Contents

[**AWS Compute** 1](#_Toc136870494)

[**1 Amazon EC2** 1](#_Toc136870495)

[**2 Amazon EC2 auto scaling** 2](#_Toc136870496)

[**2.1 AWS Auto Scaling** 2](#_Toc136870497)

[**3 Amazon Elastic load balancing** 4](#_Toc136870498)

[**4 Amazon Elasitic contanainer service (ECS)** 4](#_Toc136870499)

[**5 Amazon Elastic Kubernetes Service (Amazon EKS)** 5](#_Toc136870500)

[**6 Amazon Elasitic Beanstalk** 5](#_Toc136870501)

[**7 AWS CloudFormation** 5](#_Toc136870502)

[**8 AWS Fargate** 6](#_Toc136870503)

[**9 AWS Lambda** 6](#_Toc136870504)

[**Outros serviços** 6](#_Toc136870505)

[**Qual serviço usar** 7](#_Toc136870506)

# **AWS Compute**



## **1 Amazon EC2**

O Amazon Elastic Compute Cloud (Amazon EC2) é um serviço Web que disponibiliza capacidade computacional segura e redimensionável na nuvem. O EC2 oferece muitas opções que permitem criar e executar virtualmente qualquer aplicativo.

* as instâncias EC2 são flexíveis confiáveis e escaláveis
* O Amazon EC2 requer que você configure e gerencie sua frota de instâncias. Você é responsável por aplicar os testes de suas instâncias quando novos pacotes de software forem lançados
* É responsável também por configurar a escalabilidade dessas instâncias arquitetando as suas soluções para serem hospedadas de uma maneira altamente disponível

Com os serviços **Serverless** vc não consegue ver e nem acessar as infraestruturas subjacente. ( **lambda** é serverles)

**Modalidades de contrato**

* **On-Demand Instances – short workload, predictable pricing, pay by second** 
  + **Linux or Windows - billing per second, after the first minute (All other operating systems - billing per hour)**
  + **Has the highest cost but no upfront payment**
  + **No long-term commitment**
  + **Recommended for short-term and un-interrupted workloads, where you can't predict how the application will behave**
* **Reserved (1 & 3 years)** 
  + **Reserved Instances (RIs) – long workloads** 
    - **Up to 72% discount compared to On-demand**
    - **Payment Options – No Upfront (+discount ), Partial Upfront (++discount), All Upfront (+++discount )**
    - **Reserved Instance’s Scope – Regional or Zonal (reserve capacity in an AZ)**
    - **Recommended for steady-state usage applications (think database)**
    - **You can buy and sell in the Reserved Instance Marketplace**
  + **Convertible Reserved Instances – long workloads with flexible instances**
    - **Can change the EC2 instance type, instance family, OS, scope and tenancy**
    - **Up to 66% discount**
* **Savings Plans (1 & 3 years) – commitment to an amount of usage, long workload** 
  + **Get a discount based on long-term usage (up to 72% - same as RIs)**
  + **Savings Plans doesn´t cover Ris**
  + **Locked to a specific instance family & AWS region (e.g., M5 in us-east-1)**
  + **Flexible across:** 
    - **Instance Size (e.g., m5.xlarge, m5.2xlarge)**
    - **OS (e.g., Linux, Windows)**
    - **Tenancy (Host, Dedicated, Default)**

|  |  |
| --- | --- |
| **Saving Plans** | **IRs** |
| * **Utilização em soluções computacionais como EC2, Fargate ou Lambda.** * **Os Compute Savings Plans permitem trocas entre famílias de instâncias, sistemas operacionais, tamanhos, locações e regiões.** * **Adaptação automática ao ambiente.** * **Descontos na utilização do Fargate e/ ou SageMaker.** * **Ambientes flexíveis e menor gerenciamento/monitoração.** | * **Utilização em soluções de Database, como RDS, Redshift, Elasticsearch.** * **Planeja sua utilização em pelo menos 75% do tempo de contrato.** * **Utilize em ambientes onde você tenha aplicativos que fiquem sempre “online”.** * **Necessite de uma categoria específica de instância e um SO.** * **Menos flexível e maior gerenciamento/monitoração.** |

