Hospedagem de aplicativos web na Nuvem AWS

Setembro de 2017

Avisos

Este documento é fornecido apenas para fins informativos. Ele relaciona as atuais ofertas de produtos e práticas da AWS na data de emissão deste documento, que estão sujeitas a alterações sem aviso prévio. Os clientes são responsáveis por fazer sua própria avaliação independente das informações neste documento e de qualquer uso dos produtos ou serviços da AWS, cada um dos quais é fornecido “como está”, sem garantia de qualquer tipo, expressa ou implícita. Este documento não cria quaisquer garantias, representações, compromissos contratuais, condições ou promessas da AWS, suas afiliadas, fornecedores ou licenciadores. As responsabilidades e obrigações da AWS para com seus clientes são controladas por contratos da AWS, e este documento não modifica nem faz parte de qualquer contrato entre a AWS e seus clientes.

Sumário

Resumo

A hospedagem na web altamente disponível e escalável pode ser uma proposta complexa e cara. As arquiteturas escaláveis tradicionais da web não precisaram apenas implementar soluções complexas para garantir altos níveis de confiabilidade, mas também exigiram uma previsão precisa do tráfego para fornecer um alto nível de atendimento ao cliente. Períodos densos de picos de tráfego e variações bruscas nos padrões de tráfego resultam em taxas de utilização baixas de hardware caro. Isso gera altos custos operacionais para manter o hardware inativo e um uso ineficiente de capital para hardware subutilizado.

O Amazon Web Services (AWS) fornece uma infraestrutura confiável, escalável, segura e de alto desempenho para os aplicativos web mais exigentes. Essa infraestrutura corresponde aos custos de TI com os padrões de tráfego do cliente em tempo real.

Este whitepaper é para gerentes de TI e arquitetos de sistemas que buscam a nuvem para ajudá-los a alcançar a escalabilidade para atender às necessidades de computação sob demanda.

Uma visão geral da hospedagem tradicional na web

Hospedagem na web escalável é um espaço de problema bem conhecido. A Figura 1 descreve uma arquitetura de hospedagem tradicional na web que implementa um modelo de aplicativo web de três camadas comum. Nesse modelo, a arquitetura é separada em camadas de apresentação, aplicação e persistência. Escalabilidade é fornecida pela adição de hosts nessas camadas. A arquitetura também possui recursos integrados de desempenho, failover e disponibilidade. A arquitetura de hospedagem tradicional na web é facilmente transferida para a Nuvem AWS com apenas algumas modificações.

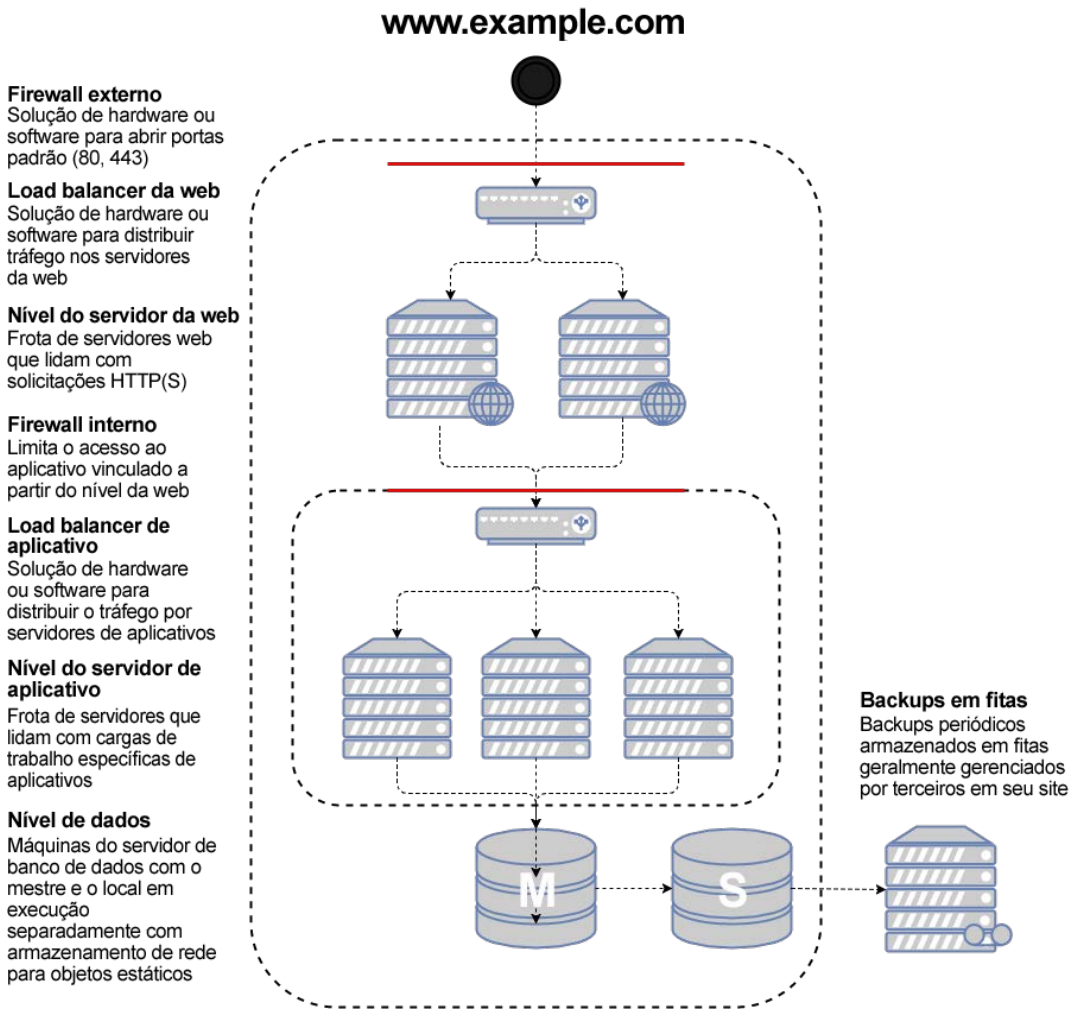

Figura 1. Uma arquitetura tradicional de hospedagem na web

Nas seções a seguir, veremos por que e como essa arquitetura deve ser e pode ser implantada na nuvem da AWS.

