

Computação em Nuvem

Qualidade de serviço em nuvem

Você sabia que seu material didático é interativo e multimídia? Isso significa que você pode interagir com o conteúdo de diversas formas, a qualquer hora e lugar. Na versão impressa, porém, alguns conteúdos interativos ficam desabilitados. Por essa razão, fique atento: sempre que possível, opte pela versão digital. Bons estudos!

Um dos obstáculos para migração de aplicações para provedores de Computação em Nuvem é o fato de que o acesso remoto aos serviços possa afetar negativamente o desempenho. De fato, sem uma conexão de rede de boa qualidade é difícil garantir uma experiência satisfatória para os usuários. Além disso, os provedores precisam de soluções para oferecer confiabilidade e escalabilidade para os serviços.

Diante disso, nesta webaula, vamos descrever as métricas de desempenho de rede e definir o conceito de Acordo de Nível de Serviço (SLA). Além disso, serão explicados mecanismos que visam aprimorar a escalabilidade, a confiabilidade e a disponibilidade de aplicações em nuvem. Por fim, vamos estudar, também, a recuperação de desastres, que consiste em mecanismos para lidar com falhas.

Qualidade de Serviço

Precisamos utilizar mecanismos para tentar assegurar uma experiência satisfatória aos usuários de aplicações em nuvem. O primeiro ponto a considerar é como caracterizar de forma objetiva a qualidade da comunicação. Isso significa que precisamos de métricas quantitativas para descrever os requisitos mínimos de desempenho. Para lidar com essas questões, há um conceito denominado Qualidade de Serviço (QoS – *Quality of Service*). Em se tratando de redes de computadores, a QoS pode ser entendida como uma abordagem utilizada para especificar parâmetros de desempenho das aplicações, assim como os mecanismos necessários para garantir os requisitos de desempenho estabelecidos (KAMIENSKI; SADOK, 2000). Os modelos de QoS podem ser utilizados para caracterizar objetivamente os requisitos de desempenho de uma aplicação. Além disso, utilizam métricas de desempenho de rede. Conheça algumas dessas métricas a seguir.

