**EC2**

**Tipos de Instancias**

**Instâncias de uso geral**

**Instâncias otimizadas para computação**

**Instâncias otimizadas para memória**

**Instâncias de computação acelerada**

**Instâncias otimizadas para armazenamento**

**Instâncias de uso geral**

equilibram os recursos de computação, memória e rede. Você pode usá-las para diversas cargas de trabalho, como:

* servidores de aplicativos
* servidores de jogos
* servidores de back-end para aplicativos empresariais
* bancos de dados pequenos e médios

Suponha que você tenha um aplicativo no qual as necessidades de recursos para computação, memória e rede sejam praticamente equivalentes. Você pode executar esse aplicativo em uma instância de uso geral porque ele não precisa de otimização em nenhuma área de recurso único.

**Instâncias otimizadas para computação**

são ideais para aplicativos vinculados à computação que se beneficiam de processadores de alto desempenho. Assim como instâncias de uso geral, você pode usar instâncias otimizadas para computação para cargas de trabalho, como servidores web, de aplicativos e de jogos.  
  
No entanto, a diferença é que aplicativos otimizados para computação são ideais para servidores web de alto desempenho, servidores de aplicativos de computação intensiva e servidores de jogos dedicados. Você também pode usar instâncias otimizadas para computação para cargas de trabalho de processamento em lote, com o processamento de muitas transações em um único grupo.

**Instâncias otimizadas para memória**

são projetadas para fornecer desempenho rápido para cargas de trabalho que processam grandes conjuntos de dados na memória. Na computação, a memória é uma área de armazenamento temporário. Ela contém todos os dados e instruções de que uma unidade central de processamento (CPU) precisa para conseguir realizar ações. Antes que um programa de computador ou aplicativo possa ser executado, ele é carregado do armazenamento para a memória. Esse processo de pré-carregamento dá à CPU acesso direto ao programa de computador.  
  
Suponha que você tenha uma carga de trabalho que precise que grandes quantidades de dados sejam pré-carregados antes de executar um aplicativo. Esse cenário pode ser de um banco de dados de alto desempenho ou uma carga de trabalho que envolva a execução de processamento em tempo real de uma grande quantidade de dados não estruturados. Nesses tipos de casos de uso, considere usar uma instância otimizada para memória. As instâncias otimizadas para memória permitem que você execute cargas de trabalho com altas necessidades de memória e tenha um ótimo desempenho.

**Instâncias de computação acelerada**

usam aceleradores de hardware, ou coprocessadores, para executar algumas funções de forma mais eficiente do que é possível em um software executado em CPUs. Exemplos dessas funções são cálculos de números com vírgula flutuante, processamento de gráficos e correspondência de padrões de dados.  
Na computação, um acelerador de hardware é um componente que pode agilizar o processamento de dados. As instâncias de computação acelerada são ideais para cargas de trabalho, como aplicativos gráficos e streaming de jogos e de aplicativos.

**As instâncias otimizadas para armazenamento**

são projetadas para cargas de trabalho que exigem alto acesso sequencial de leitura e gravação a grandes conjuntos de dados no armazenamento local. Exemplos de cargas de trabalho adequadas para instâncias otimizadas para armazenamento são sistemas de arquivos distribuídos, aplicativos de data warehouse e sistemas de processamento de transações on-line de alta frequência (OLTP).  
Na computação, o termo operações de entrada/saída por segundo (IOPS) é uma métrica que mensura o desempenho de um dispositivo de armazenamento. Ela indica quantas operações diferentes de entrada ou saída um dispositivo pode executar em um segundo. As instâncias otimizadas para armazenamento foram projetadas para fornecer dezenas de milhares de IOPS aleatórias e de baixa latência para aplicativos.  
  
Imagine as operações de entrada como dados colocados em um sistema, como registros inseridos em um banco de dados. Uma operação de saída são dados gerados por um servidor. Um exemplo de saída pode ser a análise realizada nos registros em um banco de dados. Se você tiver um aplicativo com alto requisito de IOPS, uma instância otimizada para armazenamento poderá fornecer melhor desempenho em relação a outros tipos de instâncias não otimizados para esse tipo de caso de uso.

**Instâncias sob demanda**

–

**Instâncias sob demanda** são ideais para cargas de trabalho irregulares de curto prazo que não podem ser interrompidas. Custos antecipados ou contratos mínimos não se aplicam. As instâncias são executadas continuamente até que sejam interrompidas e você paga apenas pelo tempo de computação que usar.  
  
Exemplos de casos de uso para instâncias sob demanda são desenvolvimento e teste de aplicativos e execução de aplicativos com padrões de uso imprevisíveis. As instâncias sob demanda não são recomendadas para cargas de trabalho que duram um ano ou mais, porque essas cargas de trabalho podem ser mais econômicas usando instâncias reservadas.

**Savings Plans do Amazon EC2**

–

A AWS oferece Savings Plans para vários serviços computacionais, incluindo o Amazon EC2. O **Savings Plans do Amazon EC2** permite reduzir os custos de computação ao haver o compromisso com uma quantidade consistente de uso de computação por um período de um ou três anos. Esse compromisso resulta em economias de até 72% em relação aos custos de instâncias sob demanda.  
  