Hospedagem de aplicativos web na nuvem da AWS

A primeira pergunta que deve ser sobre o valor de mover uma solução clássica de hospedagem de aplicativos web para a Nuvem AWS. Se decidir que a nuvem é ideal, você precisará de uma arquitetura adequada. Esta seção ajuda você a avaliar uma solução da Nuvem AWS. Ele compara a implantação de seu aplicativo web na nuvem em uma implantação local, apresenta uma arquitetura de Nuvem AWS para hospedar seu aplicativo e discute os principais componentes dessa solução.

Como a AWS pode resolver problemas comuns de hospedagem de aplicativos web

Se for responsável pela execução de um aplicativo web, você enfrenta uma variedade de problemas de infraestrutura e arquitetura para os quais a AWS pode fornecer soluções transparentes e econômicas. A seguir, alguns dos benefícios de usar o AWS em um modelo de hospedagem tradicional.

Uma alternativa econômica para as frotas extragrandes necessárias para lidar com os picos

No modelo de hospedagem tradicional, você precisa provisionar servidores para lidar com a capacidade de pico. Ciclos não utilizados são desperdiçados fora dos períodos de pico. Os aplicativos web hospedados pela AWS podem aproveitar o provisionamento sob demanda de servidores adicionais, para que você possa ajustar constantemente a capacidade e os custos aos padrões reais de tráfego.

Por exemplo, o gráfico a seguir mostra um aplicativo web com um pico de uso das 9h às 15h e menos uso para o restante do dia. Uma abordagem de escalabilidade automática baseada nas tendências reais de tráfego, que provisiona recursos somente quando necessário, resultaria em menos capacidade desperdiçada e uma redução de custo maior que 50%.

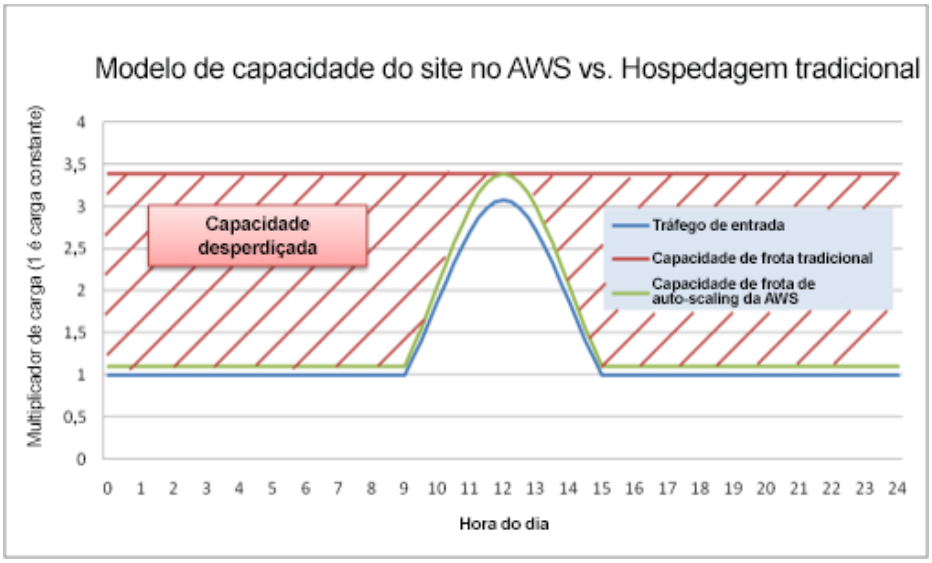

Figura 2. Um exemplo de capacidade desperdiçada em um modelo de hospedagem clássico

Uma solução escalável para lidar com picos de tráfego inesperados

Uma consequência ainda mais grave do provisionamento lento associado a um modelo de hospedagem tradicional é a incapacidade de responder a tempo aos picos de tráfego inesperados. Há muitas histórias sobre aplicativos web sendo desativados devido a um pico inesperado no tráfego depois que o site é mencionado na mídia popular. O mesmo recurso sob demanda que ajuda os aplicativos web dimensionados para corresponder a picos de tráfego regulares também pode lidar com uma carga inesperada. Novos hosts podem ser lançados e ficam prontos em questão de minutos, e podem ser colocados off-line com a mesma rapidez quando o tráfego retorna ao normal.

Uma solução sob demanda para ambientes de teste, carga, beta e pré-produção

Os custos de hardware para criar um ambiente de hospedagem tradicional para um aplicativo web de produção não param na frota de produção. Frequentemente, é necessário criar frotas de pré-produção, beta e teste para garantir a qualidade do aplicativo web em cada estágio do ciclo de vida de desenvolvimento. Embora você possa fazer várias otimizações para garantir o melhor uso possível desse hardware de teste, essas frotas paralelas nem sempre são usadas da maneira ideal: uma grande quantidade de hardware caro não é usada por longos períodos de tempo.

Na Nuvem AWS, é possível provisionar frotas de teste conforme necessário. Além disso, você pode simular o tráfego de usuários na Nuvem AWS durante o teste de carga. Você também pode usar essas frotas paralelas como um ambiente de preparação para uma nova versão de produção. Isso permite a troca rápida da produção atual para uma nova versão do aplicativo com pouca ou nenhuma interrupção de serviço.