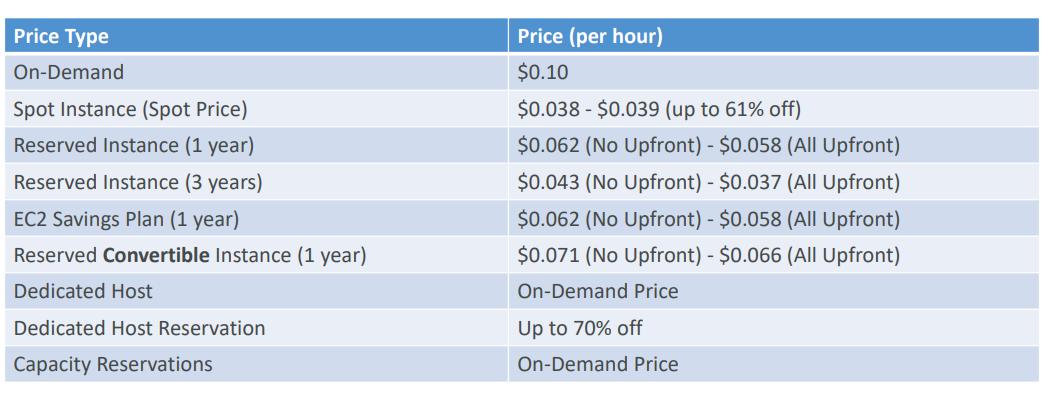
* + **Os tipos de plano Compute e EC2 Instance se aplicam a instâncias do EC2 que fazem parte dos clusters do Amazon EMR, Amazon EKS e Amazon ECS. As cobranças do Amazon EKS não serão cobertas pelos Savings Plans, mas as instâncias do EC2 subjacentes serão**
* **Spot Instances – short workloads, cheap, can lose instances (less reliable)** 
  + **Can get a discount of up to 90% compared to On-demand**
  + **Instances that you can “lose” at any point of time if your max price is less than the current spot price**
  + **The MOST cost-efficient instances in AWS**
  + **Useful for workloads that are resilient to failure**
    - **Batch jobs**
    - **Data analysis**
    - **Image processing**
    - **Any distributed workloads**
    - **Workloads with a flexible start and end time**
  + **Not suitable for critical jobs or databases**
* **Dedicated Hosts – book an entire physical server (EC2), control instance placement** 
  + **Allows you address compliance requirements and use your existing server- bound software licenses (per-socket, per-core, pe—VM software licenses)**
  + **Purchasing Options:** 
    - **On-demand – pay per second for active Dedicated Host**
    - **Reserved - 1 or 3 years (No Upfront, Partial Upfront, All Upfront)**
  + **The most expensive option**
  + **Useful for software that have complicated licensing model (BYOL – Bring Your Own License)**
  + **Or for companies that have strong regulatory or compliance needs**
* **Dedicated Instances – no other customers will share your hardware** 
  + **May share hardware with other instances in same account**



* **Capacity Reservations – reserve capacity in a specific AZ (**zona de disponibilidade) for any duration
  + **No time commitment (create/cancel anytime), no billing discounts**
  + **Combine with Regional Reserved Instances and Savings Plans to benefit from billing discounts**
  + **You’re charged at On-Demand rate whether you run instances or not**
  + **Suitable for short-term, uninterrupted workloads that needs to be in a specific AZ**

**Which purchasing option is right for me?**

* **On demand: coming and staying in resort whenever we like, we pay the full price**
* **Reserved: like planning ahead and if we plan to stay for a long time, we may get a good discount.**
* **Savings Plans: pay a certain amount per hour for certain period and stay in any room type (e.g., King, Suite, Sea View, …)**
* **Spot instances: the hotel allows people to bid for the empty rooms and the highest bidder keeps the rooms. You can get kicked out at any time**
* **Dedicated Hosts: We book an entire building of the resort**
* **Capacity Reservations: you book a room for a**

