
CLOUD
COMPUTING

Texto base

11

Alta Disponibilidade - Parte 1

Rodolfo Riyoei Goya

Resumo

A Computação em Nuvem é comumente aplicada em negócios "Missão Crítica", ou seja, de grande importantes, que exigem muita confiabilidade, e com requisitos de operação correta e de alta disponibilidade. Em ambientes complexos, formados por grande quantidade de componentes, é importante que haja recursos para repassar as tarefas de componentes que falham para outros componentes de forma imediata, imperceptível para o usuário, automatizada e, mesmo assim, sem que os custos cresçam exageradamente. Neste texto, é abordada a importância dos serviços de distribuição e balanço de carga de tarefas para a realização de alta disponibilidade e dos serviços de auto scaling automático.

11.1. Introdução

A forma mais comum de proporcionar alta disponibilidade é distribuir a demanda de serviços igualmente entre diversos componentes, divididos em grupos em diferentes zonas de disponibilidade, com o cuidado de monitorar e substituir componentes com falhas de modo transparente para o usuário. Além disso, gerenciar o dimensionamento dos grupos de modo a proporcionar elasticidade e escalabilidade também são objetivos a serem atingidos.

11.1.1. Serviços de alta disponibilidade e escalabilidade

A AWS oferece serviços para formar grupos de elasticidade com escalabilidade horizontal com o Auto Scale Group – AWS. Monitoração com métricas para ajustes de dimensionamento com o CloudWatch e distribuição de serviços com monitoração da saúde dos componentes com o Elastic Load Balance – ELB. Com estes serviços, podem-se criar e gerenciar grupos de instâncias EC2 idênticas produzidas a partir de um template e que são lançadas e terminadas conforme a necessidade, bem como configurar políticas com as quais estes grupos vão operar.

11.2. Balanço de carga

O serviço de balanço de carga opera na frente de um grupo de máquinas virtuais recebendo solicitações que são distribuídas entre elas (por exemplo, solicitações de acesso a páginas web entre diversos servidores), evitando que componentes de um grupo fiquem sobrecarregados enquanto outros fiquem ociosos.

11.2.1. AWS Elastic Load Balance

A AWS oferece balanço de carga através do AWS Elastic Load Balance – ELB. O ELB é um serviço gerenciado de modo transparente para o usuário, com alta disponibilidade, escalabilidade da infraestrutura, atualizações, patches e service packs feitos pela AWS.

O ELB recebe solicitações de serviços e as distribui dentro de um grupo de instâncias EC2. Estas instâncias podem ser de tipos diferentes, estar em mais de uma sub-rede e podem estar em diferentes zonas de disponibilidade (o que é uma boa prática para melhor tolerância a falhas).

Os dois modelos de ELB mais importante que a AWS oferece são:

a. Application Load Balance: Adequado para distribuir tráfego entre servidores web (HTTP: TCP port 80 e HTTPS: TCP port 443) para instâncias EC3.
b. Network Load Balance: Adequado para distribuir tráfego IP para serviços que não são web (TCP, UDP e ICMP).

O processo de distribuição de tráfego inclui:

a. Testes de integridade: troca de mensagens com as instâncias (ICMP para network load balance ou HTTP com resposta HTTP 200 OK para application load balance). Instâncias identificadas como funcionando mal não recebem tráfego e produzem notificações pelo ELB.
b. Afinidade: Pode-se configurar para que solicitações de mesma origem tenham preferência para ser enviada e processada em um mesmo servidor de destino.

Por ficar à frente das instâncias, o ELB apresenta um endereço IP único para o cliente final (protegendo as instâncias em endereços IP privados) e inclui serviços de segurança (como a proteção para ataque de negação de acesso) que aliviam o usuário de ter que executar esta tarefa no gerenciamento das instâncias.

O ELB também pode ser usado para armazenar o certificado digital de um site e terminar conexões SSL/TLS em sites com comunicação HTTPS com segurança com criptografia e autenticação.

11.3. Automação de escalabilidade

O uso de monitoração para ajuste na quantidade de instâncias que compõem um grupo permite escalabilidade horizontal automática. Por exemplo, um serviço pode ser configurado para iniciar uma nova instância de processamento caso a porcentagem de uso de CPU ultrapasse 80% durante um período de pelo menos 3 minutos (e termine uma instância caso essa mesma porcentagem fique abaixo de 20% por mais de 3 minutos).

11.3.1. AWS Auto Scale Group

A AWS oferece escalabilidade horizontal através da criação de Auto Scale Groups – ASG: serviço de gerenciamento de grupos de instâncias idênticas (criadas a partir de uma mesma imagem). O serviço de gerenciamento é executado em infraestrutura de alta disponibilidade e escalabilidade, com atualizações, patches e service packs gerenciados pela AWS de modo transparente para o usuário.