Uma arquitetura de Nuvem AWS para hospedagem na web

A figura a seguir fornece outra visão da arquitetura clássica de aplicativos web e como ela pode aproveitar a infraestrutura de computação em Nuvem AWS.

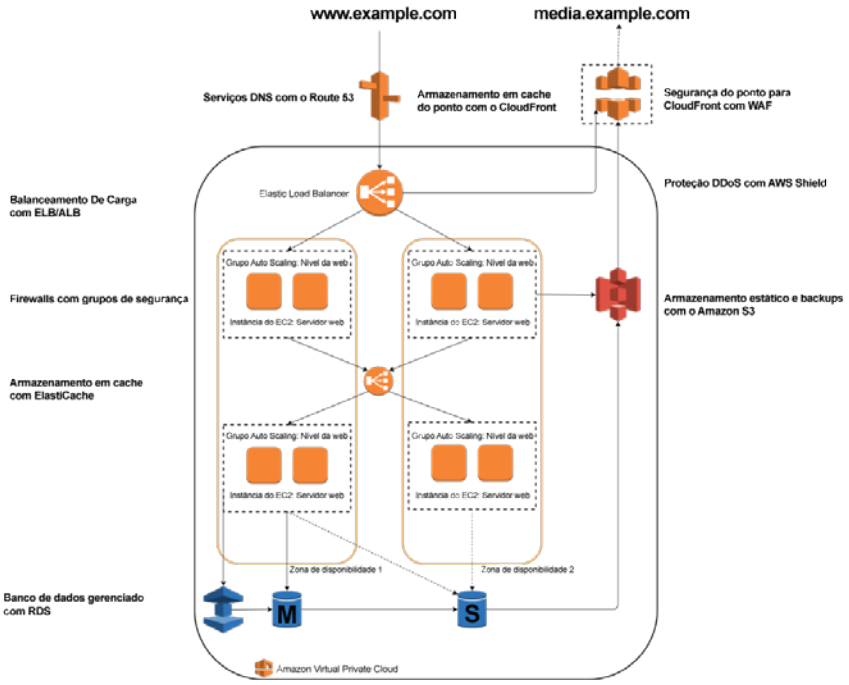

Figura 3. Um exemplo de uma arquitetura de hospedagem na web na AWS

1. **Balanceamento de carga com Elastic Load Balancing (ELB)/Application Load Balancer (ALB)** – Permite distribuir a carga por vários grupos de Zonas de disponibilidade e grupos de Auto Scaling do Amazon EC2 para redundância e desacoplamento de serviços.

2. **Firewalls com grupos de segurança** – Move a segurança para a instância para fornecer um firewall de nível de host com monitoração de estado dos servidores da Web e de aplicativos.
3. **Cache com o Amazon ElastiCache** – Fornece serviços de armazenamento em cache com o Redis ou o Memcached para remover carga do aplicativo e do banco de dados e diminuir a latência para solicitações frequentes.
4. **Banco de dados gerenciado com o Amazon RDS** – Cria uma arquitetura de banco de dados Multi-AZ altamente disponível com seis mecanismos de banco de dados possíveis.
5. **Serviços DNS com o Amazon Route 53** – Fornece serviços de DNS para simplificar o gerenciamento de domínio.
6. **O cache de ponto com o Amazon CloudFront** – O cache de ponto armazena conteúdo de alto volume para diminuir a latência para os clientes.
7. **Segurança de ponto do Amazon CloudFront com AWS WAF** – Filtra o tráfego malicioso, incluindo injeção de XSS e SQL por meio de regras definidas pelo cliente.
8. **Proteção contra DDoS com o AWS Shield** – Protege automaticamente sua infraestrutura contra os ataques DDoS de camada de rede e transporte mais comuns.
9. **Armazenamento estático e backups com o Amazon S3** – Ativa o armazenamento simples de objetos baseado em HTTP para backups e ativos estáticos, como imagens e vídeos.

Principais componentes de uma arquitetura de hospedagem na web da AWS

As seções a seguir descrevem alguns dos principais componentes de uma arquitetura de hospedagem na web implantada na Nuvem AWS e explicam como eles diferem de uma arquitetura de hospedagem tradicional na web.

Gerenciamento de rede

Em um ambiente de nuvem, como o AWS, a capacidade de segmentar sua rede a partir de outros clientes permite uma arquitetura mais segura e escalável. Embora os grupos de segurança ofereçam segurança no nível do host (consulte a

seção Segurança do host), o Amazon Virtual Private Cloud (Amazon VPC) permite ativar recursos em uma rede virtual isolada e lógica definida por você.[1]

O Amazon VPC é um serviço gratuito que oferece controle total sobre os detalhes de sua configuração de rede na AWS. Exemplos desse controle incluem a criação de sub-redes voltadas ao público para servidores web e sub-redes privadas sem acesso à Internet para seus bancos de dados. Além disso, o Amazon VPC permite que criar arquiteturas híbridas usando redes virtuais privadas (VPNs) de hardware e usar a Nuvem AWS como uma extensão do seu próprio datacenter.

O Amazon VPC também inclui suporte a IPv6, além do tradicional suporte IPv4 para sua rede.