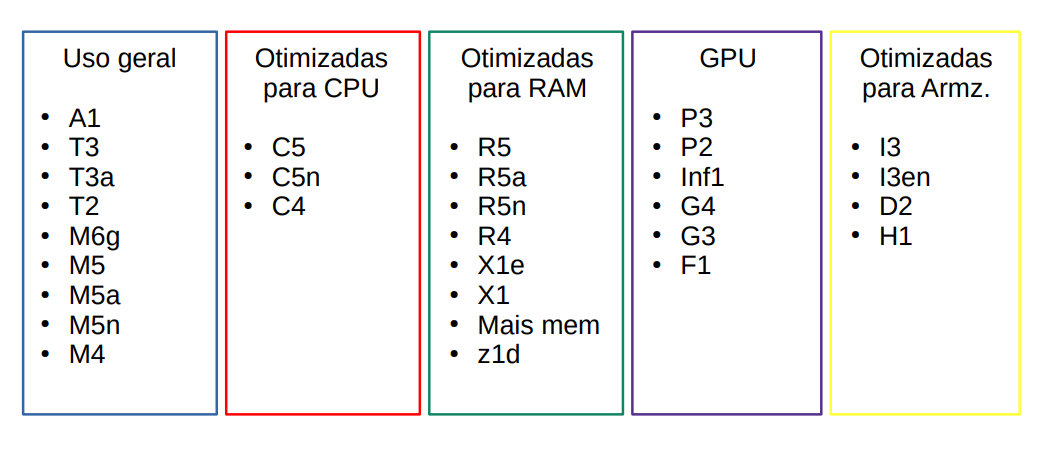
[Economias no custo da nuvem – Savings Plans – Amazon Web Services](https://aws.amazon.com/pt/savingsplans/)[O que são Savings Plans? - Savings Plans (amazon.com)](https://docs.aws.amazon.com/pt_br/savingsplans/latest/userguide/what-is-savings-plans.html)[Analisar custos com oAWS Cost Explorer - Gerenciamento de Custos da AWS (amazon.com)](https://docs.aws.amazon.com/pt_br/cost-management/latest/userguide/ce-what-is.html)

**Types instances**

<https://instances.vantage.sh/>

<https://aws.amazon.com/pt/ec2/instance-types/>

<https://docs.aws.amazon.com/pt_br/AWSEC2/latest/UserGuide/instance-types.html>



## **2 Amazon EC2 auto scaling**

O Amazon EC2 Auto Scaling permite que você adicione ou remova automaticamente instâncias do Amazon EC2 em resposta à alteração da demanda do aplicativo.

No Amazon EC2 Auto Scaling, há duas abordagens disponíveis: scaling dinâmico e scaling preditivo.

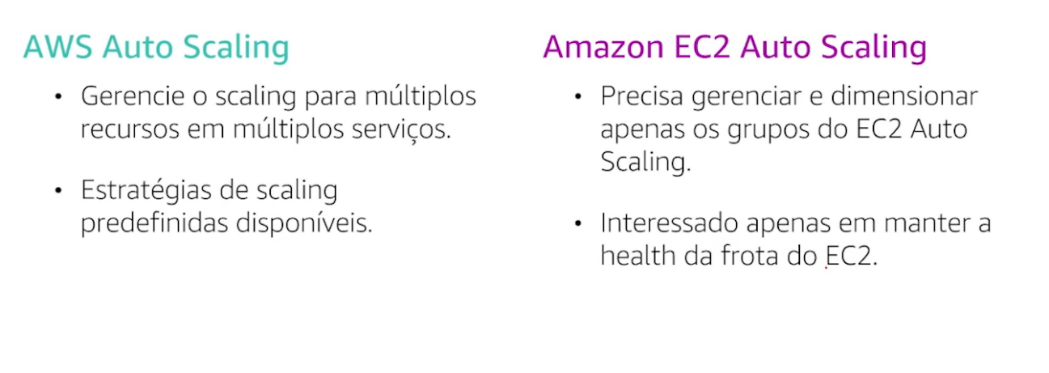
* O *scaling dinâmico* responde às alterações na demanda.
* O *scaling preditivo*programa automaticamente o número correto de instância do Amazon EC2 com base na demanda prevista.
* Executa e termina instancia de acordo com as regra definidas para aumentar ou diminuir.
* Registra as instancia automaticamente nos balanceadores de carga (ELB)
* Pode executar e escalar instancia em zonas de disponibilidades diferentes.
* **Integra-se ao Elastic Load Balancing**
* O Amazon EC2 Auto Scaling pode detectar quando uma instância não está íntegra, encerrá-la e substituí-la por uma nova.
* O Amazon EC2 Auto Scaling garante que seu aplicativo sempre tenha a quantidade certa de recursos de computação e também provisiona capacidade proativamente com a escalabilidade preditiva.
* O Amazon EC2 Auto Scaling adiciona instâncias apenas quando necessário e pode [escalar entre as opções de compra](https://aws.amazon.com/blogs/aws/new-ec2-auto-scaling-groups-with-multiple-instance-types-purchase-options) para otimizar o desempenho e o custo.
* Permite trabalhar com escalonamento agendando

