**1. Tổng quan về Cloud Computing**

Cloud Computing, là thuật ngữ mô tả việc cung cấp phân phối các tài nguyên máy tính (như máy chủ, lưu trữ, cơ sở dữ liệu, mạng, phần mềm) qua internet, giúp người dùng truy cập và sử dụng theo nhu cầu mà không cần quản lý hạ tầng vật lý.

**2. Các mô hình dịch vụ trong Cloud Computing**

* **Infrastructure as a Service (IaaS):** Cung cấp tài nguyên cơ sở hạ tầng như máy chủ, lưu trữ và mạng dưới dạng dịch vụ. Người dùng có quyền kiểm soát hệ điều hành và ứng dụng, nhưng không quản lý hạ tầng vật lý. Amazon EC2
* **Platform as a Service (PaaS):** Cung cấp nền tảng mà nhà phát triển cần để xây dựng, chạy và quản lý ứng dụng, từ máy chủ và hệ điều hành đến tất cả các mạng, lưu trữ, phần mềm trung gian, công cụ, v.v. Google App Engine
* **Software as a Service (SaaS):** Cung cấp ứng dụng phần mềm qua Internet, giúp người dùng truy cập các ứng dụng mà không cần cài đặt hoặc quản lý chúng. Microsoft 365

**3. Các mô hình triển khai Cloud Computing**

* **Public Cloud:** Dịch vụ đám mây được cung cấp qua Internet công cộng và chia sẻ giữa nhiều tổ chức.
* **Private Cloud:** Dịch vụ đám mây được sử dụng độc quyền bởi một tổ chức, cung cấp mức độ kiểm soát và bảo mật cao hơn.
* **Hybrid Cloud:** Kết hợp giữa public và private cloud, cho phép dữ liệu và ứng dụng được chia sẻ giữa chúng.

**4. Giới thiệu về AWS**

Amazon Web Services (AWS) là một nền tảng dịch vụ đám mây toàn diện và được chấp nhận rộng rãi, cung cấp hơn 200 dịch vụ đầy đủ tính năng từ các trung tâm dữ liệu trên toàn cầu.

[aws.amazon.com](https://aws.amazon.com/vi/what-is-cloud-computing/?utm_source=chatgpt.com)

**6. Cơ sở hạ tầng toàn cầu của AWS**

**AWS**

- region: chỉ các khu vực địa lý tiêng biệt thường là quy mô quốc gia hoặc vùng lãnh thổ rộng lớn

Trong Mỗi region chứa các availability zones được kết nối với nhau với độ trễ thấp và chúng thường nằm ở vị trí đủ xa về vị trí địa lý để đảm bảo sự cố xảy ra ở ez này không ảnh hưởng đến các ez khác

Bên trong az thì là các data center

Edge location là data center được dùng để lưu trữ giữ liệu như cache nhằm giảm độ trể trong việc truyển tải dự liệu tới người dùng

Edge location được phân bổ tại các thành phố lớn gần người dùng

Ví dụ ta có người dùng việt nam truy cập youtube để xem 1 video mà youtube đặt data center ở mỹ . khi video được gửi về người dùng và sau đó sẽ được cache lại edge location ở gần việt nam

Tiếp theo người dùng mới truy cập youtube xem video này thay vì nhận dự liệu ở mỹ thì nhận dự liệu từ cache edge location điều này sẽ làm giảm độ trễ cho video này

7. dịch vụ indentity and access management (IAM)

Iam là dịch vụ để xác thức và cấp quyền các tài nguyên của aws như ec2 ,s3

Iam cho các admin gán quyền user sử dùng tài nguyên aws mà không cần phải share user và pass của minh

User là người dùng tương tác với aws

Group nhóm các user thường nhóm lại có các vại trò giống nhau để dễ dàng quản lý

Policy là 1 văn bản viết dưới dạng json

Khi policicy được gắn vào user, group, role nó sẽ định nghĩa quyền hạn permission của thực thể này

Khi các thực thể này yêu cầu truy cập đến 1 tài nguyên của aws thì policy sẽ đánh giá xem yêu cầu chấp nhận hay ko

Role khác với user là có thể sử dụng bởi bất kì ai được phép

7. EC2

Elastic compute cloud

Ec nó sẽ cung cấp các máy tính để thực hiện hiện tính toán. Các máy máy tính này là máy tính ảo từ các hạ tầng của aws được cung cấp dịch qua internet

1 máy ảo chạy trên aws gọi là instance

**3. Các loại Instance EC2**

AWS cung cấp nhiều loại instance phục vụ các nhu cầu khác nhau:

| **Loại** | **Mô tả** | **Ví dụ** |
| --- | --- | --- |
| **General Purpose** | Cân bằng giữa CPU, RAM và I/O | t2.micro, m5.large |
| **Compute Optimized** | Tối ưu hiệu năng xử lý CPU | c5.large, c6g.xlarge |
| **Memory Optimized** | Dành cho ứng dụng cần nhiều RAM | r5.large, x1e.xlarge |
| **Storage Optimized** | Dành cho ứng dụng I/O cao | i3.large, d2.xlarge |
| **GPU Instances** | Dùng cho AI, Machine Learning, Đồ họa | g4dn.xlarge, p4d.24xlarge |