Entrega de conteúdo

O cache de ponto ainda é relevante na infraestrutura de computação na Nuvem AWS. Todas as soluções existentes na infraestrutura de aplicativos web devem funcionar bem na Nuvem AWS. Uma opção adicional, no entanto, é usar o Amazon CloudFront para o cache de ponto do seu site.[2]

É possível usar o CloudFront para fornecer seu site, incluindo conteúdo dinâmico, estático e de streaming contínuo usando uma rede global de pontos de presença. O CloudFront encaminha automaticamente as solicitações de seu conteúdo para o ponto de presença mais próximo, para que o conteúdo seja entregue com o melhor desempenho possível. O CloudFront é otimizado para trabalhar com outros serviços da AWS, como Amazon Elastic Compute Cloud (Amazon S3) e o Amazon Simple Storage Service[3] (Amazon EC2).[4] O CloudFront também funciona perfeitamente com qualquer servidor de origem que não seja um servidor de origem da AWS, que armazena as versões originais e definitivas de seus arquivos.

Como outros serviços da AWS, não há contratos ou compromissos mensais para usar o CloudFront - você paga apenas pelo maior ou menor conteúdo que realmente entrega através do serviço.

Gerenciando o DNS público

Mover um aplicativo web para a Nuvem AWS requer algumas alterações de DNS para aproveitar as várias Zonas de disponibilidade fornecidas pela AWS. Para ajudar a gerenciar o roteamento de DNS, a AWS fornece o Amazon Route 53,[5]

um serviço da web de DNS altamente disponível e escalável. O Amazon Route 53 encaminha automaticamente as consultas para o seu domínio para o servidor DNS mais próximo. Como resultado, as consultas são respondidas com o melhor desempenho possível. O Amazon Route 53 resolve solicitações do seu nome de domínio (por exemplo, www.example.com) para seu Classic Load Balancer, bem como para seu registro de apex de zona (example.com).

Segurança de host

Ao contrário de um modelo de hospedagem tradicional na web, a filtragem de tráfego de rede de entrada não deve ser confinada no ponto; ele também deve ser aplicado no nível do host. O Amazon EC2 fornece um recurso chamado grupos de segurança. Um grupo de segurança é análogo a um firewall de rede de entrada, para o qual você pode especificar os protocolos, as portas e os intervalos de IP de origem que podem atingir suas instâncias do EC2. É possível atribuir um ou mais grupos de segurança a cada instância do EC2. Cada grupo de segurança encaminha o tráfego apropriado para cada instância. Os grupos de segurança podem ser configurados de forma que apenas sub-redes ou endereços IP específicos tenham acesso a uma instância do EC2. Ou podem fazer referência a outros grupos de segurança para limitar o acesso a instâncias do EC2 que estão em grupos específicos.

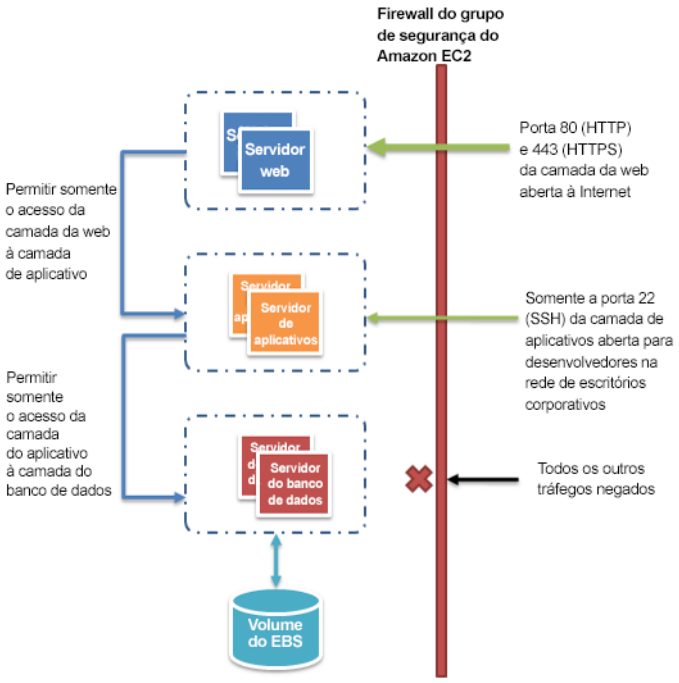

Figura 4. Grupos de segurança em um aplicativo web

No exemplo de arquitetura de hospedagem na web da AWS na Figura 4, o grupo de segurança para o cluster de servidor da web pode permitir acesso a qualquer host somente por TCP nas portas 80 e 443 (HTTP e HTTPS), e por instâncias no grupo de segurança do servidor de aplicativos na porta 22 (SSH) para gerenciamento direto do host. O grupo de segurança do servidor de aplicativos, por outro lado, pode permitir o acesso do grupo de segurança do servidor web para manipular solicitações da web e da sub-rede provenientes da sua organização por TCP na porta 22 (SSH) para gerenciamento direto do host. Nesse modelo, os engenheiros de suporte podem efetuar login diretamente nos servidores de aplicativos a partir da rede corporativa e acessar os outros clusters nas caixas do servidor de aplicativos. Para uma discussão mais aprofundada sobre segurança, consulte o Centro de Segurança da AWS[6] O centro contém boletins de segurança, informações de certificação e documentos técnicos de segurança que explicam os recursos de segurança da AWS.

