amazon web services

Dez dicas para otimização de custos na AWS

Daniel Bento de Paula

Dez dicas para otimização de custos na AWS – parte 1 de 2

Daniel Bento de Paula
Agosto de 2016

Introdução

Otimização de custos é um assunto recorrente entre os clientes da AWS: todos desejam, de alguma maneira, obter uma conta menor no final do mês. Entenda por otimizar custos a redução dos seus gastos de AWS, mas mantendo os níveis de robustez e escalabilidade de suas aplicações, sem que estes sejam impactados.

Este blog post é a primeira de uma série de duas partes, com diversas dicas de fácil implementação que irão ajudá-los a criar e manter um ambiente com custo otimizado na AWS.

As 10 Dicas

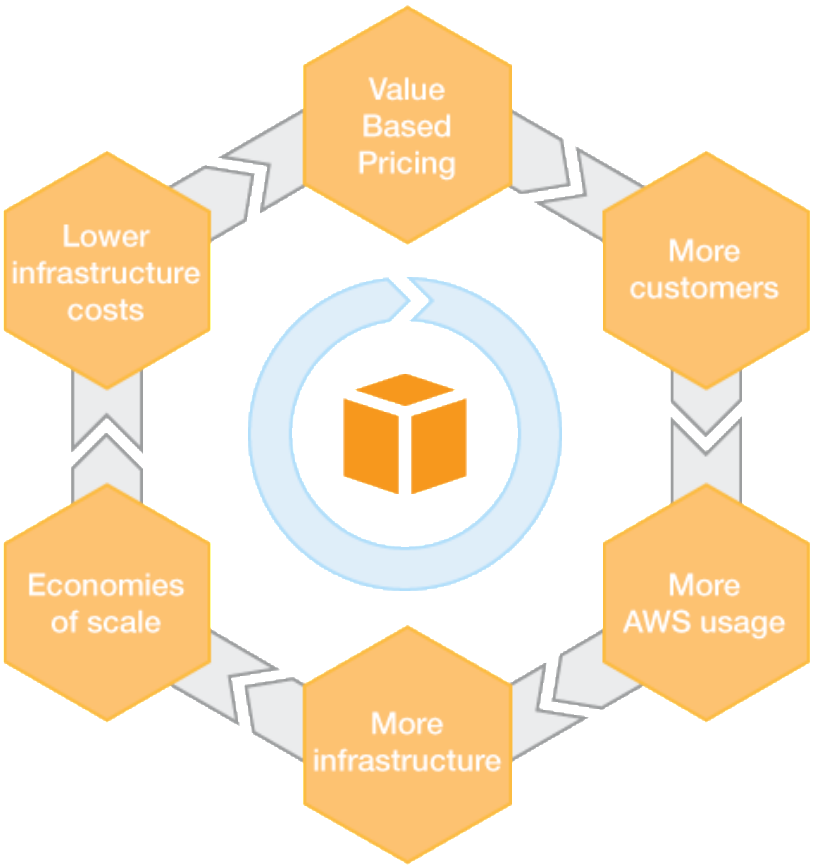
#1 Economize sem fazer esforço

Redução pró-ativa de preços

Desde a sua criação em 2006 até agosto de 2016, ocorreram 52 reduções de preços pró-ativas na AWS. Isso é uma parte normal do nosso negócio e faz parte de nossa filosofia de preço, focada em aumentar a eficiência em nossas operações e passar estas economias obtidas aos nossos clientes.

A filosofia de preços da AWS é baseada em um ciclo virtuoso:

a) os preços baixos, baseados no valor dos serviços, reduzem as barreiras para adoção da computação em nuvem;
b) novos clientes proporcionam um aumento no nível de utilização da AWS;
c) uma maior utilização alavanca os gastos em pesquisa e desenvolvimento, e em conjunto com a economia de escala e melhorias de processos internos, viabiliza a diminuição dos custos de infraestrutura;
d) estas economias são repassadas aos clientes na forma de redução de preços, completando e reiniciando o ciclo virtuoso em (a).



A troca de Capex por Opex – Pague conforme o uso

A AWS não exige comprometimentos de gastos mínimos ou contratos de longo prazo. Você pode trocar grandes capitais de investimento (CAPEX) por pequenos pagamentos variáveis que se aplicam apenas ao que você usar (OPEX). Com AWS você não está preso a acordos de vários anos ou modelos de licenciamento complicados.

A AWS oferece uma simples abordagem de preço, onde você paga apenas pelo que você usa para os mais de 50 serviços. Com a AWS, você paga apenas para os serviços que você precisa, por quanto tempo você usá-los e sem contratos de longo prazo. Todos os preços dos serviços são públicos e estão disponíveis no site da AWS.