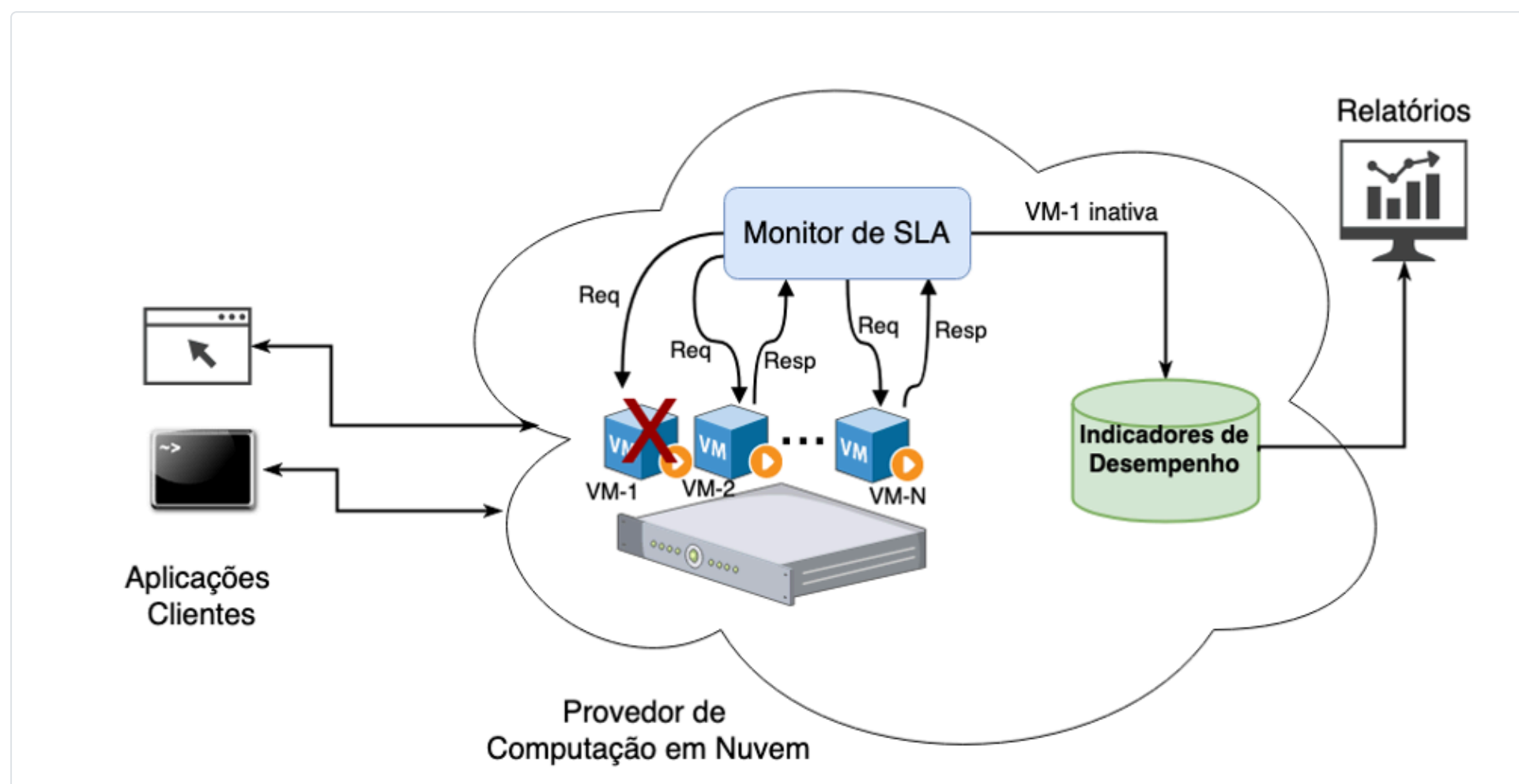
Atraso	▼
É o tempo total de transmissão de um pacote do nó remetente ao nó destinatário. Aplicações de chamadas de voz na Internet, por exemplo, requerem um atraso máximo de 150 ms (CHEN; FARLEY; YE, 2004).	
Jitter	▼
É uma medida da variação no atraso na transmissão dos pacotes. Quanto maior o jitter, pior é o desempenho de aplicações multimídia, como <i>streaming</i> de músicas na internet.	
Taxa de transmissão	▼
É o volume de dados efetivamente transmitido entre o nó remetente e o nó destinatário. Em geral, é medida em termos de megabits por segundo (Mbps). Por exemplo, a aplicação de <i>streaming</i> de vídeo sob demanda Netflix estabelece que, para assistir vídeos em resolução Ultra HD, é necessária uma conexão com uma taxa de pelo menos 25 Mbps (NETFLIX, 2019).	
Taxa de perda	▼
É a porcentagem dos pacotes que não foram entregues com sucesso para o nó destinatário. Por exemplo, se foram transmitidos 50 pacotes e apenas 40 foram efetivamente entregues ao destinatário, então a taxa de perda é de 20%, ou seja, 10 de 50 pacotes não foram entregues.	

Acordo de Nível de Serviço

Ao utilizar as métricas apresentadas e outras mais específicas, os provedores determinam condições para provisão dos serviços de Computação em Nuvem em um documento chamado Acordo de Nível de Serviço (SLA – *Service Level Agreement*) (ERL; PUTTINI; MAHMOO, 2013). O SLA descreve de forma objetiva as garantias de QoS, a confiabilidade e o desempenho de cada serviço.

Os provedores utilizam um mecanismo denominado Monitor de SLA para monitorar continuamente os indicadores de desempenho dos serviços em nuvem, a fim de verificar se eles estão de acordo com as métricas de qualidade estabelecidas no SLA (ERL; PUTTINI; MAHMOOD, 2013). O provedor mantém um banco de dados com as informações de desempenho coletadas por esse mecanismo. A figura a seguir mostra o papel do Monitor de SLA, considerando um exemplo no qual os serviços monitorados são máquinas virtuais.

Visão geral da atuação do Monitor de SLA



Fonte: elaborada pelo autor.

O Monitor de SLA verifica periodicamente se as VMs estão ativas, enviando mensagens de requisição (req) simples. Cada máquina virtual envia uma mensagem de resposta (resp). Nesse caso, a VM-1 não responde ao monitor e então é considerada inativa. As informações são armazenadas em um banco de dados com estatísticas sobre os serviços do provedor. As informações coletadas pelo monitor de SLA podem ser usadas pelo sistema de gerenciamento do provedor para gerar relatórios de análise de desempenho dos serviços e para fins de tarifação.