Balanceamento de carga nos clusters

Os load balancers de hardware são um dispositivo de rede usado em arquiteturas de aplicativos web. A AWS fornece esse recurso por meio do serviço Elastic Load Balancing (ELB).[7] O ELB é uma solução de balanceamento de carga configurável que oferece suporte a verificações de integridade em hosts, distribuição de tráfego para instâncias do EC2 em várias Zonas de disponibilidade, além de adição dinâmica e remoção de hosts do Amazon EC2 a partir da rotação de balanceamento de carga. O ELB também pode aumentar e diminuir dinamicamente a capacidade de balanceamento de carga para se ajustar às demandas de tráfego, fornecendo um ponto de entrada previsível usando um CNAME persistente. O ELB também oferece suporte às sticky sessions para atender às necessidades mais avançadas de roteamento. Se o seu aplicativo exigir recursos de balanceamento de carga mais avançados, é possível executar um pacote de balanceamento de carga de software (por exemplo, Zeus, HAProxy ou NGINX Plus) em instâncias do EC2. Em seguida, você pode atribuir endereços IP elásticos a essas instâncias EC2 de balanceamento de carga para minimizar as alterações de DNS.[8]

Encontrando Outros hosts e serviços

Na arquitetura de hospedagem tradicional na web, a maioria dos seus hosts tem endereços IP estáticos. Na nuvem, a maioria dos seus hosts terá endereços IP dinâmicos. Embora todas as instâncias EC2 possam ter entradas DNS públicas e privadas, e sejam endereçáveis pela Internet, as entradas DNS e os endereços IP são atribuídos dinamicamente quando você inicia a instância. Eles não podem

ser atribuídos manualmente. Endereços IP estáticos (endereços IP elásticos na terminologia da AWS) podem ser atribuídos a instâncias em execução após o lançamento. Você deve usar os endereços Elastic IP para instâncias e serviços que exijam endpoints consistentes, como bancos de dados mestres, servidores de arquivos centrais e balanceadores de carga hospedados pelo EC2.

As funções de servidor que podem ser dimensionadas e desdobradas facilmente, como servidores da web, devem ser disponibilizadas em seus pontos de extremidade dinâmicos, registrando seu endereço IP em um repositório central. Como a maioria das arquiteturas de aplicativos web tem um servidor de banco de dados que está sempre ativo, o servidor de banco de dados é um repositório comum para informações de descoberta. Para situações em que o endereçamento consistente é necessário, as instâncias podem ser alocadas em endereços IP elásticos de um grupo de endereços por um script de bootstrapping quando a instância for iniciada.

Usando esse modelo, hosts recém-adicionados podem solicitar a lista de endpoints necessários para comunicações do banco de dados como parte de uma fase de bootstrapping. A localização do banco de dados pode ser fornecida como dados do usuário[9] que é passado para cada instância conforme é lançado. Alternativamente, é possível usar o Amazon SimpleDB para armazenar e manter informações de configuração.[10] O SimpleDB é um serviço altamente disponível que está disponível em um endpoint conhecido.

Cache dentro do aplicativo web

Os caches de aplicativos na memória podem reduzir a carga nos serviços e melhorar o desempenho e a escalabilidade na camada do banco de dados, armazenando informações usadas com frequência em cache. O Amazon ElastiCache[11] é um serviço da web que torna fácil implantar, operar e escalar um cache na memória na nuvem. Você pode configurar o cache na memória criado, para dimensionar automaticamente com carga e substituir automaticamente os nós com falha. O ElastiCache é compatível com o protocolo Memcached e Redis, o que simplifica a migração de sua solução local atual.

Configuração do banco de dados, backup e failover

Muitos aplicativos web contêm alguma forma de persistência, geralmente na forma de um banco de dados relacional ou NoSQL. A AWS oferece infraestrutura de banco de dados relacional e NoSQL. Como alternativa, é possível implantar seu próprio software de banco de dados em uma instância do

EC2. A tabela a seguir resume essas opções e as discutimos em mais detalhes nesta seção.

	Soluções de banco de dados relacional	Soluções NoSQL
Serviço de banco de dados gerenciado	Amazon RDS - MySQL, Oracle, Servidor SQL, MariaDB, PostgreSQL, Amazon Aurora	Amazon DynamoDB
Autogerenciado	Hospedando um DBMS relacional em uma Instância do EC2	Hospedando uma solução NoSQL em uma instância do EC2

Amazon RDS

O Amazon Relational Database Service (Amazon RDS) fornece acesso aos recursos de um mecanismo familiar de banco de dados MySQL, PostgreSQL, Oracle e Microsoft SQL Server.[12] O código, os aplicativos e as ferramentas que você já usa podem ser usados com o Amazon RDS. O Amazon RDS corrige automaticamente o software de banco de dados e faz backup do banco de dados, e armazena os backups por um período de retenção definido pelo usuário. Ele também oferece suporte à recuperação point-in-time. Você se beneficia da flexibilidade de poder dimensionar os recursos de computação ou a capacidade de armazenamento associada à sua instância de banco de dados relacional, fazendo uma única chamada de API.

Além disso, as implantações Multi-AZ do Amazon RDS aumentam a disponibilidade e protegem seu banco de dados contra indisponibilidades não planejadas. As réplicas de leitura do Amazon RDS fornecem réplicas somente de leitura de seu banco de dados, para que você possa dimensionar além da capacidade de uma única implantação de banco de dados para cargas de trabalho de banco de dados com leitura intensa. Como em todos os serviços da AWS, nenhum investimento inicial é necessário e você paga apenas pelos recursos que usa.

Hospedando um sistema de gerenciamento de banco de dados relacional (RDBMS) em uma instância do Amazon EC2

Além da oferta gerenciada do Amazon RDS, você pode instalar sua opção de RDBMS (como MySQL, Oracle, SQL Server ou DB2) em uma instância do EC2 e gerenciá-la por conta própria. Os clientes da AWS que hospedam um banco de dados no Amazon EC2 usam com êxito vários modelos mestre/escravo e de

replicação, incluindo espelhamento para cópias de somente leitura e envio de logs para escravos passivos sempre prontos.

