
CLOUD
COMPUTING

Texto base

12

Alta Disponibilidade - Parte 2

Rodolfo Riyoei Goya

Resumo

Com a universalização da Internet e a "Computação em Nuvem", serviços oferecidos através passaram a estar disponíveis para o mundo todo. Como isso, conteúdo específico para cada localização (idioma, moeda, conjunto de caracteres etc.) ou distância entre servidores e clientes passaram a ser relevantes. Abordam-se aqui, como se lidam com tais questões nas implementações reais de nuvens através de exemplos na nuvem da AWS.

12.1. Introdução

Quando se produz uma aplicação disponibilizada através da Internet, ela pode ser usada em diferentes países. Por isso, diferentes versões podem ser disponibilizadas. Como dar tolerância a falhas nesta situação? Como o DNS escolhe o endereço da aplicação a resolver se há diferentes versões da aplicação? Como evitar que clientes tenham respostas ruins por se conectarem a servidores muito distantes? No caso da AWS, isto implica em instalações em múltiplas regiões.

12.1.1. Internacionalização

Ao se produzir uma aplicação para disponibilizar pela Internet, é preciso se levar em conta que os clientes podem estar em qualquer lugar do mundo. Assim, além do idioma do conteúdo, o uso de moeda local, a obediência a legislação, a logística de estoques e entregas de produtos pode ser peculiar de cada local. Por isso, o conteúdo deve ser específico para os locais atendidos.

Um cliente de um local distante pode ser atendido por um servidor muito distante e isso leva a um atraso de comunicação elevado o que resulta em uma experiência ruim.

Para dar conta destes requisitos, é possível implantar diversas versões da aplicação adaptadas para diferentes clientes em regiões distintas e que direcione o cliente para a versão da aplicação que lhe é mais adequada.

12.1.2. Elasticidade e tolerância a falhas

Ao se produzir uma aplicação com múltiplas versões localizadas pode não ser econômico criar várias implantações com escalabilidade horizontal. Assim, situações de falha ou queda de desempenho em uma região podem ser transbordadas para outras regiões (o atendimento em caso de queda ou congestionamento no site em holandês pode ser redirecionado para o site em inglês).

12.2. Domain Name System

O Domain Name System – DNS é o sistema de registro e resolução de nomes na Internet. Suas finalidades principais são a de gerenciar os nomes (de modo a não haver dúvidas sobre quais nomes estão ocupados por quem ou estão livres) e resolver nomes em endereços IP.

Quando um nome associado a uma aplicação pode ser resolvido em diversos endereços IP diferentes, os padrões do DNS permitem diversas soluções, por exemplo um rodízio de endereços (denominado “round-robin” que resulta em distribuição da carga entre estes endereços) ou entregar todos os endereços para o cliente e delegar para ele a decisão de qual endereço usar). Tais soluções são usadas para proporcionar balanço de carga e failover usando múltiplos servidores em endereços IP diferentes.

12.2.1. Route 53 e DNS na AWS

O serviço de DNS da AWS é chamado Route 53. Ele é totalmente gerenciado pela AWS com alta disponibilidade em todas as regiões, escalabilidade e tolerância a falhas. Ele faz parte do registro de nomes e é integrado ao DNS mundial, permitindo verificar se um nome já está ocupado na Internet, resolver qual o endereço IP correspondente a esse nome e registrá-lo caso seja desejado.

O Route 53 pode ser configurado para resolver solicitações de nomes registrados nele para endereços IPv4 e IPv6 tanto internos da nuvem como vindos de usuários externos na Internet.

Quando o nome corresponde a diversos possíveis servidores, a escolha do endereço pode ser feita usando diversas políticas:

a. Round-robin: A cada consulta, um dos endereços é entregue, de modo a promover um rodízio entre os servidores.
b. Round-robin ponderado: A cada consulta, um dos endereços é entregue. Mas pode-se controlar a porcentagem das respostas para cada servidor. É útil quando se inclui um servidor com uma versão para teste para receber uma pequena porcentagem de tráfego.
c. Múltiplos valores: A consulta retorna todos os endereços registrados no DNS para o cliente, que decide qual deles usar.

d. Fail over: A consulta retorna o endereço do servidor principal da aplicação. O Route 53 testa periodicamente a integridade deste servidor e, em caso de queda, passa a entregar o endereço de um servidor alternativo (que frequentemente é apenas o endereço de uma página estática informando que o serviço está fora do ar).
e. Geolocalização: A consulta retorna um endereço escolhido baseado na localização da origem da consulta. Por exemplo, se a consulta vem de um país de língua espanhola escolhe um endereço de um servidor com conteúdo nessa língua.
f. Geo-proximidade: A consulta retorna o endereço do servidor mais próximo da localização da origem da consulta.
g. Latência: A consulta retorna o endereço do servidor com menor latência de acesso.