Qualquer uso até o compromisso é cobrado de acordo com o preço de Savings Plan com desconto (por exemplo, USD 10 por hora). Qualquer uso além do compromisso é cobrado de acordo com os preços normais de instâncias sob demanda.  
  
Mais adiante neste curso, você analisará o AWS Cost Explorer, uma ferramenta que permite visualizar, interpretar e gerenciar o uso e os custos da AWS ao longo do tempo. Se você estiver considerando o modelo Savings Plans, o AWS Cost Explorer poderá analisar seu uso do Amazon EC2 nos últimos 7, 30 ou 60 dias. O AWS Cost Explorer também dá recomendações personalizadas para Savings Plans. Essas recomendações calculam o quanto você pode economizar mensalmente com o Amazon EC2, com base no uso anterior do Amazon EC2 e no valor do compromisso por hora em um Savings Plan de um ou três anos.

**Instâncias reservadas**

–

**Instâncias reservadas** são um desconto de cobrança aplicado ao uso de instâncias sob demanda em sua conta. Você pode adquirir instâncias reservadas comuns e instâncias reservadas conversíveis por um período de um ou três anos, e instâncias reservadas agendadas por um período de um ano. Você tem mais economia com a opção de três anos.  
  
Ao final do período da instância reservada, você pode continuar usando a instância do Amazon EC2 sem interrupção. No entanto, são cobrados os preços de instâncias sob demanda até que um dos procedimentos a seguir seja feito:

* Encerrar a instância.
* Adquirir uma nova instância reservada que corresponda aos atributos da instância (tipo de instância, Região, locação e plataforma).

**Instâncias spot**

–

As **instâncias spot** são ideais para cargas de trabalho com horários de início e término flexíveis ou que toleram interrupções. As instâncias spot usam a capacidade de computação não utilizada do Amazon EC2 e têm uma economia de até 90% de desconto em relação aos preços das instâncias sob demanda.  
  
Suponha que você tenha um trabalho de processamento em segundo plano que pode ser iniciado e interrompido conforme for necessário (como o trabalho de processamento de dados para uma pesquisa de cliente). Você deseja iniciar e interromper o trabalho de processamento sem afetar as operações gerais de seus negócios. Se você fizer uma solicitação spot e a capacidade do Amazon EC2 estiver disponível, a instância spot será iniciada. No entanto, se você fizer uma solicitação spot e a capacidade do Amazon EC2 estiver indisponível, a solicitação não terá sucesso até que a capacidade seja disponibilizada. A capacidade indisponível pode atrasar a execução do trabalho de processamento em segundo plano.  
  
Depois de executar uma instância spot, se a capacidade não estiver mais disponível ou a demanda por instâncias spot aumentar, sua instância poderá ser interrompida. Isso pode não representar problemas para o trabalho de processamento em segundo plano. No entanto, no exemplo anterior de desenvolvimento e teste de aplicativos, é provável que você queira evitar interrupções inesperadas. Portanto, escolha um tipo de instância do EC2 diferente que seja ideal para essas tarefas.

**Hosts dedicados**

–

**Hosts dedicados**são servidores físicos com capacidade de instância do Amazon EC2 totalmente dedicada ao uso do cliente.  
  
Você pode usar suas licenças de software por soquete, por núcleo ou por VM existentes para manter a conformidade da licença. Você pode adquirir hosts dedicados sob demanda e reservas de hosts dedicados. De todas as opções do Amazon EC2 que foram abordadas, os hosts dedicados são os mais caros.