Ao gerenciar seu próprio software de banco de dados diretamente no Amazon EC2, é possível considerar a disponibilidade de armazenamento tolerante a falhas e persistente. Para tanto, recomendamos que os bancos de dados executados no Amazon EC2 usem volumes do Amazon Elastic Block Store (Amazon EBS), que sejam semelhantes ao armazenamento anexado à rede.[13] Para instâncias do EC2 que executam um banco de dados, você deve colocar todos os dados e logs do banco de dados nos volumes do EBS. Estes permanecerão disponíveis mesmo se o host do banco de dados falhar. Essa configuração permite um cenário de failover simples, no qual uma nova instância do EC2 pode ser iniciada se um host falhar e os volumes existentes do EBS puderem ser associados à nova instância. O banco de dados pode então continuar de onde parou.

Os volumes do EBS fornecem automaticamente redundância dentro da Zona de disponibilidade, o que aumenta sua disponibilidade em discos simples. Se o desempenho de um único volume do EBS não for suficiente para suas necessidades de bancos de dados, os volumes poderão ser distribuídos para aumentar o desempenho de IOPS no seu banco de dados. Para cargas de trabalho exigentes, você também pode usar IOPS provisionadas pelo EBS, onde especifica as IOPS necessárias. Se você usa o Amazon RDS, o serviço gerencia seu próprio armazenamento para que você possa se concentrar no gerenciamento de seus dados.

Soluções NoSQL

Além do suporte a bancos de dados relacionais, a AWS também oferece o Amazon DynamoDB, um serviço de banco de dados NoSQL totalmente gerenciado que fornece desempenho rápido e previsível com escalabilidade perfeita.[14] Usando o Console de Gerenciamento da AWS ou a API do DynamoDB, é possível aumentar ou diminuir a capacidade sem tempo de inatividade ou degradação do desempenho. Como o DynamoDB lida com as cargas administrativas de operação e escalabilidade de bancos de dados distribuídos para AWS, não é necessário se preocupar com provisionamento de hardware, instalação e configuração, replicação, correção de software ou escalabilidade de clusters.

O Amazon SimpleDB fornece um serviço de banco de dados não relacional leve, altamente disponível e tolerante a falhas, que oferece consulta e indexação de dados sem o requisito de um esquema fixo. O SimpleDB pode ser uma substituição muito eficaz para os bancos de dados em cenários de acesso a dados que exijam uma tabela de esquemas grande, altamente indexada e flexível.

Além disso, você pode usar o Amazon EC2 para hospedar muitas outras tecnologias emergentes no movimento NoSQL, como Cassandra, CouchDB e MongoDB.

Armazenamento e backup de dados e ativos

Existem inúmeras opções dentro da Nuvem AWS para armazenar, acessar e fazer backup dos dados e ativos do seu aplicativo web. O Amazon S3 fornece um depósito de objetos altamente disponível e redundante. O Amazon S3 é uma ótima solução de armazenamento para objetos pouco estáticos ou que mudam lentamente, como imagens, vídeos e outras mídias estáticas. O Amazon S3 também suporta cache de ponto e streaming desses ativos, interagindo com o CloudFront.

Para armazenamento semelhante a um sistema de arquivos anexado, as instâncias do EC2 podem ter volumes do EBS conectados. Eles atuam como discos montáveis para executar instâncias do EC2. O Amazon EBS é ótimo para dados que precisam ser acessados como armazenamento em bloco e que requeiram persistência além da duração da instância em execução, como partições de banco de dados e logs de aplicativos.

Além de ter uma vida útil independente da instância do EC2, é possível obter instantâneos de volumes do EBS e armazená-los no Amazon S3. Como os snapshots do EBS fazem backup apenas das alterações feitas desde o último snapshot, os snapshots mais frequentes podem reduzir o tempo do snapshot. Também é possível usar um snapshot do EBS como uma linha de base para replicar dados em vários volumes do EBS e associa-los a outras instâncias em execução.

Os volumes do EBS podem ter até 16 TB, e vários volumes do EBS podem ser distribuídos para volumes ainda maiores ou para aumentar o desempenho de E/S. Para maximizar o desempenho de seus aplicativos com uso intensivo de E/S, é possível usar volumes de IOPS provisionados. Os volumes de IOPS provisionados foram criados para atender às necessidades de cargas de trabalho

intensivas de E/S, particularmente cargas de trabalho de banco de dados sensíveis ao desempenho de armazenamento e à consistência na taxa de transferência de E/S de acesso aleatório. Você especifica uma taxa de IOPS ao criar o volume e as provisões do Amazon EBS que classificam durante a vida útil do volume. O Amazon EBS atualmente suporta até 20.000 IOPS por volume. É possível distribuir vários volumes para entregar milhares de IOPS por instância ao seu aplicativo.

Escalabilidade automática da frota

Uma das principais diferenças entre a arquitetura de Nuvem AWS e o modelo de hospedagem tradicional é que a AWS pode dimensionar automaticamente a frota de aplicativos web sob demanda para lidar com as alterações no tráfego. No modelo de hospedagem tradicional, os modelos de previsão de tráfego geralmente são usados para provisionar hosts antes do tráfego projetado. Na AWS, as instâncias podem ser provisionadas dinamicamente de acordo com um conjunto de triggers para redimensionar a frota e recuar. O serviço de Auto Scaling pode criar grupos de capacidade de servidores que podem aumentar ou diminuir sob demanda.[15] O Auto Scaling também funciona diretamente com o Amazon CloudWatch para dados de métricas e com o Elastic Load Balancing para adicionar e remover hosts para distribuição de carga. Por exemplo, se os servidores web estiverem relatando mais de 80% de utilização da CPU durante um período de tempo, um servidor web adicional poderá ser implantado rapidamente e adicionado automaticamente ao load balancer para inclusão imediata na rotação de balanceamento de carga.