## **2.1 AWS Auto Scaling**

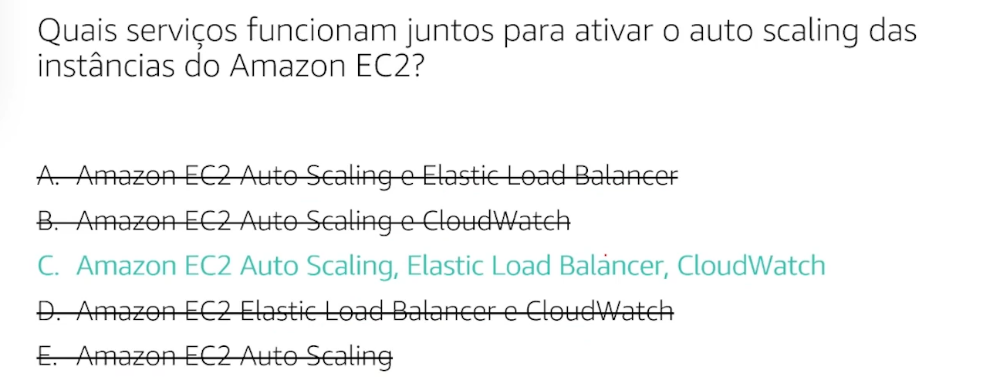
O AWS Auto Scaling monitora os aplicativos e ajusta automaticamente a capacidade para manter um desempenho constante e previsível pelo menor custo possível. Com o AWS Auto Scaling, é fácil configurar a escalabilidade de aplicativos para vários recursos em diversos serviços em questão de minutos. O serviço oferece uma interface de usuário simples e eficiente que permite criar planos de escalabilidade para recursos, como instâncias e frotas spot do [Amazon EC2](https://aws.amazon.com/ec2/), **tarefas do**[**Amazon ECS**](https://aws.amazon.com/ecs/)**, tabelas e índices do**[**Amazon DynamoDB**](https://aws.amazon.com/dynamodb/)**e réplicas do**[**Amazon Aurora**](https://aws.amazon.com/aurora/)**.**

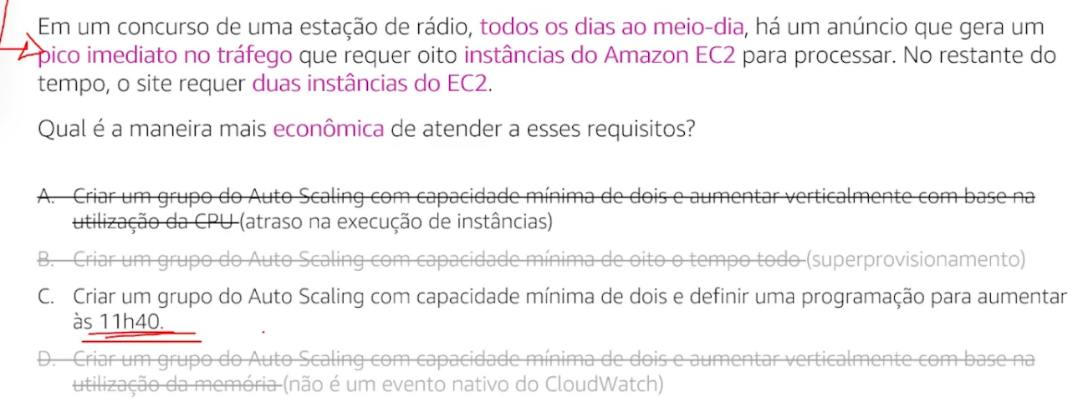
É fácil começar a usar o AWS Auto Scaling utilizando o [Console de gerenciamento da AWS](https://console.aws.amazon.com/console/home), a Interface da linha de comando (ILC) ou o SDK. **O AWS Auto Scaling está disponível sem custos adicionais**. Você paga apenas pelos recursos da AWS necessários para executar os aplicativos e pelas taxas de monitoramento do [Amazon CloudWatch](https://aws.amazon.com/cloudwatch/).