Saiba mais sobre os princípios de definição de preços na AWS em https://aws.amazon.com/pt/pricing/.

#2 Remova recursos não utilizados ou ociosos

Desligue instâncias não utilizadas

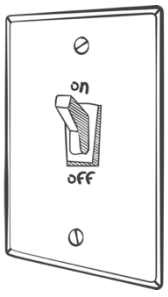
Desligue suas instâncias enquanto não estiverem sendo utilizadas e tome vantagem do modelo de custo onde você paga apenas pelo que usa.

Esta dica é bem aderente para ambientes onde não há a necessidade de ficarem ligados 24 horas nos 7 dias da semana. Ambientes de desenvolvimento, testes, treinamento, etc., geralmente não são utilizados nos períodos noturnos ou finais de semana e podem ser desligados. Uma menor utilização implica diretamente em economia de custos.

Uma semana possui 44 horas úteis/comerciais de um total de 168 horas, representando 26% do tempo da semana, se consideradas 8 horas trabalhadas por dia útil somado com 4 horas do sábado. Desta maneira, um ambiente que deve estar no ar apenas no horário comercial, tem potencial de ficar ligado apenas 26% do tempo em uma semana, representando uma economia de 74% no custo de computação.

Para tal, suas instâncias podem ser iniciadas e desligadas manualmente através da AWS Console ou de maneira automatizada através de scripts ou serviços como AWS Lambda ou AWS Data Pipeline. Você também pode desligar ou subir todos os recursos de seu ambiente utilizando o AWS CloudFormation.

Saiba como parar e iniciar suas instâncias EC2 em horários agendados utilizando o AWS Lambda em
https://aws.amazon.com/premiumsupport/knowledge-center/start-stop-lambda-cloudwatch/
Ou através do AWS Data Pipeline: https://aws.amazon.com/premiumsupport/knowledge-center/stop-start-ec2-instances/



Remova os demais recursos ociosos ou de baixa utilização

A dicas acima não se aplicam somente a instâncias EC2. É comum o cenário onde clientes não verificam a utilização e ficam com diversos recursos ociosos em seu ambiente na AWS. O esquecimento destes recursos geram gastos desnecessários. Veja abaixo uma lista com os principais recursos esquecidos:

- Baixa utilização de instâncias EC2: desligue-as ou faça a consolidação das aplicações em uma única instância;
- Elastic Load Balancing ociosos, que não possuem alguma instância de backend associada. Se o ELB não recebe e distribui tráfego, então pode ser desligado;
- Volumes EBS que não estão ligados a alguma instância ou estão subutilizados. Ao invés de deixar um volume EBS ocioso, caso queira recuperá-lo futuramente, tire e armazene o snapshot deste volume, que possui custo inferior;
- Elastic IPs não associados a instâncias, sendo cobrados por esse período;
- Conexões VPN não utilizadas;
- NAT Gateways que não recebem tráfego;
- Instâncias RDS subutilizadas.

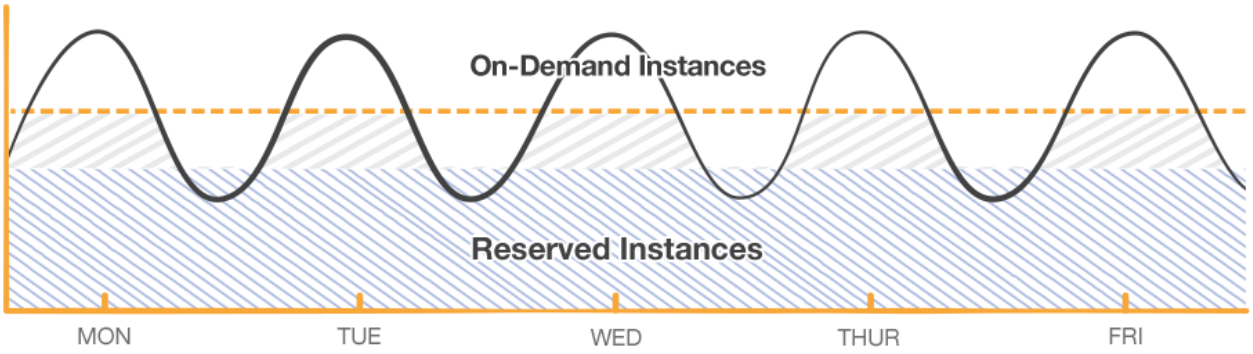
Para monitorar o consumo dos recursos de seu ambiente utilize o AWS CloudWatch. Faça a análise de diferentes métricas, de consumo atual e passado através dos dashboards, e detecte possíveis ociosidades. Saiba mais em https://aws.amazon.com/pt/cloudwatch/.