12.3. Aceleradores de conteúdo

Uma diferença importante entre colocar uma aplicação em infraestrutura on premises e colocar em um provedor de serviços em nuvem é que os provedores têm PRESENÇA GLOBAL, ou seja, possuem pontos de presença em locais espalhados pelo mundo.

Isto tem grande impacto quando seus clientes estão espalhados pelo mundo: quando a comunicação entre o cliente e o servidor tem que percorrer uma grande distância, o atraso resulta em desempenho ruim.

Quando se instala uma aplicação em um provedor de nuvem, o cliente pode ser atendido entregando a ele uma cópia do conteúdo a partir de um ponto de presença próximo a ele.

O serviço de aceleração de conteúdo funciona instalando-se múltiplas cópias de conteúdo em diversos pontos de presença. Como resultado disso, há um serviço de melhor qualidade, um tráfego menor nos servidores e uma proteção maior para ataques de negação de acesso (uma vez que um agressor precisa atingir um número bem maior de provedores de conteúdo).

12.3.1. CloudFront

O produto da AWS para aceleração de entrega de conteúdo estático e dinâmico (arquivos html, css, imagens e código Javascript do lado do cliente) é o CloudFront. Quando uma solicitação de um cliente é recebida por um servidor, ela é redirecionada para o ponto de presença mais próximo do cliente de modo a minimizar a latência. Se o conteúdo existir no ponto de presença, ele é entregue ao cliente. Se não existir, uma cópia será enviada pelo servidor para atender esta solicitação e as solicitações futuras.

O tempo para permanência das cópias nos pontos de presença é, por default, 24 horas podendo ser configurado.

O Cloudfront suporta a entrega de certificado para o protocolo SSL/TLS e podendo processar conexões HTTPS criptografadas com os clientes.

Além da aceleração na entrega, o Cloudfront pode ser usado para executar código Lambda em pontos de presença remotos e para coletar e fazer upload de dados enviados de usuários para aplicações. Para os dois casos, o atendimento é mais próximo dos usuários e com tempo de resposta melhor.

12.4. Vamos praticar?

12.4.1. Faça o seu site estático na AWS ter alto desempenho no mundo inteiro

Coloque um site no ar e acelere a entrega de seu conteúdo com o CloudFront. Veja mais detalhes no link:
https://docs.aws.amazon.com/pt_br/Route53/latest/DeveloperGuide/getting-started-cloudfront-overview.html

12.4.2. Configure o nome do seu site estático na AWS

Configure o Route 53 para registrar e resolver o nome que você desejar para o seu site. Veja mais detalhes no link:
https://docs.aws.amazon.com/pt_br/Route53/latest/DeveloperGuide/getting-started-cloudfront-overview.html

12.5. Você quer ler?

12.5.1. Quer saber mais sobre o AWS CloudFront?

O serviço CloudFront da AWS cria cópias de conteúdo (cache) nos pontos de presença, onde houve solicitação por usuários, para acelerar a entrega de conteúdo com melhor desempenho e reduzir a carga nos serviços de nuvem. Veja mais detalhes no link:
https://docs.aws.amazon.com/pt_br/AmazonCloudFront/latest/DeveloperGuide/Introduction.html

12.5.2. Como prover tolerância a falhas com o AWS Route 53?

O serviço Route 53 AWS implementa mecanismos de teste que permitem redirecionar acessos em caso de falha. Veja mais detalhes nos links:
https://docs.aws.amazon.com/pt_br/Route53/latest/DeveloperGuide/disaster-recovery-resiliency.html
https://docs.aws.amazon.com/pt_br/Route53/latest/DeveloperGuide/monitoring-health-checks.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, 2012.

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. **Amazon Route 53 Developer Guide.** Amazon. 701p. Disponível em: <https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/route53-dg.pdf>. Acesso em: 16 mar. 2022.

Official Amazon Web Services (AWS) Documentation. **Amazon CloudFront Developer Guide. Amazon.** 563. Disponível em: <https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/AmazonCloudFront_DevGuide.pdf>. Acesso em: 16 mar. 2022.