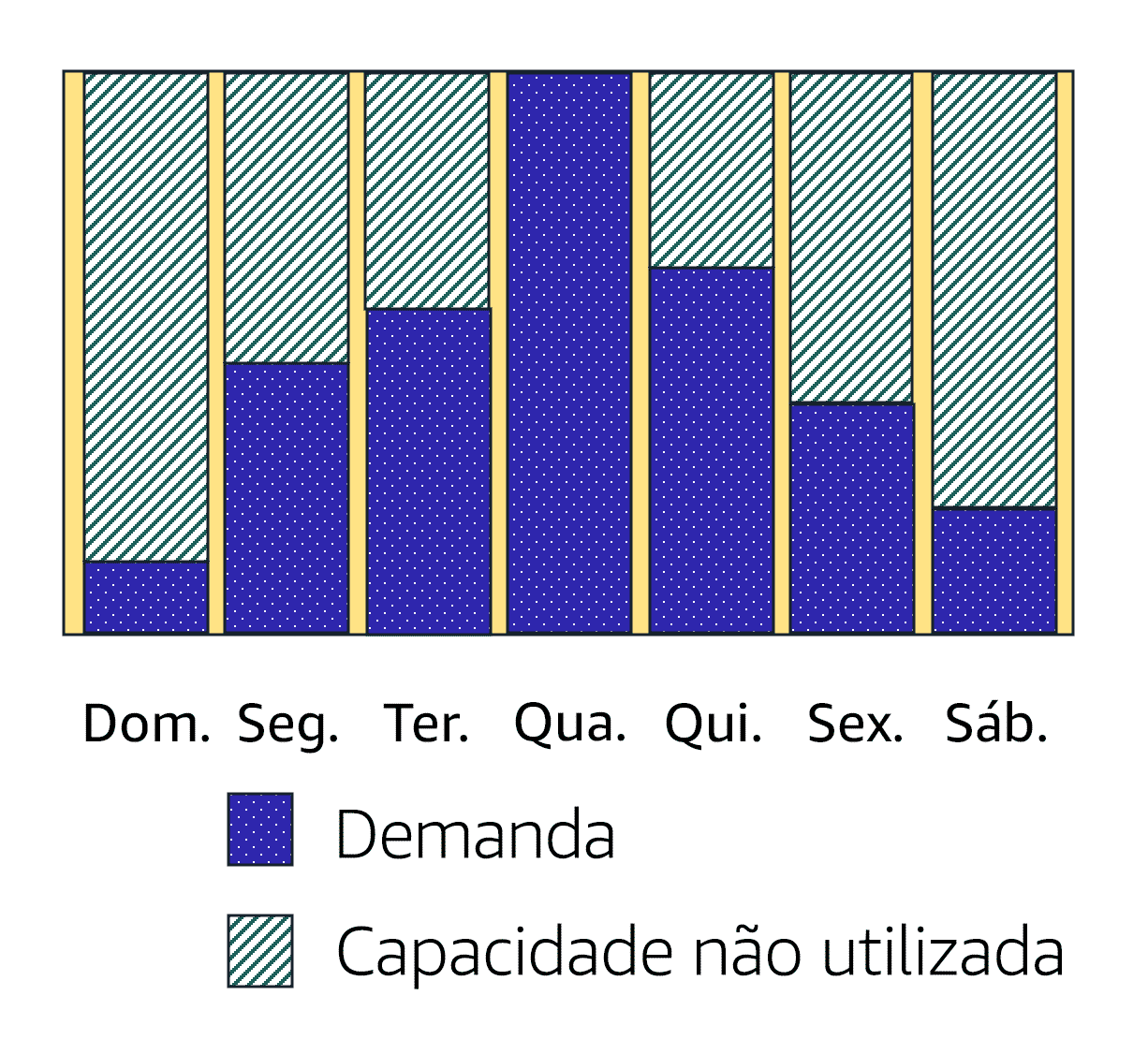
* [Computação na AWS](https://aws.amazon.com/products/compute)
* [Blog de computação na AWS](https://aws.amazon.com/blogs/compute/)
* [Serviços computacionais da AWS](https://docs.aws.amazon.com/whitepapers/latest/aws-overview/compute-services.html)
* [Tutoriais práticos: computação](https://aws.amazon.com/getting-started/hands-on/?awsf.getting-started-category=category%23compute&awsf.getting-started-content-type=content-type%23hands-on)
* [Aprofundamento na categoria: sem servidor](https://aws.amazon.com/getting-started/deep-dive-serverless/)
* [Histórias de clientes da AWS: sem servidor](https://aws.amazon.com/solutions/case-studies/?customer-references-cards.sort-by=item.additionalFields.publishedDate&customer-references-cards.sort-order=desc&awsf.customer-references-location=*all&awsf.customer-references-segment=*all&awsf.customer-references-product=product%23vpc%7Cproduct%23api-gateway%7Cproduct%23cloudfront%7Cproduct%23route53%7Cproduct%23directconnect%7Cproduct%23elb&awsf.customer-references-category=category%23serverless)