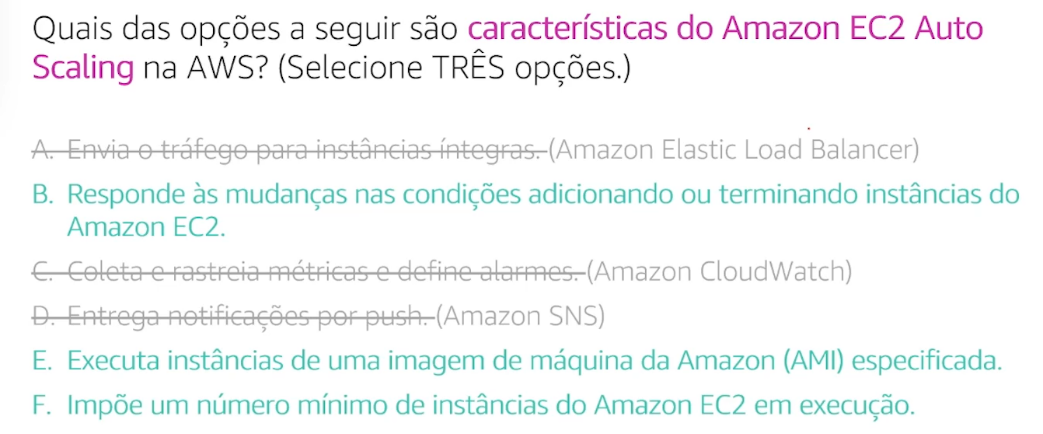
* Gerencia o scaling de múltiplos recusros e serviços gerenciados da aws
* Atua com estratégica de scaling predefinidas disponíveis.



**Questão de exemplo referente ao auto scaling**







<https://aws.amazon.com/pt/autoscaling/>

<https://aws.amazon.com/pt/autoscaling/faqs/>

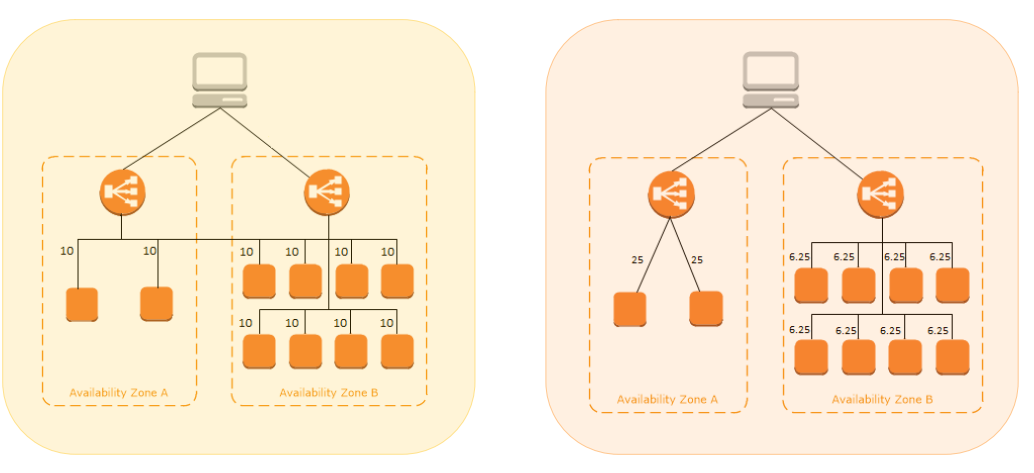
<https://aws.amazon.com/pt/ec2/autoscaling/>