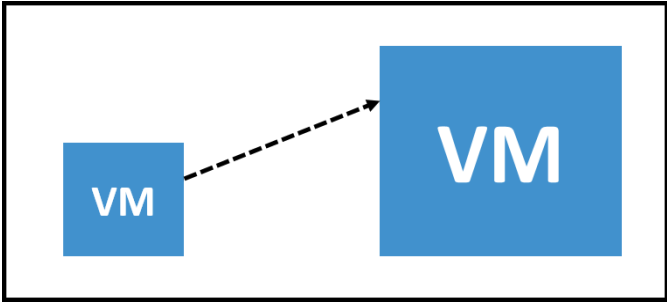
Mecanismo de dimensionamento automático

O mecanismo de dimensionamento automático (*automated scaling*) é responsável por ajustar a capacidade de um serviço em função das demandas. Se a carga de trabalho aumenta, o mecanismo aloca mais recursos, para manter a performance do serviço. Por exemplo, esse mecanismo pode automaticamente criar uma réplica de um banco de dados, para lidar com um aumento no número de consultas ao banco. Se a carga de trabalho diminui, o mecanismo libera recursos ociosos para reduzir custos. Assim, esse mecanismo confere escalabilidade aos serviços em nuvem de forma automatizada, buscando otimizar a relação entre custo e performance (MAO; HUMPHREY, 2011). O redimensionamento (escalonamento) pode ser vertical ou horizontal. Saiba mais a seguir.

Escalonamento vertical

O escalonamento vertical corresponde a aumentar a configuração de um recurso, por exemplo, reconfigurar uma máquina virtual com 8GB de memória para 16GB.

Escalonamento vertical

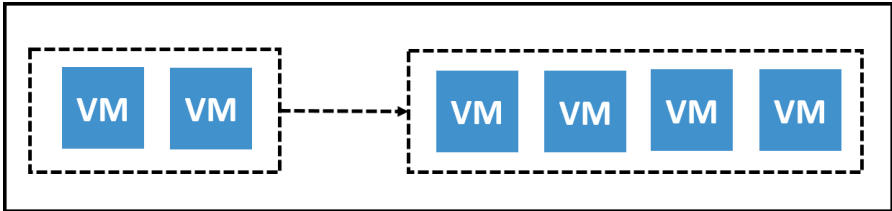


Fonte: elaborada pelo autor.

Escalonamento horizontal

O escalonamento horizontal corresponde a criar réplicas de uma instância e é bastante utilizado em aplicações web. Nesse caso, são criadas réplicas do servidor web ou do servidor de bancos de dados para atender a um aumento no número de requisições recebidas pela aplicação.