**Escalabilidade**

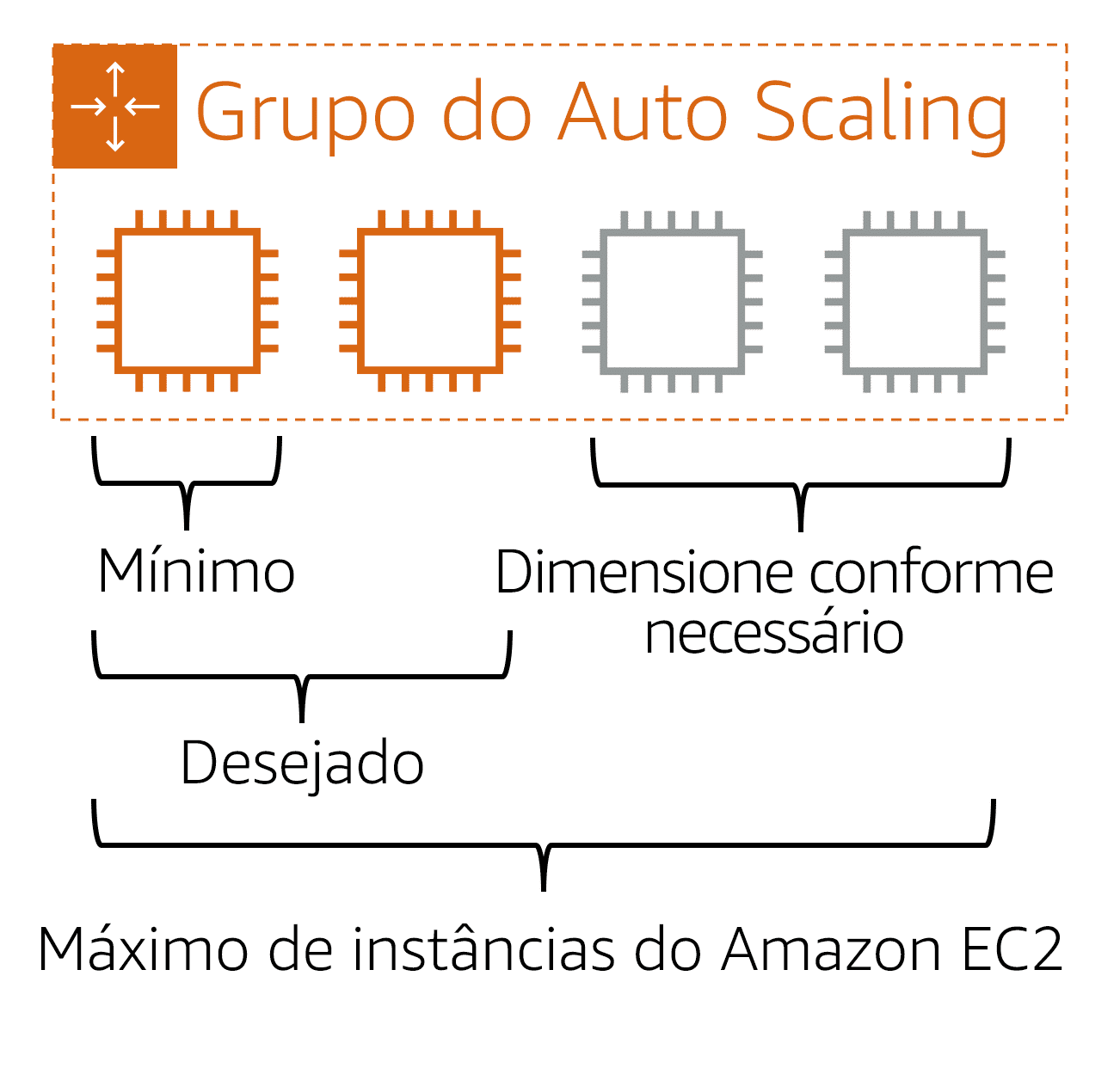
A **escalabilidade** envolve começar apenas com os recursos de que você precisa e projetar sua arquitetura para responder automaticamente às alterações de demanda, fazendo aumentos ou reduções. Como resultado, você paga apenas pelos recursos que usa. Você não precisa se preocupar com a falta de capacidade de computação para atender às necessidades de seus clientes.  
  
Se você quisesse que o processo de scaling acontecesse automaticamente, qual serviço AWS você usaria? O serviço AWS que fornece essa funcionalidade para instâncias do Amazon EC2 é o **Amazon EC2 Auto Scaling**.

**Amazon EC2 Auto Scaling**

Se você já tentou acessar um site que não carregava e atingiu o tempo limite algumas vezes, ele pode ter recebido mais solicitações do que conseguia atender. Essa situação é semelhante a esperar em uma longa fila em uma cafeteria quando há apenas um barista disponível para registrar os pedidos dos clientes.



O Amazon EC2 Auto Scaling permite que você adicione ou remova automaticamente instâncias do Amazon EC2 em resposta à alteração da demanda do aplicativo. Ao fazer auto scaling de suas instâncias, aumentando ou reduzindo conforme a necessidade, você consegue manter uma sensação maior de disponibilidade de aplicativos.  
  
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.
* **Exemplo: Amazon EC2 Auto Scaling**
* Já que na nuvem a capacidade computacional é um recurso programático, você pode adotar uma abordagem mais flexível para o problema de scaling. Ao adicionar o Amazon EC2 Auto Scaling a um aplicativo, você pode adicionar novas instâncias ao aplicativo quando for necessário e encerrá-las quando não forem mais necessárias.  
    
  Suponha que você esteja se preparando para executar um aplicativo em instâncias do Amazon EC2. Ao configurar o tamanho do seu grupo do Auto Scaling, você pode definir o número mínimo de instâncias do Amazon EC2 como sendo um. Isso significa que, em qualquer momento, deve haver pelo menos uma instância do Amazon EC2 em execução.
* 
* Ao criar um grupo do Auto Scaling, você pode definir o número mínimo de instâncias do Amazon EC2. A **capacidade mínima** é o número de instâncias do Amazon EC2 que são executadas imediatamente após a criação do grupo do Auto Scaling. Neste exemplo, o grupo do Auto Scaling tem uma capacidade mínima de uma instância do Amazon EC2.
* Em seguida, você pode definir a **capacidade desejada** como duas instâncias do Amazon EC2, mesmo que o aplicativo precise de um mínimo de uma única instância do Amazon EC2 para que seja executado.
* Se você não especificar o número desejado de instâncias do Amazon EC2 em um grupo do Auto Scaling, a capacidade desejada se tornará a capacidade mínima regular.
* A terceira configuração que você pode definir em um grupo do Auto Scaling é a **capacidade máxima**. Por exemplo, você pode configurar o grupo do Auto Scaling para aumentar em resposta à demanda elevada, mas apenas para um máximo de quatro instâncias do Amazon EC2.  
    
  Como o Amazon EC2 Auto Scaling usa instâncias do Amazon EC2, você paga apenas pelas instâncias que usar, e somente quando elas forem usadas. Você agora tem uma arquitetura econômica que fornece a melhor experiência do cliente e ao mesmo tempo reduz custos.

