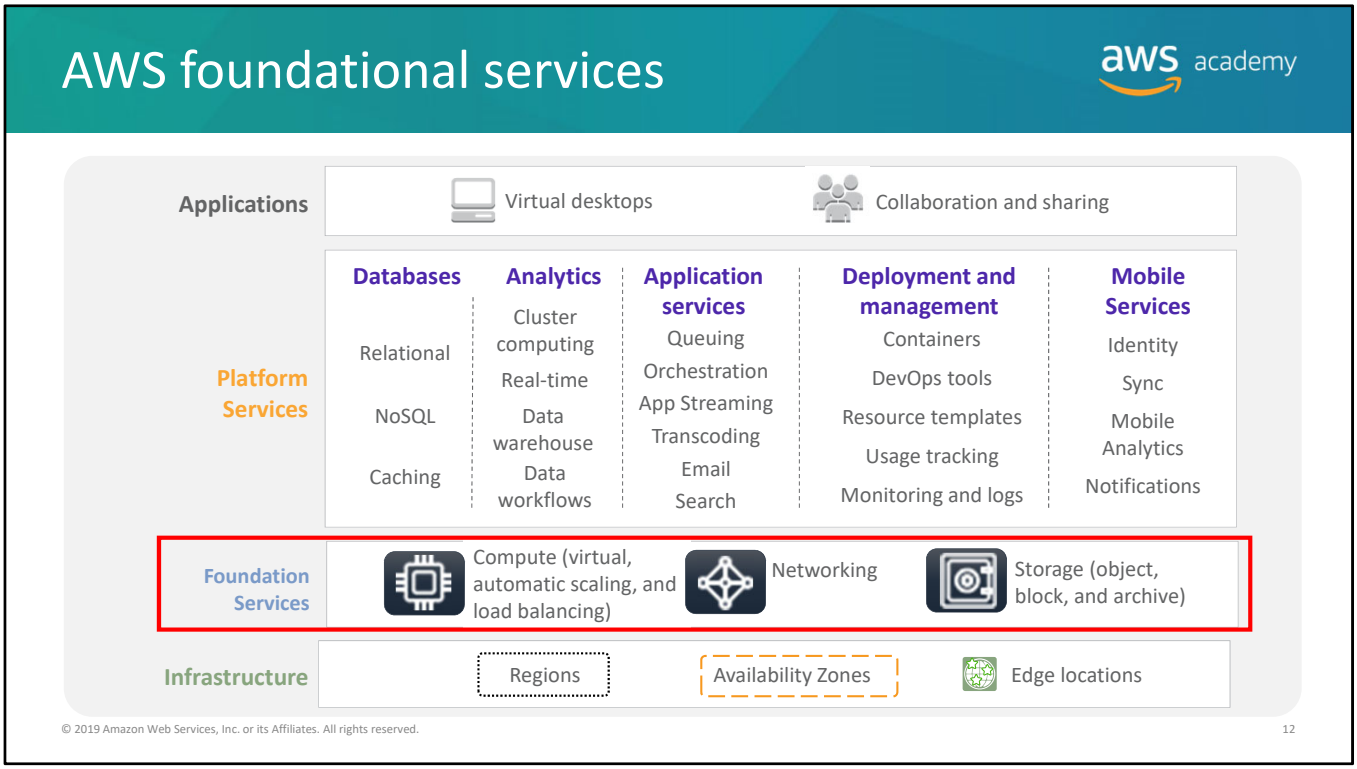
Module 3: AWS Global Infrastructure Overview