Escalonamento horizontal



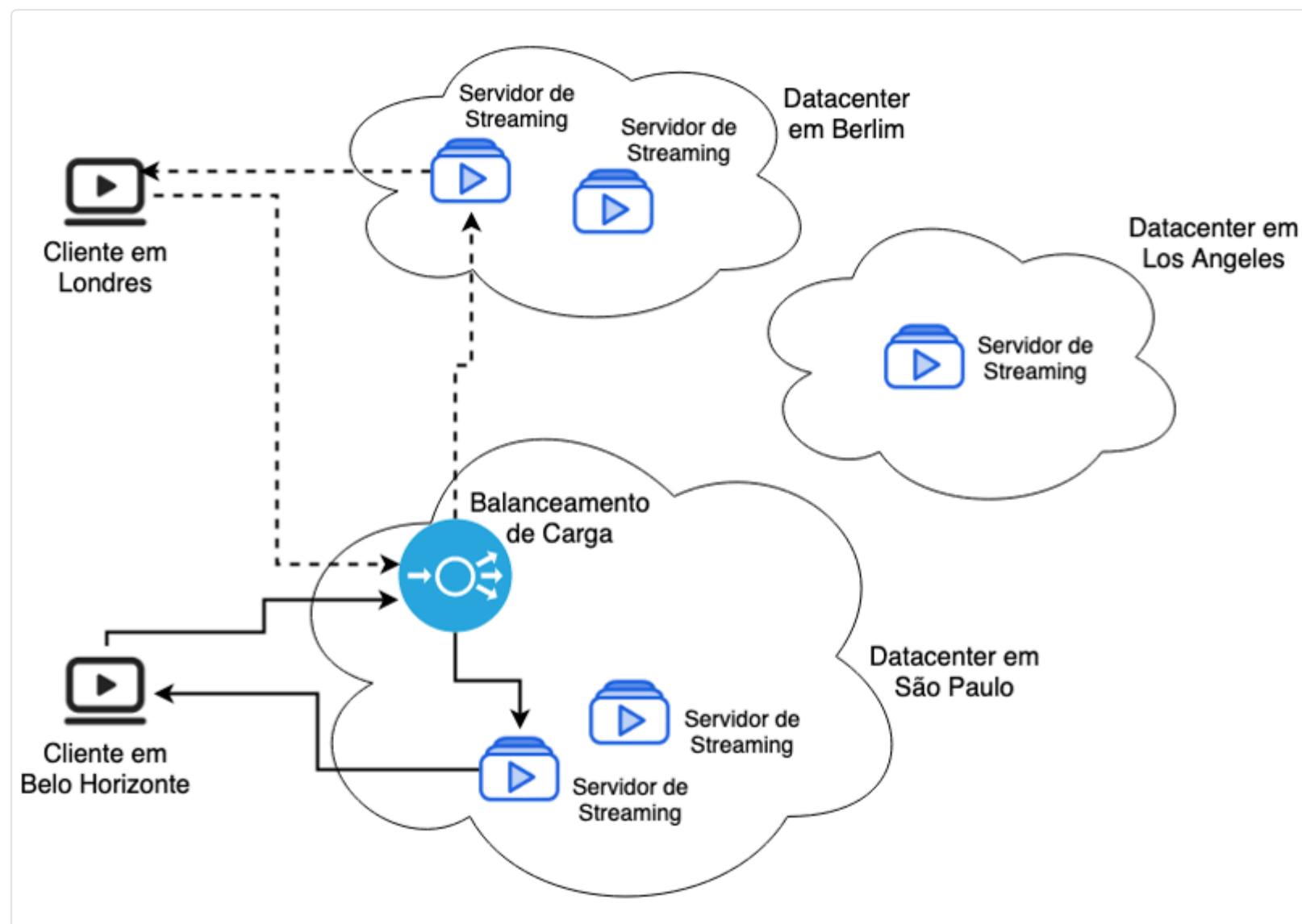
Fonte: elaborada pelo autor.

Balanceamento de carga

O balanceamento de carga é um mecanismo para distribuir a demanda de trabalho entre as réplicas de um serviço. Por exemplo, cada nova requisição que chega à rede do provedor pode ser encaminhada para a réplica menos sobrecarregada. Isso resulta em melhor performance, pois diminui o tempo de resposta das requisições. Em conjunto com o dimensionamento automático, o balanceamento de carga também favorece a escalabilidade e a disponibilidade do serviço, pois se uma instância de um serviço se tornar inativa (indisponível, devido a uma falha, por exemplo) as requisições a esse serviço podem ser encaminhadas para outra réplica.

A figura a seguir mostra um exemplo de um mecanismo de balanceamento de carga para um serviço de *streaming* de vídeo. Suponha que um provedor tenha três *data centers* em diferentes regiões: Alemanha, Califórnia e São Paulo. O mecanismo de balanceamento de carga escolhe a réplica do servidor de *streaming* mais próxima do cliente para melhorar a performance do serviço. De fato, existem várias técnicas para implementar o balanceamento de carga, como distribuição uniforme, balanceamento ponderado, entre outras (KANSAL; CHANA, 2012).

Balanceamento de carga baseado em localização



Fonte: elaborada pelo autor.