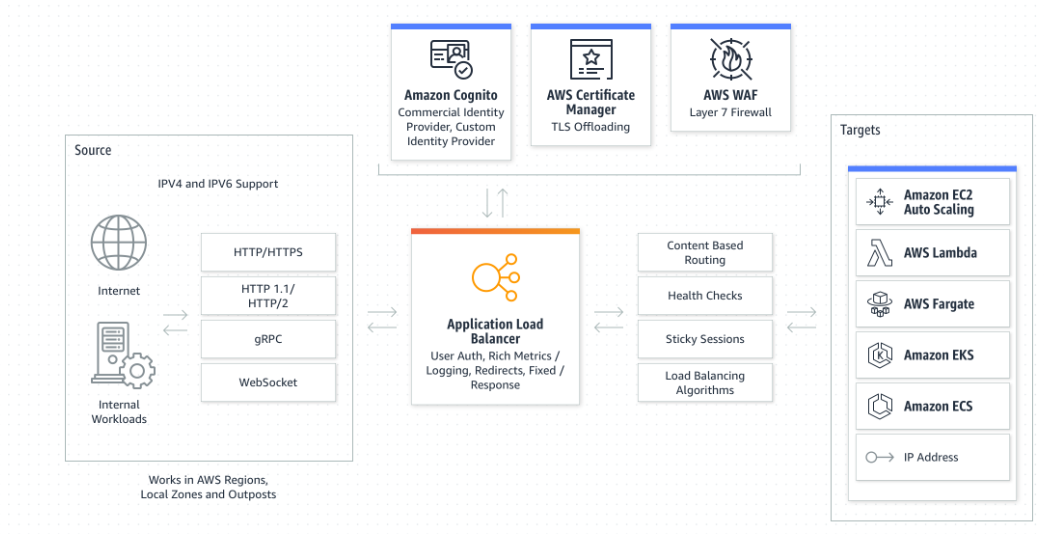
<https://aws.amazon.com/pt/ec2/autoscaling/features/>

<https://aws.amazon.com/pt/ec2/autoscaling/faqs/>

## **3 Amazon Elastic load balancing**

O Elastic Load Balancing é o serviço AWS que distribui automaticamente o tráfego de entrada de aplicativos entre vários recursos, como instâncias do Amazon EC2. Isso ajuda a garantir que nenhum recurso único seja usado em excesso.





Existem 3 tipos de Load Balancer

1. **Aplication Load Balancer:** Http/Https vai até a 7 camanda

* Atuam na camada 7 (aplicação) do modelo de referência OSI
* Categoria mais recomendada para balanceamento de tráfego HTTP e HTTPS
* Permitem roteamento avançado
* Escala automaticamente
* Endereço IP NÃO estático
* O endereço IP do cliente NÃO é visualizado nos Targets. Apenas o endereço IP do LB é disponibilizado.
* Multi-AZ
* Permite a utilização de stickiness

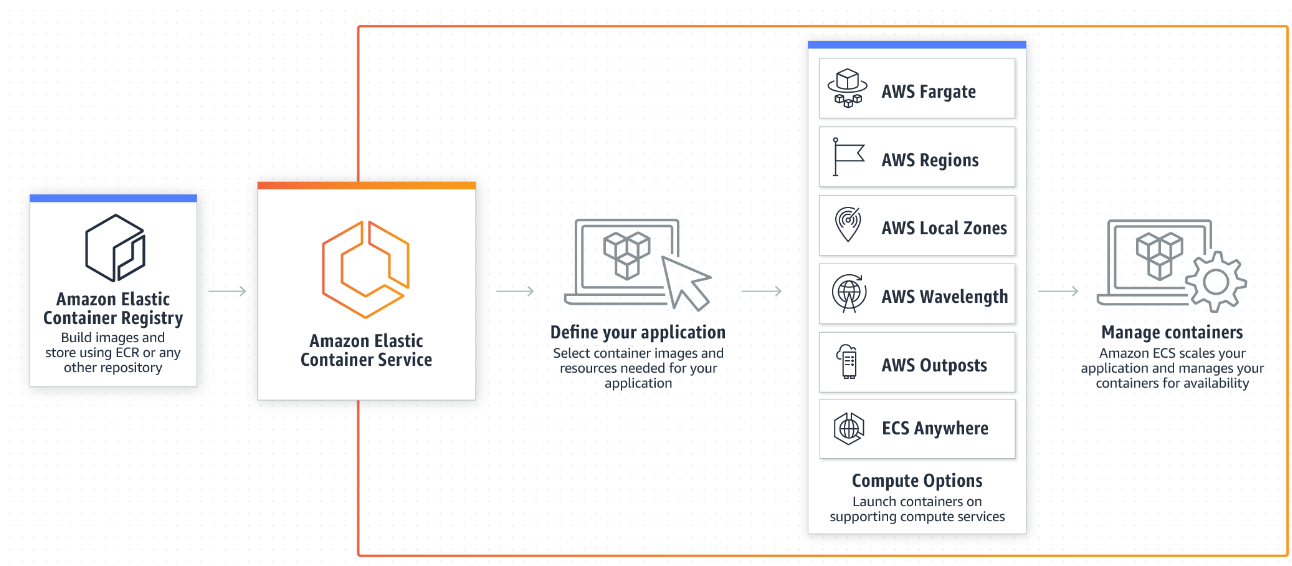
1. **Network Load Balancer:** Http/Https vai até a 4 camanda