Quando o ASG recebe eventos entregues pelo CloudWatch indicando que o número de instâncias em um grupo não é suficiente para atender a demanda por serviços, ele lança uma ou mais novas instâncias no grupo e, do mesmo modo, se a monitoração do CloudWatch indicar que as instâncias estão com baixo nível de uso, uma ou mais instâncias são terminadas.

A configuração do ASG deve especificar:

a. Imagem da instância usada como template de replicação para as instâncias do grupo.
b. Número mínimo e máximo (para evitar que algum ataque de negação de acesso por acessos massivos cause um custoso uso excessivo) de instâncias.
c. Valores de métrica para iniciar (por exemplo, uso de CPU das instâncias acima de 80%) e encerrar (por exemplo, uso de CPU das instâncias abaixo de 20%) instâncias.
d. Política de contratação da instância (pode minimizar custos escolhendo instâncias "spot" contratadas a preço de mercado ou instâncias reservadas com pagamento por demanda).
e. Política de término de instância (pode minimizar custos escolhendo instâncias que faltam menor tempo para completar horas cheias – o tempo de uso cobrado é arredondado para a hora inteira seguinte).

Os ASG normalmente trabalham conjuntamente com serviços de balanço de carga (como o Elastic Load Balance – ELB da AWS), de modo que as instâncias de um mesmo grupo dividam a carga de trabalho.

O serviço ASG inclui testes de integridade das instâncias como parte do processo de seleção de instâncias a remover.

Atualizações, com patches e service packs das instâncias, podem ser feitas atualizando a imagem de template de lançamento. O ASG promove a atualização pela remoção das instâncias não atualizadas e lançamento de instâncias com templates atualizados.

11.4. Vamos praticar?

10.4.1. Faça o seu site com grupo de auto scale usando ASG na AWS

Faça do seu servidor Apache de uma instância EC2 um template para um grupo de auto scale. Use este template para o seu site na Internet operar em um ASG e se dimensionar automaticamente com a demanda. Veja um roteiro completo de configuração no link: https://docs.aws.amazon.com/autoscaling/ec2/userguide/GettingStartedTutorial.html

11.4.2. Coloque o seu site em um grupo de servidores com ELB na AWS

Faça um grupo com vários servidores Apache idênticos em instâncias EC2 com balanço de carga por ELB. Veja mais detalhes sobre como esta configuração pode ser feita nos links:
https://docs.aws.amazon.com/elasticloadbalancing/latest/application/application-load-balancer-getting-started.html;
https://docs.aws.amazon.com/pt_br/elasticloadbalancing/latest/application/create-application-load-balancer.html

11.5. Você quer ler?

11.5.1. Como monitorar o Elastic Load Balance?

O ELB gera diversas métricas de monitoração que podem ser capturadas e exibidas pelo CloudWatch. Veja mais detalhes com exemplos de configuração no link:
https://docs.aws.amazon.com/pt_br/elasticloadbalancing/latest/application/load-balancer-cloudwatch-metrics.html

11.5.2. Elastic Load Balance para aplicações

O ELB, além de distribuir solicitações HTTP/HTTPS, no modo de Application Load Balance, também pode ser utilizado para distribuição de carga de processamento entre microsserviços baseadas em containers. Veja mais detalhes no link:
https://docs.aws.amazon.com/pt_br/AmazonECS/latest/developerguide/service-load-balancing.html

Referências

TAURION, Cezar. **Cloud Computing:** computação em nuvem: transformando o mundo da tecnologia da informação. Rio de Janeiro: Brasport, 2009.

VELTE, Anthony T.; VELTE, Toby J.; ELSENPRETER, Robert. **Cloud Computing:** a practical approach. EUA:McGraw-Hill, 2011.

MARSHALL, Nick; BROWN, Mike; BLAIR FRITZ, G.; JOHNSON, Ryan. **Mastering VMware vSphere 6.7.** New Jersey: Sybex, 2019. 848p.

SANTOS, Tiago. **Fundamentos da computação em nuvem.** São Paulo: Editora Senac, 2018. 211p. (Série Universitária).

ANDREWS, Joshua; HALL, Jon. **VMware Certified Professional Data Center Virtualization on vSphere 6.7 Study Guide:** Exam 2V0-21.19. New Jersey: Sybex, 2020. 640p.

Official Amazon Web Services (AWS) Documentation. **Elastic Load Balancing User Guide. Amazon.** 33p. Disponível em: <https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/elb-ug.pdf>. Acesso em: 17 mar. 2022.

Official Amazon Web Services (AWS) Documentation. **Amazon EC2 Auto Scaling User Guide.** Amazon. 368p. Disponível em: <https://docs.aws.amazon.com/autoscaling/ec2/userguide/as-dg.pdf>. Acesso em: 17 mar. 2022.

Official Amazon Web Services (AWS) Documentation. **Amazon CloudWatch User Guide.** Amazon. 985p. Disponível em: <https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/acw-ug.pdf>. Acesso em: 17 mar. 2022.