Mecanismo de recuperação a falhas

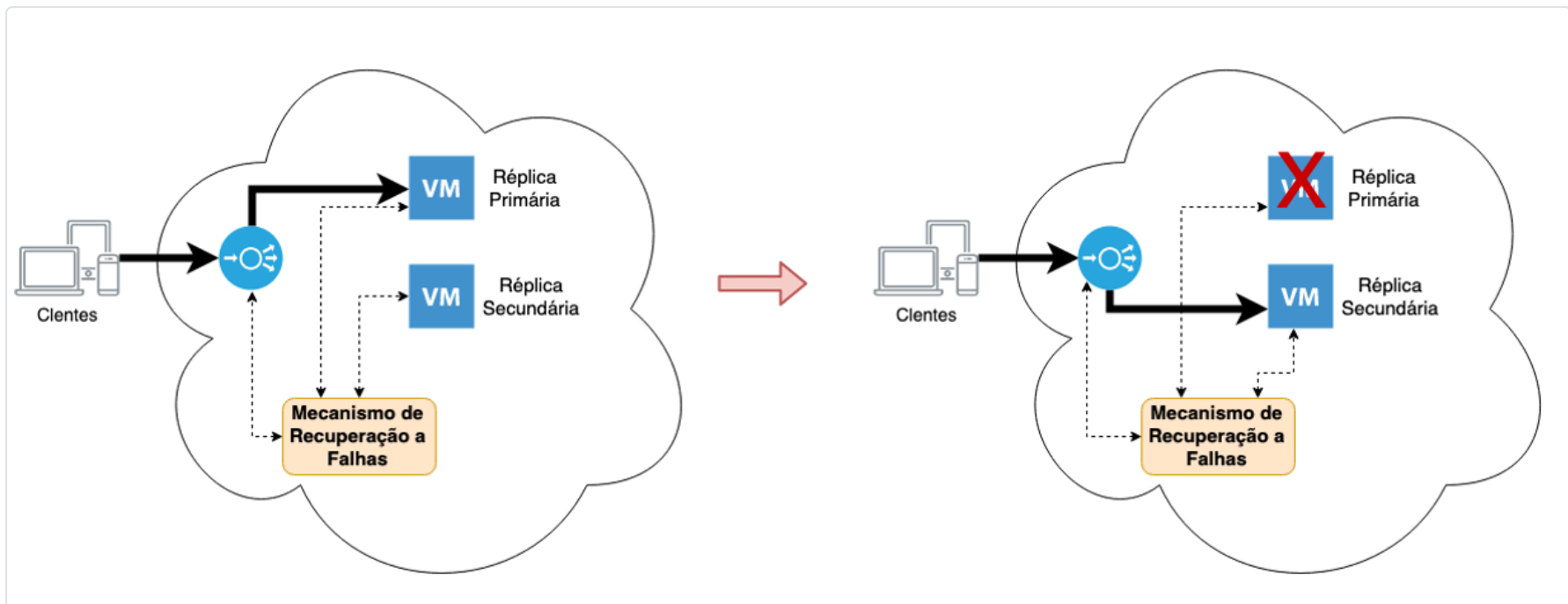
O mecanismo de recuperação a falhas trabalha em conjunto com os demais mecanismos. O objetivo desse mecanismo é identificar a ocorrência de falhas para que as requisições sejam redirecionadas somente para as réplicas do serviço que estejam ativas e funcionando corretamente. Quando uma instância falha, o mecanismo de balanceamento de carga é avisado para não redirecionar requisições para essa instância. Assim, a implementação e recuperação a falhas contribui também para aumentar a confiabilidade e a disponibilidade do serviço.

Observe que o mecanismo de recuperação a falhas depende da redundância (replicação) de recursos. Por exemplo, quando uma instância de um serviço falha, deve haver uma réplica (secundária) desse serviço já preparada para receber as requisições. De acordo com a forma como a réplica secundária é utilizada, existem dois modelos básicos de recuperação a falhas: modelo ativo-passivo e modelo ativo-ativo (ERL; PUTTINI; MAHMOO, 2013). Saiba mais a seguir.

Modelo ativo-passivo

No modelo ativo-passivo, a réplica secundária não é utilizada para atender requisições regularmente. Ela só é acionada quando a réplica principal falha. A figura a seguir exemplifica essa estratégia. Quando a máquina virtual primária falha, as requisições dos clientes são direcionadas para sua réplica.