#3 Utilize instâncias reservadas

Com o objetivo de diminuir os custos nos cenários onde há uma previsibilidade da utilização de instâncias, a AWS criou o modelo de instâncias reservadas. Nele é possível realizar a reserva de uma instância, com comprometimento de utilização de 1 ou 3 anos e receber descontos de até 75% se comparados aos custos do modelo on-demand.

Alguns clientes criaram regras de decisão de quanto se reservar uma instância. Por exemplo: “Se determinada instância ficou ligada 100% do tempo a mais de duas semanas e não pode ser desligada neste exato momento, então ela tem potencial para ser reservada”.

Ambientes podem ser compostos tanto por instâncias reservadas como por instâncias on-demand. Tome como exemplo um e-commerce tradicional, que possui picos de acesso durante o dia. O baseline de utilização poderia ser suportado por instâncias reservadas. Já os picos de utilização poderiam ser suportados por instâncias on-demand.



Saiba como as instâncias reservadas funcionam em:
https://aws.amazon.com/premiumsupport/knowledge-center/ec2-ri-basics/

#4 Utilize instâncias Spot

Economize até 90% dos custos de EC2 utilizando instâncias Spot. Com o objetivo de aproveitar a capacidade ociosa de infraestrutura dos datacenters, a AWS criou um modelo de preço dinâmico, chamado instâncias Spot. O preço das instâncias spot é baseado na oferta e demanda desta infraestrutura e oscila analogamente a uma bolsa de valores.

Funciona da seguinte maneira: você define o preço máximo por hora que está disposto a pagar pela instância. Sua instância será inicializada caso seu preço escolhido seja maior que o preço de mercado das instâncias Spot. O preço do mercado é dinâmico e oscila conforme outros clientes solicitam instâncias Spot.

Instâncias Spot tradicional

Para o modelo de instâncias Spot tradicional, caso o preço de mercado suba e fique maior que o seu preço inicial, sua instância receberá uma notificação e será finalizada dentro de dois minutos. Como existem riscos associados ao desligamentos das instâncias, é desejável que suas aplicações sejam planejadas a serem tolerantes a estes tipos de eventos.

Existem alguns tipos aplicações que se beneficiam mais facilmente de instâncias Spot. Por exemplo:

- Aplicações stateless;
- Aplicações com AutoScaling, com a criação de dois grupos: um com instâncias Spot e outro com instâncias on-demand;
- Processamento Batch;
- Processamento utilizando Amazon Elastic MapReduce;
- Servidores de Integração contínua (CI);
- High performance computing (HPC);
- Renderização e transcoding de mídia.

Spot Blocks

Spot Blocks são instâncias spots que irão rodar continuamente por um período finito, que pode ser definido de 1 a 6 horas de duração. O preço é baseado no período solicitado e a capacidade disponível, e é geralmente 30% a 45% menor do que o modelo on-demand.

Você submete uma requisição e define a quantidade de horas que você quer que suas instâncias rodem, juntamente com o preço máximo que está disposto a pagar. Quando existe capacidade disponível para o período solicitado, suas instâncias serão iniciadas e rodarão de forma contínua a um preço fixo. Elas serão encerradas automaticamente no final do período.

Este modelo é indicado para situações onde você tem jobs que precisam rodar continuamente por até 6 horas.

Saiba mais em https://aws.amazon.com/pt/ec2/spot/

#5 Escalando seus recursos

Escalando verticalmente

Escalar verticalmente significa acrescentar ou reduzir capacidade de um recurso em um mesmo nó e geralmente está relacionado a alterar o número de vCPUs, memória, storage, rede, etc. de uma instância.

A estratégia de otimização de custo através de escalabilidade vertical consiste em aumentar os recursos somente para lidar com um pico de consumo e reduzi-los conforme a diminuição da demanda. Assim, evita-se o super provisionamento de um recurso pelo pior cenário, que estaria ocioso maior parte do tempo, e utiliza-se recursos menores, de menor custo, com capacidade bem justa, para atender o baseline da aplicação.

Tome como exemplo uma aplicação de recursos humanos que necessita calcular a folha de pagamento dos funcionário no final de cada mês. Usualmente um processo desses implica em um pico de processamento que pode durar algumas horas. Uma estratégia consistiria em aumentar o tamanho das instâncias de servidores de aplicação e banco de dados para lidar com cálculo e, após o término, voltar ao tamanho original.