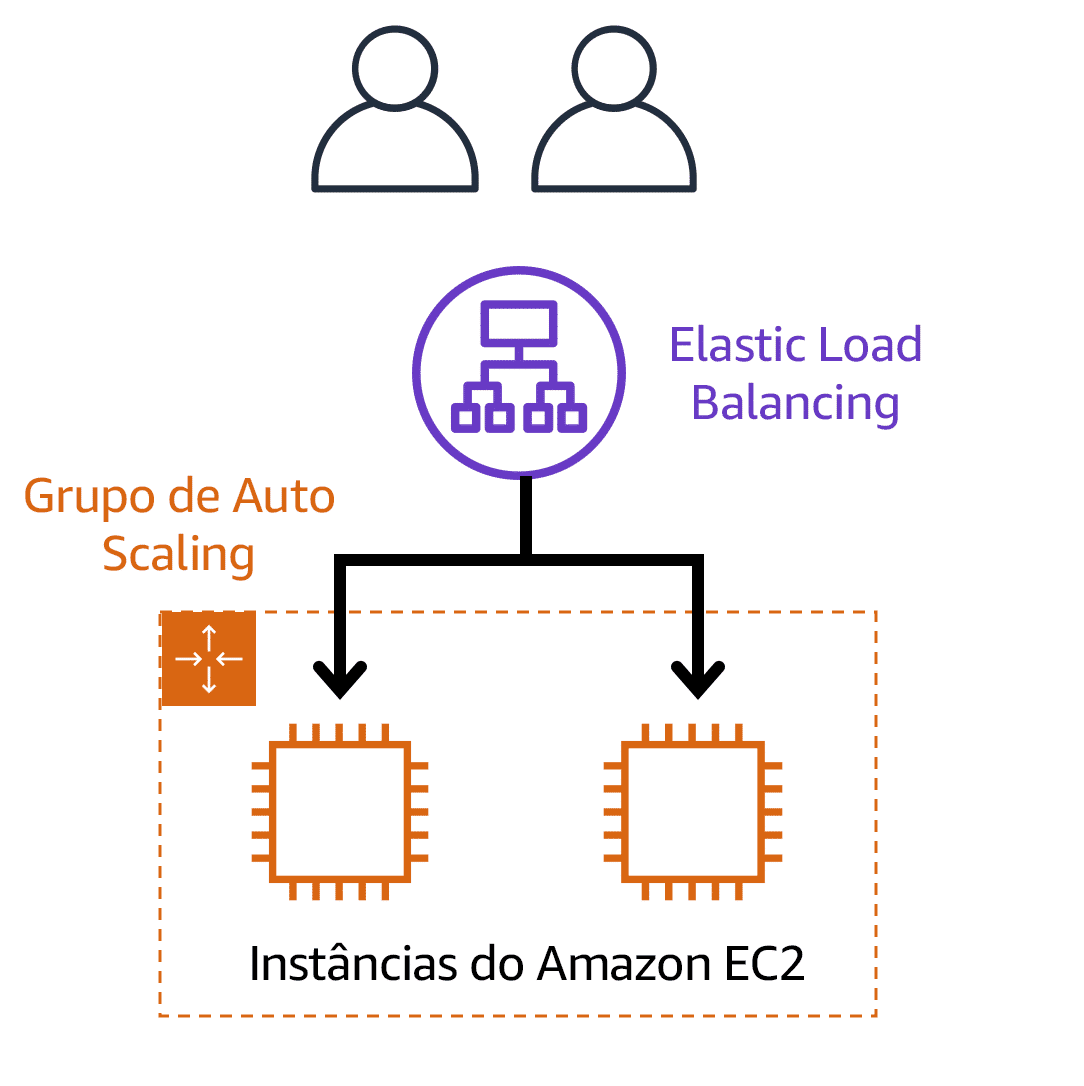
**Direcionamento de tráfego com o Elastic Load Balancing**

**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.  
  
Um balanceador de carga atua como um ponto único de contato para todo o tráfego da web de entrada no seu grupo do Auto Scaling. Isso significa que, à medida que você adiciona ou remove instâncias do Amazon EC2 em resposta à quantidade de tráfego de entrada, essas solicitações são direcionadas para o balanceador de carga primeiro. Em seguida, as solicitações se espalham por vários recursos que lidarão com elas. Por exemplo, se você tiver várias instâncias do Amazon EC2, o Elastic Load Balancing distribuirá a carga de trabalho entre elas para que nenhuma instância tenha que carregar a maior parte.  
  
Embora o Elastic Load Balancing e o Amazon EC2 Auto Scaling sejam serviços separados, eles trabalham juntos para que os aplicativos executados no Amazon EC2 possam fornecer alto desempenho e disponibilidade.