Section 2: AWS services and service category overview

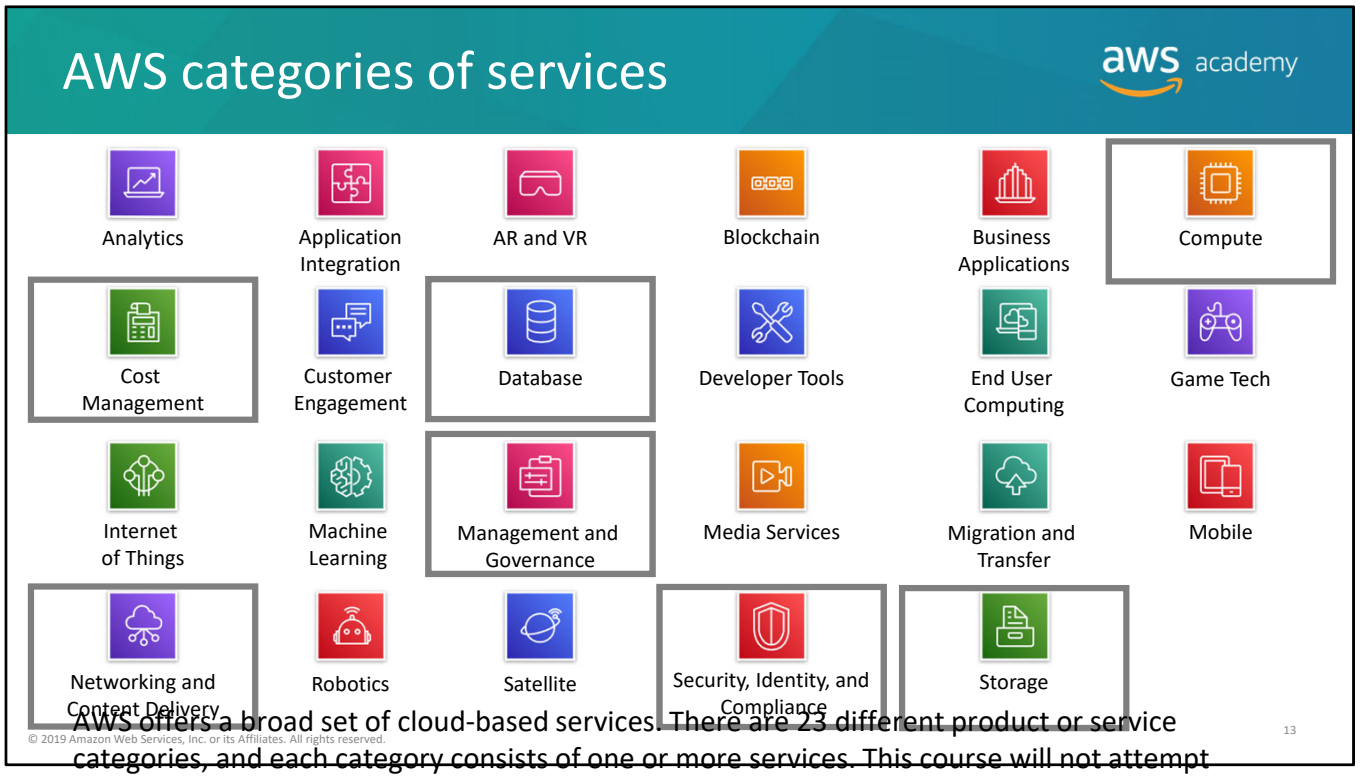
aws academy

Introducing Part 2: AWS Service and Service Category Overview.

AWS offers a broad set of global cloud-based products that can be used as building blocks for common cloud architectures. Here is a look at how these cloud based products are organized.



As discussed previously, the AWS Global Infrastructure can be broken down into three elements: Regions, Availability Zones, and Points of Presence, which include edge locations. This infrastructure provides the platform for a broad set of services, such as networking, storage, compute services, and databases—and these services are delivered as an on-demand utility that is available in seconds, with pay-as-you-go pricing.



AWS offers a broad set of cloud-based services. There are 23 different product or service categories, and each category consists of one or more services. This course will not attempt to introduce you to each service. Rather, the focus of this course is on the services that are most widely used and offer the best introduction to the AWS Cloud. This course also focuses on services that are more likely to be covered in the AWS Certified Cloud Practitioner exam.

The categories that this course will discuss are highlighted on the slide: Compute, Cost Management, Database, Management and Governance, Networking and Content Delivery, Security, Identity, and Compliance, and Storage.

To learn more about AWS products, see: http://aws.amazon.com/products. All AWS products are organized into the service categories that are shown here. For example, if you click **Compute**, you will see that Amazon Elastic Compute Cloud (Amazon EC2) is first on the list. The compute category also lists many other products and services.

If you click **Amazon EC2**, it takes you to the Amazon EC2 page. Each product page provides a detailed description of the product and lists some of its benefits.