Visão geral do modelo ativo-passivo

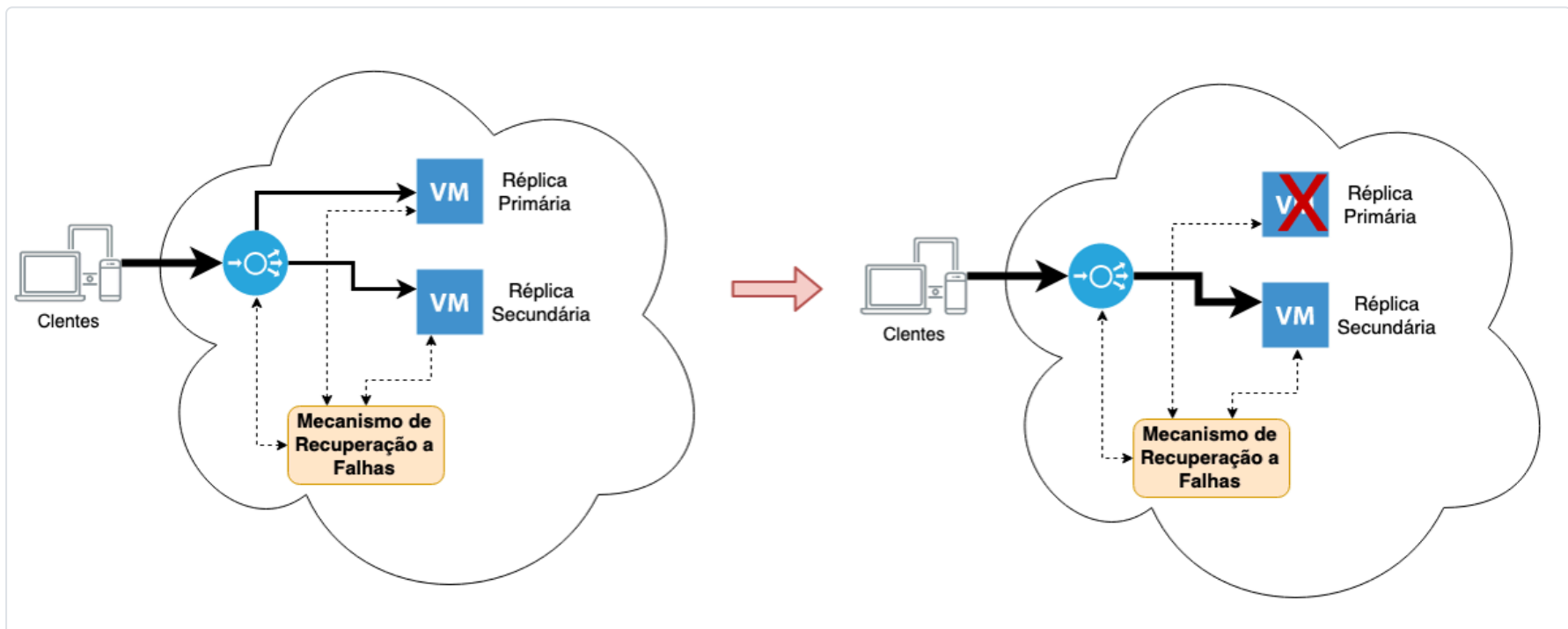


Fonte: elaborada pelo autor.

Modelo ativo-ativo

No modelo ativo-ativo, ambas as réplicas recebem requisições (sem distinção). Quando uma das réplicas falha, então todas as requisições são redirecionadas para a outra, até que a réplica que falhou seja corrigida ou uma nova réplica seja instanciada. A figura a seguir mostra um cenário baseado nesse modelo.

Visão geral do modelo ativo-ativo



Fonte: elaborada pelo autor.

Recuperação de desastres

Os mecanismos de recuperação a falhas, escalonamento automático e balanceamento de carga, aliados à escalabilidade dos serviços de Computação em Nuvem, são fundamentais para implementação de soluções de recuperação de desastres (CSA, 2011). Essas soluções são importantes para garantir a confiabilidade do serviço e a proteção dos dados. Além disso, o provedor deve manter um plano de recuperação de desastres que, entre outros aspectos, especifica as prioridades, serviços e dados mais críticos, além das estratégias de recuperação de falhas que devem ser usadas em cada caso.

Como exemplos de desastres podemos citar falhas nos sistemas de distribuição de energia elétrica que resultam em blackouts, podendo durar horas ou até mesmo dias. Também podem ocorrer catástrofes naturais como inundações. Nesses casos, as instalações de um *data center* poderiam ficar completamente inoperantes, o que afetaria a continuidade dos negócios que dependem da disponibilidade dos serviços do provedor. Imagine os prejuízos decorrentes da interrupção dos serviços para as aplicações dos clientes do provedor.

Nesta webaula, você compreendeu como podemos medir a qualidade das aplicações de forma objetiva, assim como as principais soluções para melhoria de desempenho e escalabilidade em ambientes de nuvem. Continue seus estudos para se aprofundar nessas temáticas!