Conforme mostrado no modelo de arquitetura de hospedagem na web da AWS, você pode criar vários grupos de Auto Scaling para diferentes camadas da arquitetura, para que cada camada possa ser dimensionada independentemente. Por exemplo, o grupo de Auto Scaling do servidor web pode acionar a entrada e a saída de escala em resposta a alterações na E/S da rede, enquanto o grupo de Auto Scaling do servidor de aplicativos pode ser dimensionado e utilizado de acordo com a utilização da CPU. Você pode definir os valores mínimos e máximos para ajudar a garantir a disponibilidade 24 horas por dia, sete dias por semana e limitar o uso dentro de um grupo.

Os triggers do Auto Scaling podem ser configurados para aumentar e reduzir a frota total em uma determinada camada para corresponder a utilização de recursos à demanda real. Além do serviço Auto Scaling, é possível escalar as

frotas do Amazon EC2 diretamente por meio da API do Amazon EC2, que permite iniciar, encerrar e inspecionar instâncias.

Recursos adicionais de segurança

O número e a sofisticação dos Ataques DDoS estão aumentando. Tradicionalmente, esses ataques são difíceis de se defender. Eles geralmente acabam sendo caros, tanto em questão de tempo de mitigação quanto gasto de energia, bem como custo de oportunidade de visitas perdidas ao seu site durante o ataque. Há vários fatores e serviços da AWS que podem ajudá-lo a se defender contra esses ataques. A primeira é a escala da rede da AWS. A infraestrutura da AWS é muito grande e permitimos que você aproveite nossa escala para otimizar sua defesa. Vários serviços, incluindo o Elastic Load Balancing, o Amazon CloudFront e o Amazon Route 53, são eficazes na escalabilidade de seu aplicativo web em resposta a um grande aumento no tráfego.

Dois serviços em particular ajudam na sua estratégia de defesa. O AWS Shield é um serviço de proteção contra DDoS gerenciado que ajuda a proteger contra várias formas de vetores de ataque de DDoS.[16] A oferta padrão do AWS Shield é gratuita e ativa automaticamente em toda a sua conta. Essa oferta padrão ajuda a defender contra os ataques mais comuns na camada de rede e de transporte. Além desse nível, a oferta avançada concede níveis mais altos de proteção contra seu aplicativo web, fornecendo visibilidade quase em tempo real de um ataque contínuo, bem como integrando em níveis mais altos com os serviços mencionados anteriormente. Além disso, você obtém acesso à Equipe de Resposta DDoS (DRT) da AWS para ajudar a reduzir os ataques em grande escala e sofisticados contra seus recursos.

O AWS WAF (firewall de aplicativo web) foi projetado para proteger seus aplicativos web contra ataques que possam comprometer a disponibilidade ou a segurança ou, de outra forma, consumir recursos excessivos.[17] O AWS WAF funciona em linha com o CloudFront ou o Application Load Balancer, juntamente com suas regras personalizadas, para se defender contra ataques como scripts entre sites, injeção de SQL e DDoS. Assim como a maioria dos serviços da AWS, o AWS WAF vem com uma API com todos os recursos que pode ajudar a automatizar a criação e a edição de regras para o WAF conforme a mudanças de suas necessidades de segurança.

Failover com a AWS

Outra vantagem importante da AWS em relação à hospedagem tradicional na web são as Zonas de disponibilidade, que oferecem acesso fácil a locais de implantação redundantes. As Zonas de disponibilidade são locais fisicamente distintos projetados para serem isolados de falhas em outras zonas de disponibilidade. Elas fornecem conectividade de rede barata e de baixa latência para outras Zonas de disponibilidade na mesma região da AWS. Como mostra o diagrama de arquitetura de hospedagem na web da AWS na Figura 3, recomendamos implantar hosts EC2 em várias Zonas de disponibilidade para tornar seu aplicativo web mais tolerante a falhas. É importante garantir que haja provisões para migrar pontos únicos de acesso pelas Zonas de disponibilidade em caso de falha. Por exemplo, você deve configurar um escravo de banco de dados em uma segunda Zona de disponibilidade para que a persistência dos dados permaneça consistente e altamente disponível, mesmo durante um cenário de falha improvável. Você pode fazer isso no Amazon EC2 ou no Amazon RDS com o clique de um botão.

Embora algumas alterações arquiteturais sejam frequentemente necessárias ao mover um aplicativo web existente para a Nuvem AWS, há melhorias significativas na escalabilidade, confiabilidade e relação custo-benefício que fazem com que o uso da Nuvem AWS valha a pena. Na próxima seção, discutiremos essas melhorias.

Principais considerações ao usar o AWS para hospedagem na web

Existem algumas diferenças importantes entre a Nuvem AWS e um modelo de hospedagem de aplicativos tradicional na web. A seção anterior destacou muitas das principais áreas que você deve considerar ao implantar um aplicativo web na nuvem. Esta seção indica algumas das principais mudanças arquitetônicas que você precisa considerar ao trazer qualquer aplicativo para a nuvem.

Não há mais aparelhos de rede física

Você não pode implantar dispositivos de rede física na AWS. Por exemplo, firewalls, roteadores e load balancers para seus aplicativos da AWS não podem mais residir em dispositivos físicos, mas devem ser substituídos por soluções de software. Existe uma ampla variedade de soluções de software de qualidade corporativa, seja para balanceamento de carga (por exemplo, Zeus, HAProxy,

NGINX Plus e Pound) ou para estabelecer uma conexão VPN (por exemplo, OpenVPN, OpenSwan e Vyatta). Essa não é uma limitação do que pode ser executado na Nuvem AWS, mas é uma alteração de arquitetura do seu aplicativo se usar esses dispositivos hoje.