* Atuam na camada 4 (transporte) do modelo de referência OSI
* Categoria mais recomendada para balanceamento de tráfego TCP e UDP
* Capaz de lidar com milhões de requisições por segundo
* Escala automaticamente
* Possui 1 endereço IP estático por subnet
* Multi-AZ
* Sem terminações SSL

1. **Elastic Load Balancer (classic LB):** legado será descontinuado

## **4 Amazon Elasitic contanainer service (ECS)**

O Amazon ECS é um serviço de orquestração de contêineres totalmente gerenciado que facilita a implantação, o gerenciamento e a escala de aplicações em contêineres.



## **5 Amazon Elastic Kubernetes Service (Amazon EKS)**

O Amazon EKS é um serviço totalmente gerenciado do Kubernetes. O Kubernetes é um software de código aberto que permite implantar e gerenciar aplicativos em contêineres em grande escala.

## **6 Amazon Elasitic Beanstalk**

O AWS Elastic Beanstalk é um serviço que ajuda você a provisionar ambientes baseados no Amazon EC2. Em vez de clicar pela console ou escrever vários comandos para criar sua rede, instâncias EC2, escalabilidade e os load balances, você vai fornecer o seu código da aplicação com as configurações desejadas no Beanstalk e, em seguida, ele cria o seu ambiente e faz o seu deploy. O Elastic Beanstalk também permite salvar as configurações do ambiente para que elas possam ser implantadas novamente. O AWS Elastic Beanstalk fornece a conveniência de não ter que provisionar e gerenciar todos esses componentes separadamente e ainda te provê a visibilidade e o controle dos recursos que compõem o seu ambiente. Você irá focar, então, na sua aplicação e não na infraestrutura.

Com o **AWS Elastic Beanstalk,** você fornece definições de código e configuração, e o Elastic Beanstalk implanta os recursos necessários para executar as seguintes tarefas:

* Ajustar capacidade
* Balancear carga
* Dimensionar de forma automática
* Monitorar a integridade do aplicativo

## **7 AWS CloudFormation**

O AWS CloudFormation é uma ferramenta que trata a infraestrutura como código, permitindo que você defina uma ampla variedade de recursos da AWS usando JSON ou Yaml, chamados de Templates do CloudFormation. Um formato declarativo como esse permite definir o que você deseja criar sem especificar o passo a passo exatamente de como fazer essa criação. Portanto, o CloudFormation te permite definir o que você deseja e o mecanismo dele vai se preocupar com os detalhes sobre como chamar as APIs para fazer essa criação.  
  
Também não é limitado apenas para soluções básicas que contemplem instâncias EC2. O CloudFormation é compatível com diversos recursos diferentes da AWS, como armazenamento, banco de dados, machine learning e muito mais. Depois de definir os seus recursos em um template do CloudFormation, o serviço vai analisar o template e começar a provisionar todos os recursos que você definiu em paralelo. O CloudFormation gerencia todas as chamadas de API para o back end da AWS para você. Você pode executar o mesmo template do CloudFormation em várias contas ou em várias regiões e ele criará ambientes idênticos em todos eles. Há menos espaço para erro humano uma vez que é um processo totalmente automatizado.