Explore the different service groups to understand the categories and services within them. Now that you know how to locate information about different services, this module will discuss the highlighted service categories. **The next seven slides list the individual services —within each of the categories highlighted above—that this course will discuss.**

Storage service category

Photo from https://www.pexels.com/photo/black-and-grey-device-159282/

AWS storage services

Amazon Simple Storage Service (Amazon S3)

Amazon Elastic Block Store (Amazon EBS)

Amazon Elastic File System (Amazon EFS)

Amazon Simple Storage Service Glacier

AWS storage services include the services listed here, and many others.

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers scalability, data availability, security, and performance. Use it to store and protect any amount of data for websites, mobile apps, backup and restore, archive, enterprise applications, Internet of Things (IoT) devices, and big data analytics.

Amazon Elastic Block Store (Amazon EBS) is high-performance block storage that is designed for use with Amazon EC2 for both throughput and transaction intensive workloads. It is used for a broad range of workloads, such as relational and non-relational databases, enterprise applications, containerized applications, big data analytics engines, file systems, and media workflows.

Amazon Elastic File System (Amazon EFS) provides a scalable, fully managed elastic Network File System (NFS) file system for use with AWS Cloud services and on-premises resources. It is built to scale on demand to petabytes, growing and shrinking automatically as you add and remove files. It reduces the need to provision and manage capacity to accommodate growth.

Amazon Simple Storage Service Glacier is a secure, durable, and extremely low-cost Amazon S3 cloud storage class for data archiving and long-term backup. It is designed to deliver 11 9s of durability, and to provide comprehensive security and compliance capabilities to meet stringent regulatory requirements.

Compute service category

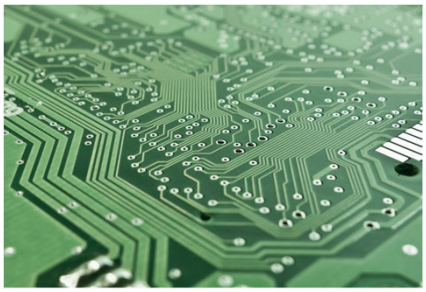


Photo from https://www.pexels.com/photo/technology-computer-lines-board-50711/

AWS Compute services

- Amazon EC2
- Amazon EC2 Auto Scaling
- Amazon Elastic Container Service (Amazon ECS)
- Amazon EC2 Container Registry
- AWS Elastic Beanstalk
- AWS Lambda
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Fargate

AWS compute services include the services listed here, and many others.

Amazon Elastic Compute Cloud (Amazon EC2) provides resizable compute capacity as virtual machines in the cloud.

Amazon EC2 Auto Scaling enables you to automatically add or remove EC2 instances according to conditions that you define.

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container orchestration service that supports Docker containers.

Amazon Elastic Container Registry (Amazon ECR) is a fully-managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images.

AWS Elastic Beanstalk is a service for deploying and scaling web applications and services on familiar servers such as Apache and Microsoft Internet Information Services (IIS).

AWS Lambda enables you to run code without provisioning or managing servers. You pay only for the compute time that you consume. There is no charge when your code is not running.

Amazon Elastic Kubernetes Service (Amazon EKS) makes it easy to deploy, manage, and scale containerized applications that use Kubernetes on AWS.

AWS Fargate is a compute engine for Amazon ECS that allows you to run containers without having to manage servers or clusters.

Database service category

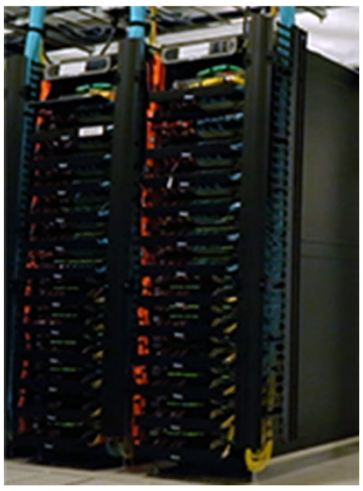


Photo from https://aws.amazon.com/compliance/data-center/data-centers/

AWS Database services

Amazon Relational Database Service

Amazon Aurora

Amazon Redshift

Amazon DynamoDB

AWS database services include the services listed here, and many others.

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides resizable capacity while automating time-consuming administration tasks such as hardware provisioning, database setup, patching, and backups.

Amazon Aurora is a MySQL and PostgreSQL-compatible relational database. It is up to five times faster than standard MySQL databases and three times faster than standard PostgreSQL databases.

Amazon Redshift enables you to run analytic queries against petabytes of data that is stored locally in Amazon Redshift, and directly against exabytes of data that are stored in Amazon S3. It delivers fast performance at any scale.

Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale, with built-in security, backup and restore, and in-memory caching.

Networking and content delivery service category

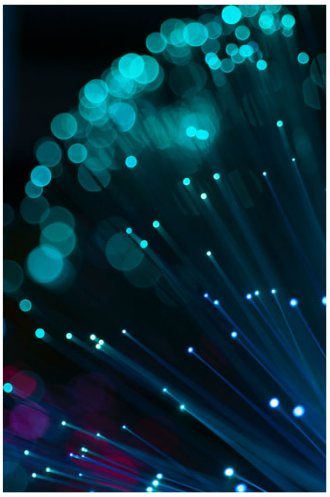


Photo by Umberto on Unsplash

AWS networking and content delivery services

- Amazon VPC
- Elastic Load Balancing
- Amazon CloudFront
- AWS Transit Gateway
- Amazon Route 53
- AWS Direct Connect
- AWS VPN

AWS networking and content delivery services include the services listed here, and many others.

Amazon Virtual Private Cloud (Amazon VPC) enables you to provision logically isolated sections of the AWS Cloud.

Elastic Load Balancing automatically distributes incoming application traffic across multiple targets, such as Amazon EC2 instances, containers, IP addresses, and Lambda functions.

Amazon CloudFront is a fast content delivery network (CDN) service that securely delivers data, videos, applications, and application programming interfaces (APIs) to customers globally, with low latency and high transfer speeds.

AWS Transit Gateway is a service that enables customers to connect their Amazon Virtual Private Clouds (VPCs) and their on-premises networks to a single gateway.

Amazon Route 53 is a scalable cloud Domain Name System (DNS) web service designed to give you a reliable way to route end users to internet applications. It translates names (like *www.example.com*) into the numeric IP addresses (like *192.0.2.1*) that computers use to connect to each other.

AWS Direct Connect provides a way to establish a dedicated private network connection from your data center or office to AWS, which can reduce network costs and increase bandwidth throughput.

AWS VPN provides a secure private tunnel from your network or device to the AWS global network.

Security, identity, and compliance service category

Photo by Paweł Czerwiński on Unsplash

AWS security, identity, and compliance services

- AWS Identity and Access Management (IAM)
- AWS Organizations
- Amazon Cognito
- AWS Artifact
- AWS Key Management Service
- AWS Shield

AWS security, identity, and compliance services include the services listed here, and many others.

AWS Identity and Access Management (IAM) enables you to manage access to AWS services and resources securely. By using IAM, you can create and manage AWS users and groups. You can use IAM permissions to allow and deny user and group access to AWS resources.

AWS Organizations allows you to restrict what services and actions are allowed in your accounts.

Amazon Cognito lets you add user sign-up, sign-in, and access control to your web and mobile apps.

AWS Artifact provides on-demand access to AWS security and compliance reports and select online agreements.

AWS Key Management Service (AWS KMS) enables you to create and manage keys. You can use AWS KMS to control the use of encryption across a wide range of AWS services and in your applications.

AWS Shield is a managed Distributed Denial of Service (DDoS) protection service that safeguards applications running on AWS.

AWS cost management service category

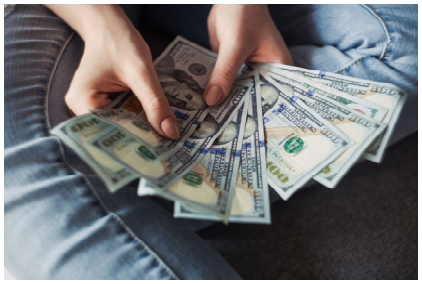


Photo by Alexander Mils on Unsplash

AWS cost management services

- AWS Cost and Usage Report
- AWS Budgets
- AWS Cost Explorer

AWS cost management services include the services listed here, and others.

The AWS Cost and Usage Report contains the most comprehensive set of AWS cost and usage data available, including additional metadata about AWS services, pricing, and reservations.

AWS Budgets enables you to set custom budgets that alert you when your costs or usage exceed (or are forecasted to exceed) your budgeted amount.

AWS Cost Explorer has an easy-to-use interface that enables you to visualize, understand, and manage your AWS costs and usage over time.