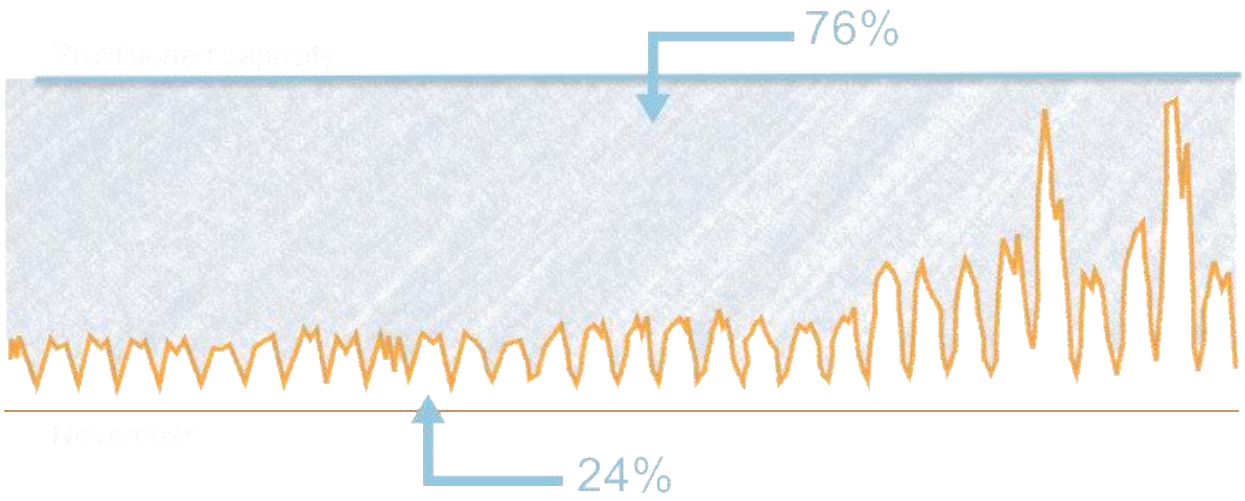
Saiba alterar o tamanho de sua instância EC2 em
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-resize.html

AutoScaling – Escalando horizontalmente de maneira automática

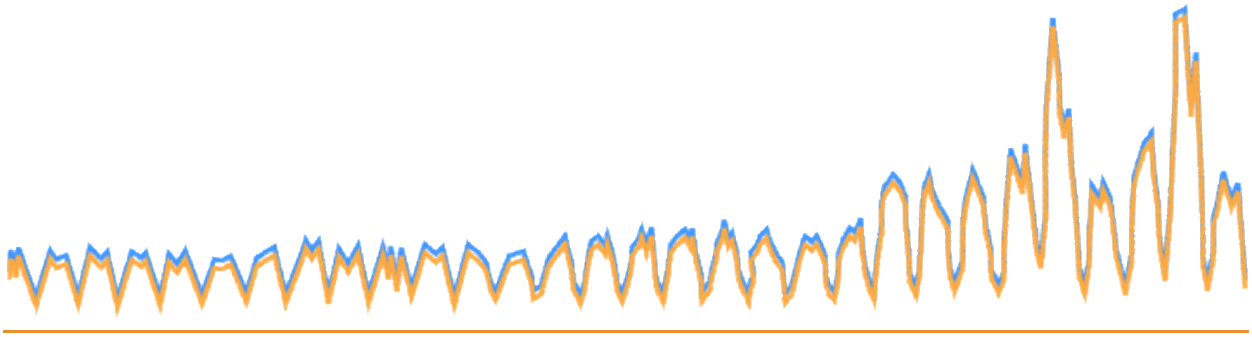
Escalar horizontalmente significa acrescentar ou reduzir o número de nós em um conjunto de servidores para lidar com a variação de demanda. AutoScaling é o serviço da AWS que realiza essa tarefa de maneira automática.

Assim, não há a necessidade de se provisionar a quantidade de instâncias baseadas no cenário de pico de acesso. Como boa prática, é mantido um número mínimo de instâncias para lidar com o baseline de acessos de sua aplicação e, para lidar com os picos, novas instâncias são lançadas através do AutoScaling.

O gráfico abaixo mostra o potencial de desperdício de infraestrutura em um cenário tradicional de e-commerce, 76% da capacidade computacional é desperdiçada, enquanto somente 24% em media é utilizado durante o mês.



O gráfico abaixo mostra como o AutoScaling ajusta o provisionamento de infraestrutura (em azul) em função da demanda (em amarelo). Note que há pouco desperdício de infraestrutura, implicando em economia e diminuição de custos.



Saiba mais sobre o AutoScaling em: https://aws.amazon.com/pt/autoscaling/

Na parte dois desta série continuaremos com as 5 dicas finais de otimização de custos na AWS.

#6 Alinhe o provisionamento de recursos com a demanda
#7 Aproveite os diferentes tipos de storage
#8 Faça o Offload de sua arquitetura
#9 Aproveite os serviços gerenciados
#10 Utilize as ferramentas de monitoramento e análise de custos