**6. Chi phí sử dụng EC2 💰**

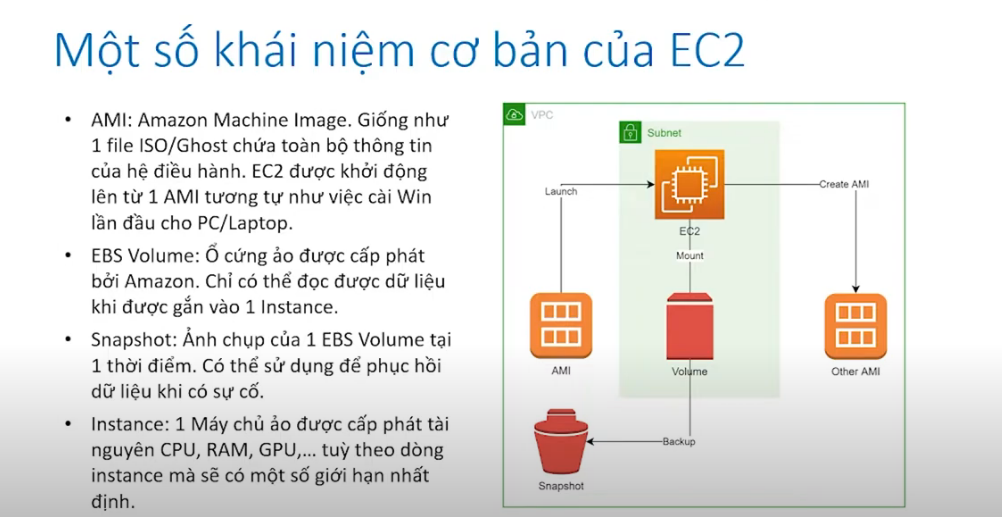
Bạn chỉ trả tiền cho tài nguyên sử dụng, bao gồm:

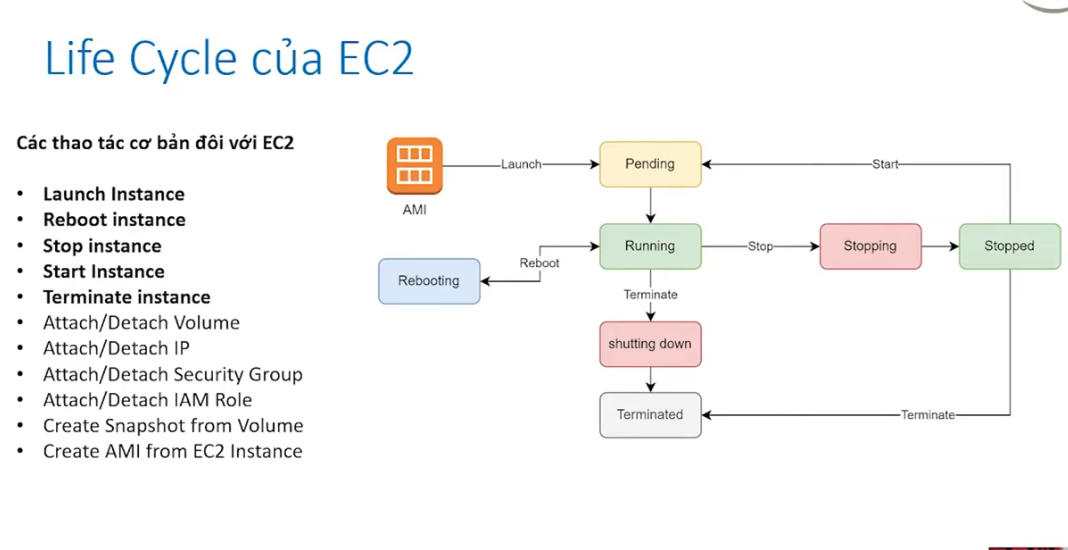
* **On-Demand Pricing** – Trả theo giờ/giây (linh hoạt, giá cao hơn).
* **Spot Pricing** – Giá rẻ hơn nhưng có thể bị AWS thu hồi nếu tài nguyên thiếu hụt.

**Reserved Pricing – Trả trước theo năm, giảm đến 75%.**

**Auto Scaling là gì?**

* Auto Scaling là quy trình tự động add mới hoặc remove EC2 instance dựa trên yêu cầu traffic của application hoặc của hệ thống.
* **ELB (Elastic Load Balancer) là gì? 🚀**
* **ELB (Elastic Load Balancer)** là một dịch vụ của AWS giúp **phân phối lưu lượng mạng (traffic)** giữa nhiều EC2 instances hoặc container để **cân bằng tải**, giúp hệ thống hoạt động ổn định và có khả năng mở rộng tốt hơn.
* **Mô hình cơ bản của Auto Scaling**





Tạo mới instance từ ami sau đó vô trạng thái pending và sau khi hoàn tất quá trình khởi động sẽ vào trạng thái running. bạn có thể thực hiện các hành động khác trên phiên bản. Chẳng hạn như khởi động lại (**reboot**), chấm dứt (**terminate**), dừng (**stop**) và dừng chế độ ngủ đông (**stop-hibernate**).

**Tổng quan về AWS ECS (Elastic Container Service)**

**1. AWS ECS là gì?**

**Amazon ECS (Elastic Container Service)** là dịch vụ quản lý container của AWS, giúp bạn chạy, dừng và quản lý **Docker container** trên cụm **EC2 instances** hoặc **AWS Fargate** mà không cần tự quản lý cơ sở hạ tầng.

**2. Kiến trúc ECS**

AWS ECS bao gồm các thành phần chính sau:

**1️⃣ Cluster**

* Là tập hợp các tài nguyên tính toán để chạy container (**EC2 instances hoặc Fargate**).

**2️⃣ Task Definition**

* Là cấu hình mô tả container, gồm:  
  ✅ Image Docker  
  ✅ CPU & RAM  
  ✅ Network (bridge, host, awsvpc)  
  ✅ Logging (CloudWatch Logs)  
  ✅ Các biến môi trường

**3️⃣ Task**

* Là một instance của **Task Definition** đang chạy trên ECS.

**4️⃣ Service**

* Đảm bảo **Task** chạy liên tục (có thể mở rộng hoặc tự phục hồi khi bị lỗi).