Management and governance service category

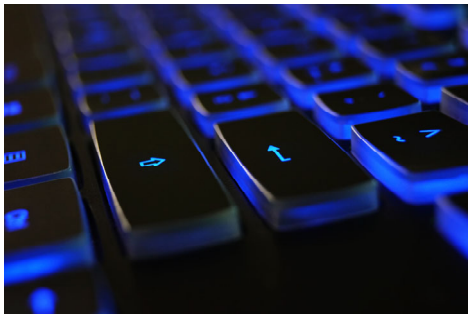


Photo by Marta Branco from Pexels

AWS management and governance services

- AWS Management Console
- AWS Config
- Amazon CloudWatch
- AWS Auto Scaling
- AWS Command Line Interface
- AWS Trusted Advisor
- AWS Well-Architected Tool
- AWS CloudTrail

AWS management and governance services include the services listed here, and others.

The AWS Management Console provides a web-based user interface for accessing your AWS account.

AWS Config provides a service that helps you track resource inventory and changes.

Amazon CloudWatch allows you to monitor resources and applications.

AWS Auto Scaling provides features that allow you to scale multiple resources to meet demand.

AWS Command Line Interface provides a unified tool to manage AWS services.

AWS Trusted Advisor helps you optimize performance and security.

AWS Well-Architected Tool provides help in reviewing and improving your workloads.

AWS CloudTrail tracks user activity and API usage.

Activity: AWS Management Console clickthrough

Photo by Pixabay from Pexels.

In this educator-led activity, you will be asked to log in to the AWS Management Console. The activity instructions are on the next slide. You will be challenged to answer five questions. The educator will lead the class in a discussion of each question, and reveal the correct answers.

Hands-on activity: AWS Management Console clickthrough

1. Launch the Sandbox hands-on environment and connect to the AWS Management Console**.**
2. Explore the AWS Management Console.
   A. Click the **Services** menu.
   B. Notice how services are grouped into service categories. For example, the **EC2** service appears in the **Compute** service category.
      Question #1: Under which service category does the **IAM** service appear?
      Question #2: Under which service category does the **Amazon VPC** service appear?
   C. Click the **Amazon VPC** service. Notice that the dropdown menu in the top-right corner displays an AWS Region (for example, it might display *N. Virginia*).
   D. Click the Region menu and switch to a different Region. For example, choose **EU (London)**.
   E. Click **Subnets** (on the left side of the screen). The Region has three subnets in it. Click the box next to one of the subnets. Notice that the bottom half of the screen now displays details about this subnet.
      Question #3: Does the subnet you selected exist at the level of the Region or at the level of the Availability Zone?
   F. Click **Your VPCs**. An existing VPC is already selected.
      Question #4: Does the VPC exist at the level of the Region or the level of the Availability Zone?
      Question #5: Which services are global instead of Regional? Check Amazon EC2, IAM, Lambda, and Route 53.

The purpose of this activity is to expose you to the AWS Management Console. You will gain experience navigating between AWS service consoles (such as the Amazon VPC console). You will also practice navigating to services in different service categories. Finally, the console will help you distinguish whether a given service or service resource is global or Regional.

Follow the instructions on the slide. After most or all students have completed the steps document above, the educator will review the questions and answers with the whole class.

Activity answer key

- Question #1: Under which service category does the **IAM** service appear?
  - **Answer**: **Security, Identity, & Compliance**.
- Question #2: Under which service category does the **Amazon VPC** service appear?
  - **Answer**: **Networking & Content Delivery**
- Question #3: Does the subnet that you selected exist at the level of the Region or the level of the Availability Zone?
  - **Answer**: Subnets exist at the **level of the Availability Zone.**
- Question #4: Does the VPC exist at the level of the Region or the level of the Availability Zone?
  - **Answer**: VPCs exist at the **Region level.**
- Question #5: Which of the following services are global instead of Regional? Check Amazon EC2, IAM, Lambda, and Route 53.
  - **Answer: IAM and Route 53 are global**. Amazon EC2 and Lambda are Regional.

This slide provides an answer key to the questions that were asked in the activity on the previous slide. The educator will use this slide to lead a discussion and debrief the hands-on activity.

Additional resources

- AWS Global Infrastructure
- AWS Global Infrastructure Region Table
- AWS Cloud Products

The following resources provide more detail on the topics discussed in this module:

- AWS Global Infrastructure
- AWS Global Infrastructure Region Table
- AWS Cloud Products