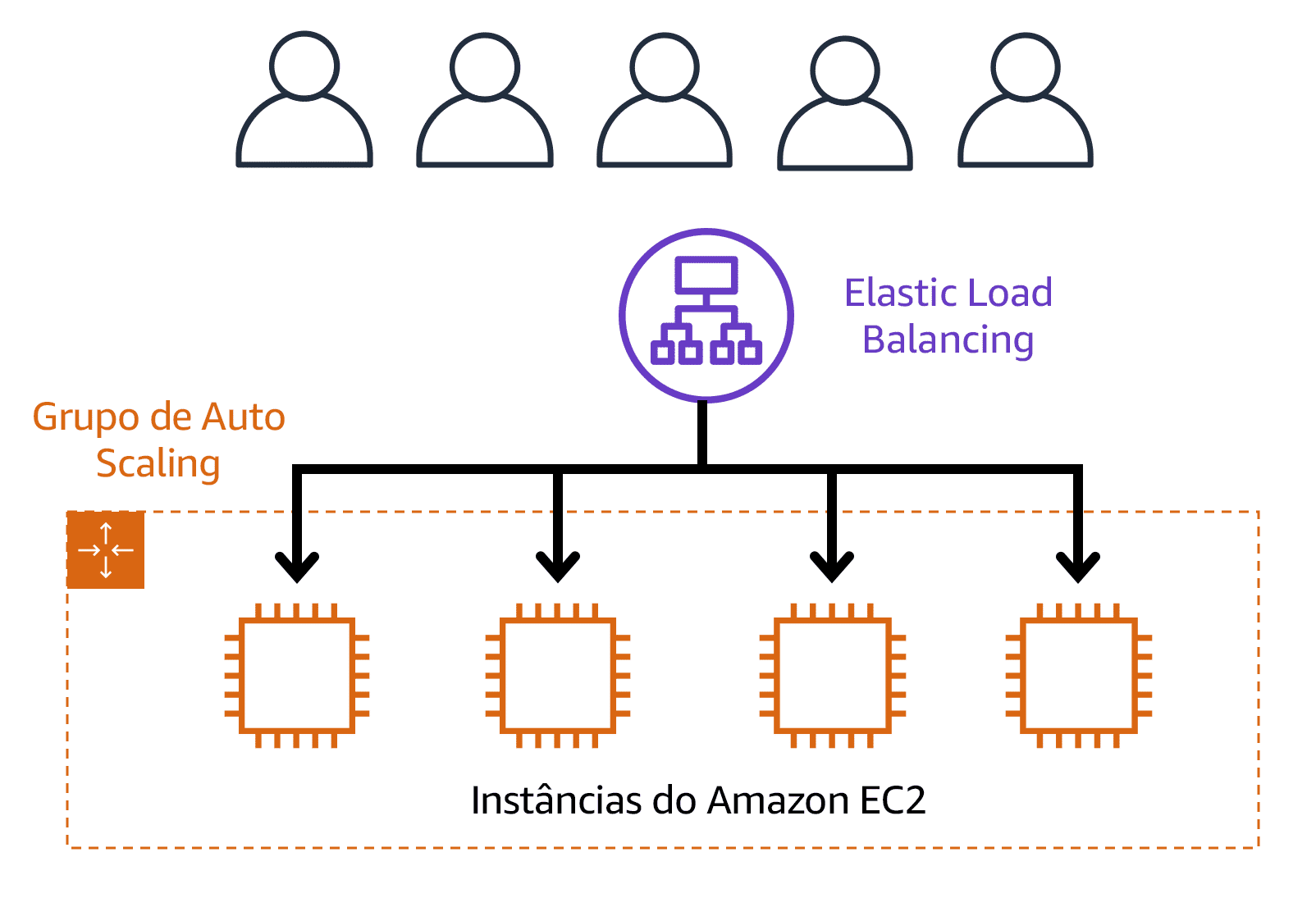
**Exemplo: Elastic Load Balancing**

**Exemplo: Elastic Load Balancing**



**Período de baixa demanda**

Aqui está um exemplo de como o Elastic Load Balancing funciona. Suponha que alguns clientes vieram à cafeteria e estão prontos para fazer seus pedidos.  
  
Se apenas algumas caixas registradoras estiverem abertas, isso corresponde à demanda dos clientes que precisam do serviço. A cafeteria tem menos probabilidade de ter caixas registradoras abertas sem clientes. Nesse exemplo, você pode pensar nas caixas registradoras como instâncias do Amazon EC2.



**Período de alta demanda**

Ao longo do dia, à medida que o número de clientes aumenta, a cafeteria abre mais caixas registradoras para acomodá-los. No diagrama, o grupo do Auto Scaling representa isso.  
  
Além disso, um funcionário da cafeteria direciona os clientes para a caixa registradora mais adequada para que o número de solicitações possa ser distribuído uniformemente entre as caixas abertas. Você pode pensar nesse funcionário da cafeteria como um balanceador de carga.

**Amazon Simple Queue Service (Amazon SQS)**

O **Amazon Simple Queue Service (Amazon SQS)** é um serviço de enfileiramento de mensagens.

Use o Amazon SQS para enviar, armazenar e receber mensagens entre componentes de software, sem perder mensagens ou precisar que outros serviços estejam disponíveis. No Amazon SQS, um aplicativo envia mensagens para uma fila. Um usuário ou serviço recupera uma mensagem da fila, processa-a e a exclui da fila.



**Exercicios**

Você deseja usar uma instância do Amazon EC2 para uma carga de trabalho de processamento em lote. Qual seria o melhor tipo de instância do Amazon EC2 a ser usado?

Uso geral

Otimizada para memória

Otimizada para computação

Otimizada para armazenamento

A resposta correta é **Otimizada para computação**.  
  