## **8 AWS Fargate**

O AWS Fargate é um mecanismo de computação sem servidor para contêineres.

## **9 AWS Lambda**

é um serviço que permite executar código sem provisionar ou gerenciar servidores.

O Lambda é um serviço que permite que você faça o upload do seu código no que chamamos de função lambda, configure um trigger, ou seja, um evento que vai iniciar a execução da sua função Lambda e o serviço vai aguardar o trigger. Quando o Trigger é detectado, o código é automaticamente executado em um ambiente gerenciado

O Lambda foi projetado para executar um código em até 15 minutos, então, não foi criado para o processamento de longa execução, como o Deep learning, por exemplo.

## **Outros serviços**

* O Amazon Simple Queue Service (Amazon SQS) é um serviço que permite enviar, armazenar e receber mensagens entre componentes de software por uma fila.
* O Amazon Simple Notification Service (Amazon SNS) é um serviço de publicação/assinatura. Usando tópicos do Amazon SNS, um editor publica mensagens para assinantes.

## **Qual serviço usar**

Se você ainda não está pronto para trabalhar de forma serverless ou precisa acessar e gerenciar o seu ambiente mas ainda assim quer eficiência e portabilidade deve olhar para serviços de contêiner, como, por exemplo, o Amazon Elastic Container Service, também conhecido como ECS, ou o Amazon Elastic Kubernetes Service, também conhecido como EKS.  
  
Ambos os serviços são ferramentas de orquestração de container

**Se você estiver tentando hospedar suas aplicações tradicionais e desejar sucesso total o sistema operacional como Linux ou Windows, você vai optar por instâncias EC2.**

**Se você deseja trabalhar com processamento de curta duração, orientados a eventos e serviços ou aplicações e não deseja gerenciar o seu ambiente, cheque a AWS Lambda.**

**Se você deseja executar workloads baseados em containers Docker na AWS, primeiro precisa escolher a sua ferramenta de orquestração. O que você prefere: o ECS ou o EKS? Depois de escolher a sua ferramenta, você precisa escolher a sua plataforma. Você prefere executar seus containers em instâncias do EC2 que você gerencia ou utilizando o Fargate que é gerenciado pra você?**

Escalabilidade

Você pode redimensionar as instâncias EC2 verticalmente, aumentando o tamanho da instância, ou horizontalmente, adicionando novas instâncias no grupo. Você pode configurar a escalabilidade horizontal automatizada utilizando o Amazon EC2 Auto Scaling.  
  
**Depois de redimensionar as suas instâncias EC2 horizontalmente, Você precisa de algo para distribuir o tráfego de entrada entre todas as instâncias. E é ai que entra o Elastic Load Balancing. As instâncias EC2 têm diferentes modelos de preços: há o sob demanda, que é on demand, que é mais flexível e não tem um compromisso. Tem um Modelo spot, que permite a utilização da capacidade computacional que não está sendo utilizada no momento por um valor inferior ao on demand. Saving plans em instâncias reservadas, que permitem que você obtenha um desconto ao se comprometer a um determinado nível de uso, e Compute Savings plans que se aplicam ao AWS Lambda, Fargate, bem como às instâncias EC2.**

**ambém abordamos os serviços de mensageria, o Amazon Simple Queue Service, ou SQS. Esse serviço permite deaacoplar componentes do seu sistema. As mensagens permanecem na fila até que sejam consumidas ou excluídas. Vimos também o Amazon Simple Notification Service, ou SMS, que e é usado para enviar mensagens para assinantes de e-mails, mensagens de texto, notificações push ou até mesmo endpoints HTTP e HTTPS. Depois que uma mensagem é publicada em um tópico, ela é enviada para todos esses assinantes.**

Os **armazenamentos de instâncias** são ideais para dados temporários que não precisam ser mantidos a longo prazo

Quando uma instância do Amazon EC2 é interrompida ou encerrada, todos os dados que foram gravados no armazenamento de instâncias anexado são excluídos.