Firewalls em todos os lugares

Onde você já teve uma DMZ simples e, em seguida, abriu as comunicações entre seus hosts em um modelo de hospedagem tradicional, a AWS aplica um modelo mais seguro, no qual cada host é bloqueado. Uma das etapas no planejamento de uma implantação da AWS é a análise do tráfego entre hosts. Essa análise orientará as decisões sobre exatamente quais portas precisam ser abertas. É possível criar grupos de segurança no Amazon EC2 para cada tipo de host em sua arquitetura. Além disso, é possível criar uma grande variedade de modelos de segurança simples e em camadas para permitir o acesso mínimo entre os hosts em sua arquitetura. O uso de listas de controle de acesso à rede no Amazon VPC pode ajudar a bloquear sua rede no nível de sub-rede.

Considere a disponibilidade de vários datacenters

Pense em Zonas de disponibilidade dentro de uma região da AWS como vários datacenters. As instâncias do EC2 em diferentes Zonas de disponibilidade são lógica e fisicamente separadas, e fornecem um modelo fácil de usar para implantar seu aplicativo nos datacenters, proporcionando alta disponibilidade e confiabilidade. O Amazon VPC como um serviço regional permite aproveitar as zonas de disponibilidade, mantendo todos os seus recursos na mesma rede lógica.

Tratar hosts como efêmeros e dinâmicos

Provavelmente, a mudança mais importante com a qual você pode arquitetar seu aplicativo da AWS é que os hosts do Amazon EC2 devem ser considerados efêmeros e dinâmicos. Qualquer aplicativo criado para a Nuvem AWS não deve presumir que um host estará sempre disponível e deve ser desenvolvido com o conhecimento de que os dados que não estiverem em um volume do EBS serão perdidos se uma instância do EC2 falhar. Além disso, quando um novo host é criado, você não deve fazer suposições sobre o endereço IP ou o local dentro de uma zona de disponibilidade do host. Seu modelo de configuração deve ser flexível e sua abordagem para inicializar um host deve levar em conta a natureza dinâmica da nuvem. Essas técnicas são críticas para construir e executar um aplicativo altamente escalável e tolerante a falhas.

Considere uma arquitetura sem servidor

Este whitepaper foca principalmente em uma arquitetura web mais tradicional. No entanto, serviços mais novos como o AWS Lambda[18] e o Amazon API Gateway[19] permitem criar um aplicativo web sem servidor que abstrai o uso de máquinas virtuais para executar computação. Nesses casos, o código é executado por solicitação, e você paga apenas pelo número e tamanho das solicitações. Você pode descobrir mais sobre as arquiteturas sem servidor aqui.

Conclusões

Existem inúmeras considerações arquitetônicas e conceituais quando você está pensando em migrar seu aplicativo web para a Nuvem AWS. Os benefícios de ter uma infraestrutura econômica, altamente escalável e tolerante a falhas que cresce com seus negócios superam em muito os esforços de migração para a Nuvem AWS.

Contribuidores

As pessoas e organizações a seguir contribuíram com este documento:

- Jack Hemion, arquiteto associado de soluções, AWS
- Matt Tavis, arquiteto principal de soluções, AWS
- Philip Fitzsimons, gerente sênior do Well-Architected, AWS

Outras leituras

- Guia de conceitos básicos - Hospedagem de aplicativos web da AWS para Linux
- Guia de conceitos básicos - Hospedagem de aplicativos web da AWS para Linux
- Conceitos básicos da série de vídeos: Aplicativos web Linux na Nuvem AWS
- Conceitos básicos da série de vídeos: Aplicativos web .NET na Nuvem AWS

Revisões do documento

Data	Descrição
Julho de 2017	Várias seções adicionadas e atualizadas para novos serviços. Diagramas atualizados para maior clareza e serviços. Adição de VPC como método de rede padrão na AWS em “Gerenciamento de rede”. Foi adicionada a seção sobre proteção e mitigação de DDoS em “Recursos de segurança adicionais”. Foi adicionada uma pequena seção sobre arquiteturas sem servidor para hospedagem na web.
Setembro de 2012	Várias seções atualizadas para melhorar a clareza. Diagramas atualizados para usar os ícones da AWS. Adição da seção “Gerenciando DNS públicos” para obter detalhes sobre o Amazon Route 53. Seção “Encontrando Outros Hosts e Serviços” atualizada para maior clareza. Seção “Configuração do banco de dados, backup e failover” atualizada para maior clareza e DynamoDB. Seção “Armazenamento e backup de dados e ativos” expandida para cobrir os volumes de IOPS provisionadas do EBS.
Maio de 2010	Primeira publicação

Observações

[1] https://aws.amazon.com/vpc/

[2] https://aws.amazon.com/cloudfront/

[3] https://aws.amazon.com/s3/

[4] https://aws.amazon.com/ec2/

[5] https://aws.amazon.com/route53/

[6] https://aws.amazon.com/security/

[7] https://aws.amazon.com/elasticloadbalancing/

[8] Endereços IP elásticos são endereços IP estáticos projetados para computação em nuvem dinâmica, que você pode mover de uma instância para outra.

[9] http://docs.aws.amazon.com/AWSEC2/latest/APIReference/Welcome.html

[10] https://aws.amazon.com/simpledb/

[11] https://aws.amazon.com/elasticache/

[12] https://aws.amazon.com/rds/

[13] https://aws.amazon.com/ebs/

[14] https://aws.amazon.com/dynamodb/

[15] https://aws.amazon.com/autoscaling/

[16] https://aws.amazon.com/shield/

[17] https://aws.amazon.com/waf/

[18] https://aws.amazon.com/lambda/

[19] https://aws.amazon.com/api-gateway/