As outras respostas estão incorretas porque:

* instâncias de uso geral equilibram os recursos de computação, memória e rede. Essa família de instâncias não seria a melhor escolha para o aplicativo nesse cenário. As instâncias otimizadas para computação são mais adequadas para cargas de trabalho de processamento em lote do que as instâncias de uso geral.
* As instâncias otimizadas para memória são ideais para cargas de trabalho que processam grandes conjuntos de dados na memória, como bancos de dados de alto desempenho.
* As instâncias otimizadas para armazenamento foram projetadas para cargas de trabalho que exigem alto acesso sequencial de leitura e gravação a grandes conjuntos de dados no armazenamento local. A questão não especifica o tamanho dos dados que serão processados. O processamento em lote envolve o processamento de dados em grupos. Uma instância otimizada para computação é ideal para esse tipo de carga de trabalho, que se beneficiaria de um processador de alto desempenho.

**Saiba mais:**

* [Tipos de instâncias do Amazon EC2](https://aws.amazon.com/ec2/instance-types/)

Você tem uma carga de trabalho que será executada por um total de seis meses e consegue suportar interrupções. Qual seria a opção de compra mais econômica do Amazon EC2?

Instância reservada

Instância spot

Instância dedicada

Instância sob demanda

As outras respostas estão incorretas porque:

* Instâncias reservadas exigem um contrato com duração de 1 ou 3 anos. A carga de trabalho neste cenário será executada apenas por 6 meses.
* Instâncias dedicadas são executadas em um Virtual Private Cloud (VPC) em hardware dedicado a um único cliente. Elas têm um custo mais alto do que outras respostas, que são executadas em hardware compartilhado.
* As instâncias sob demanda cumprem os requisitos de execução por apenas seis meses e são resistentes a interrupções. No entanto, uma instância spot seria a melhor escolha porque não requer contrato com duração mínima, consegue suportar interrupções e custa menos do que uma instância sob demanda.

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.  
  
As outras respostas são exemplos de Auto Scaling.

**Saiba mais:**

* [Elastic Load Balancing](https://aws.amazon.com/elasticloadbalancing)
* [Amazon EC2 Auto Scaling](https://aws.amazon.com/ec2/autoscaling)

<https://aws.amazon.com/pt/getting-started/fundamentals-core-concepts/?e=gs2020&p=fundoverview>

https://aws.amazon.com/pt/types-of-cloud-computing/

**Amazon Elastic Compute Cloud (Amazon EC2)**

**Instâncias sob demanda** são ideais para cargas de trabalho irregulares de curto prazo que não podem ser interrompidas.

**Tipos de instância**

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

**Instâncias de uso geral** equilibram os recursos de computação, memória e rede. Você pode usá-las para diversas cargas de trabalho, como:

* servidores de aplicativos
* servidores de jogos
* servidores de back-end para aplicativos empresariais
* bancos de dados pequenos e médios

Suponha que você tenha um aplicativo no qual as necessidades de recursos para computação, memória e rede sejam praticamente equivalentes. Você pode executar esse aplicativo em uma instância de uso geral porque ele não precisa de otimização em nenhuma área de recurso único.

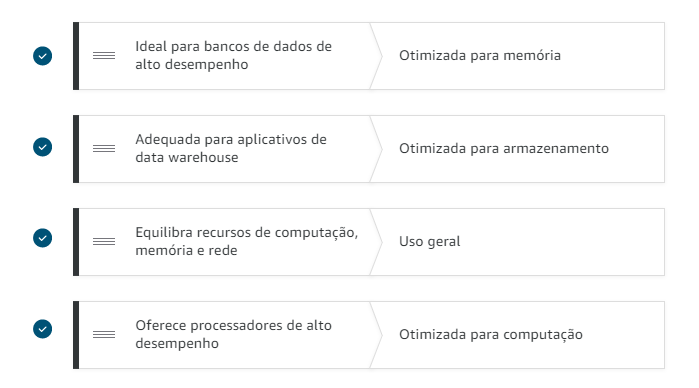
**Instâncias otimizadas para computação** são ideais para aplicativos vinculados à computação que se beneficiam de processadores de alto desempenho. Assim como instâncias de uso geral, você pode usar instâncias otimizadas para computação para cargas de trabalho, como servidores web, de aplicativos e de jogos.  
  
No entanto, a diferença é que aplicativos otimizados para computação são ideais para servidores web de alto desempenho, servidores de aplicativos de computação intensiva e servidores de jogos dedicados. Você também pode usar instâncias otimizadas para computação para cargas de trabalho de processamento em lote, com o processamento de muitas transações em um único grupo.

**Instâncias otimizadas para memória**

são projetadas para fornecer desempenho rápido para cargas de trabalho que processam grandes conjuntos de dados na memória. Na computação, a memória é uma área de armazenamento temporário. Ela contém todos os dados e instruções de que uma unidade central de processamento (CPU) precisa para conseguir realizar ações. Antes que um programa de computador ou aplicativo possa ser executado, ele é carregado do armazenamento para a memória. Esse processo de pré-carregamento dá à CPU acesso direto ao programa de computador.  
  
Suponha que você tenha uma carga de trabalho que precise que grandes quantidades de dados sejam pré-carregados antes de executar um aplicativo. Esse cenário pode ser de um banco de dados de alto desempenho ou uma carga de trabalho que envolva a execução de processamento em tempo real de uma grande quantidade de dados não estruturados. Nesses tipos de casos de uso, considere usar uma instância otimizada para memória. As instâncias otimizadas para memória permitem que você execute cargas de trabalho com altas necessidades de memória e tenha um ótimo desempenho.

**Instâncias de computação acelerada** usam aceleradores de hardware, ou coprocessadores, para executar algumas funções de forma mais eficiente do que é possível em um software executado em CPUs. Exemplos dessas funções são cálculos de números com vírgula flutuante, processamento de gráficos e correspondência de padrões de dados.  
  
Na computação, um acelerador de hardware é um componente que pode agilizar o processamento de dados. As instâncias de computação acelerada são ideais para cargas de trabalho, como aplicativos gráficos e streaming de jogos e de aplicativos.

**As instâncias otimizadas para armazenamento** são projetadas para cargas de trabalho que exigem alto acesso sequencial de leitura e gravação a grandes conjuntos de dados no armazenamento local. Exemplos de cargas de trabalho adequadas para instâncias otimizadas para armazenamento são sistemas de arquivos distribuídos, aplicativos de data warehouse e sistemas de processamento de transações on-line de alta frequência (OLTP).  
  
Na computação, o termo operações de entrada/saída por segundo (IOPS) é uma métrica que mensura o desempenho de um dispositivo de armazenamento. Ela indica quantas operações diferentes de entrada ou saída um dispositivo pode executar em um segundo. As instâncias otimizadas para armazenamento foram projetadas para fornecer dezenas de milhares de IOPS aleatórias e de baixa latência para aplicativos.  
  
Imagine as operações de entrada como dados colocados em um sistema, como registros inseridos em um banco de dados. Uma operação de saída são dados gerados por um servidor. Um exemplo de saída pode ser a análise realizada nos registros em um banco de dados. Se você tiver um aplicativo com alto requisito de IOPS, uma instância otimizada para armazenamento poderá fornecer melhor desempenho em relação a outros tipos de instâncias não otimizados para esse tipo de caso de uso.



**Definição de preço do Amazon EC2**

**Instâncias sob demanda** são ideais para cargas de trabalho irregulares de curto prazo que não podem ser interrompidas. Custos antecipados ou contratos mínimos não se aplicam. As instâncias são executadas continuamente até que sejam interrompidas e você paga apenas pelo tempo de computação que usar.

Exemplos de casos de uso para instâncias sob demanda são desenvolvimento e teste de aplicativos e execução de aplicativos com padrões de uso imprevisíveis. As instâncias sob demanda não são recomendadas para cargas de trabalho que duram um ano ou mais, porque essas cargas de trabalho podem ser mais econômicas usando instâncias reservadas.

**Savings Plans do Amazon EC2** permite reduzir os custos de computação ao haver o compromisso com uma quantidade consistente de uso de computação por um período de um ou três anos. Esse compromisso resulta em economias de até 72% em relação aos custos de instâncias sob demanda.  
  
Qualquer uso até o compromisso é cobrado de acordo com o preço de Savings Plan com desconto (por exemplo, USD 10 por hora). Qualquer uso além do compromisso é cobrado de acordo com os preços normais de instâncias sob demanda.

**Instâncias reservadas** são um desconto de cobrança aplicado ao uso de instâncias sob demanda em sua conta. Você pode adquirir instâncias reservadas comuns e instâncias reservadas conversíveis por um período de um ou três anos, e instâncias reservadas agendadas por um período de um ano. Você tem mais economia com a opção de três anos.  
  
Ao final do período da instância reservada, você pode continuar usando a instância do Amazon EC2 sem interrupção. No entanto, são cobrados os preços de instâncias sob demanda até que um dos procedimentos a seguir seja feito:

* Encerrar a instância.
* Adquirir uma nova instância reservada que corresponda aos atributos da instância (tipo de instância, Região, locação e plataforma).

As **instâncias spot** são ideais para cargas de trabalho com horários de início e término flexíveis ou que toleram interrupções. As instâncias spot usam a capacidade de computação não utilizada do Amazon EC2 e têm uma economia de até 90% de desconto em relação aos preços das instâncias sob demanda.  
  
Suponha que você tenha um trabalho de processamento em segundo plano que pode ser iniciado e interrompido conforme for necessário (como o trabalho de processamento de dados para uma pesquisa de cliente). Você deseja iniciar e interromper o trabalho de processamento sem afetar as operações gerais de seus negócios. Se você fizer uma solicitação spot e a capacidade do Amazon EC2 estiver disponível, a instância spot será iniciada. No entanto, se você fizer uma solicitação spot e a capacidade do Amazon EC2 estiver indisponível, a solicitação não terá sucesso até que a capacidade seja disponibilizada. A capacidade indisponível pode atrasar a execução do trabalho de processamento em segundo plano.

**Hosts dedicados**são servidores físicos com capacidade de instância do Amazon EC2 totalmente dedicada ao uso do cliente.  
  
Você pode usar suas licenças de software por soquete, por núcleo ou por VM existentes para manter a conformidade da licença. Você pode adquirir hosts dedicados sob demanda e reservas de hosts dedicados. De todas as opções do Amazon EC2 que foram abordadas, os hosts dedicados são os mais caros.