SÃO
PAULO
TECH
SCHOOL

Infraestrutura em Nuvem

Aula 07

Marcio Santana

marcio.santana@sptech.school

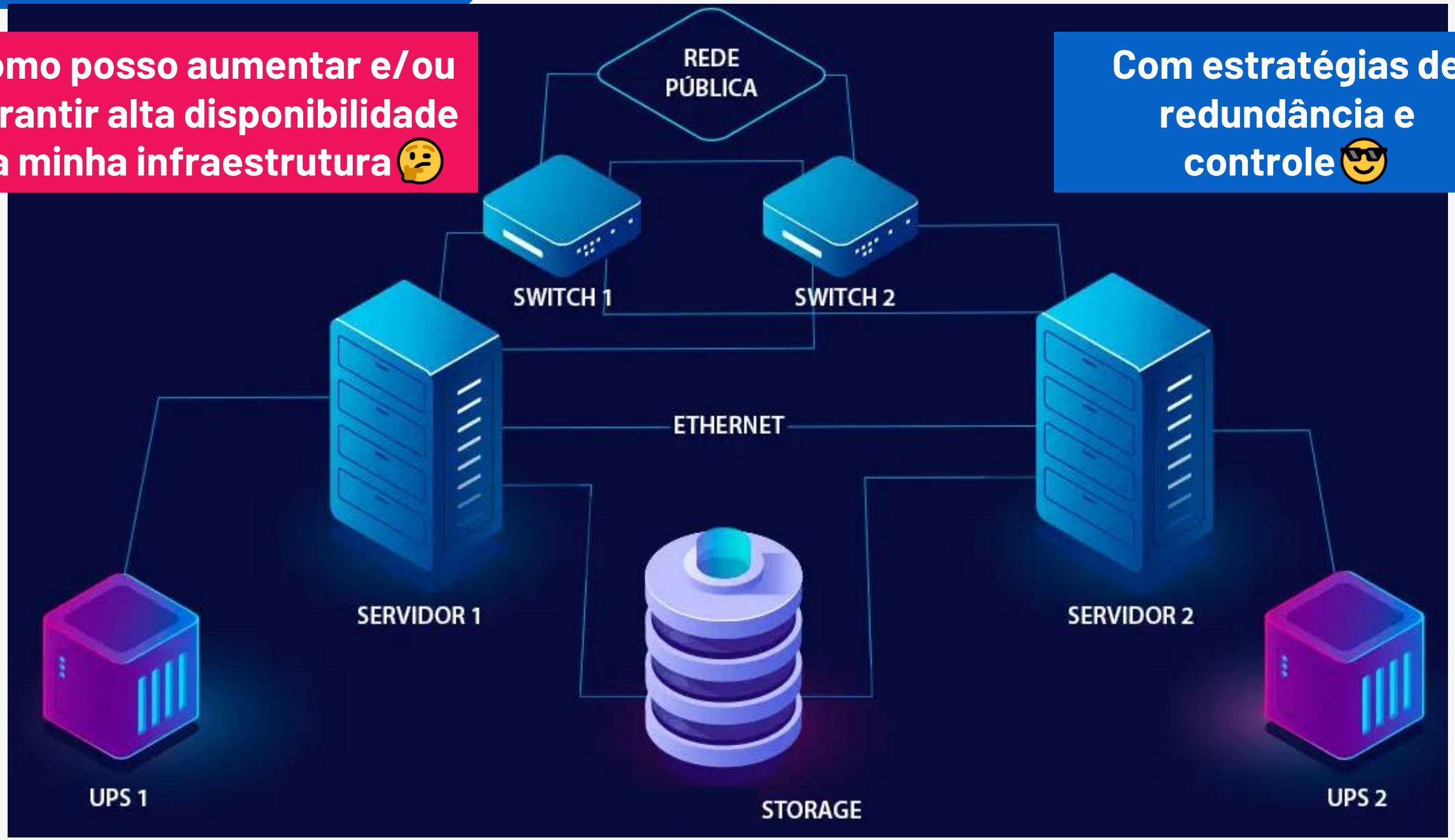
Agenda da Aula

- Quiz
- HA e Load Balance
- Intervalo
- Desafio-Atividade 09
- Apoio PI

SUA INFRA TEM HA?

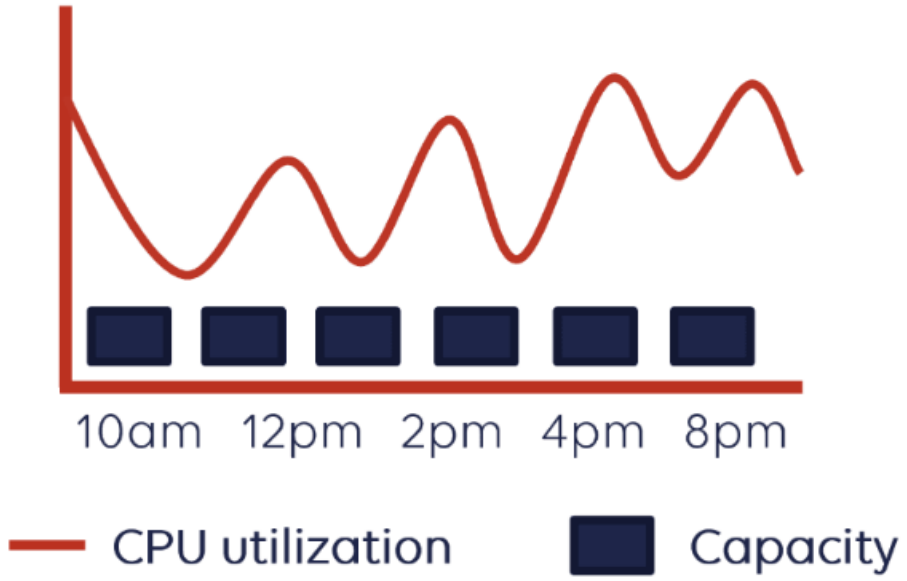
Como posso aumentar e/ou garantir alta disponibilidade na minha infraestrutura🤔

Com estratégias de redundância e controle😎

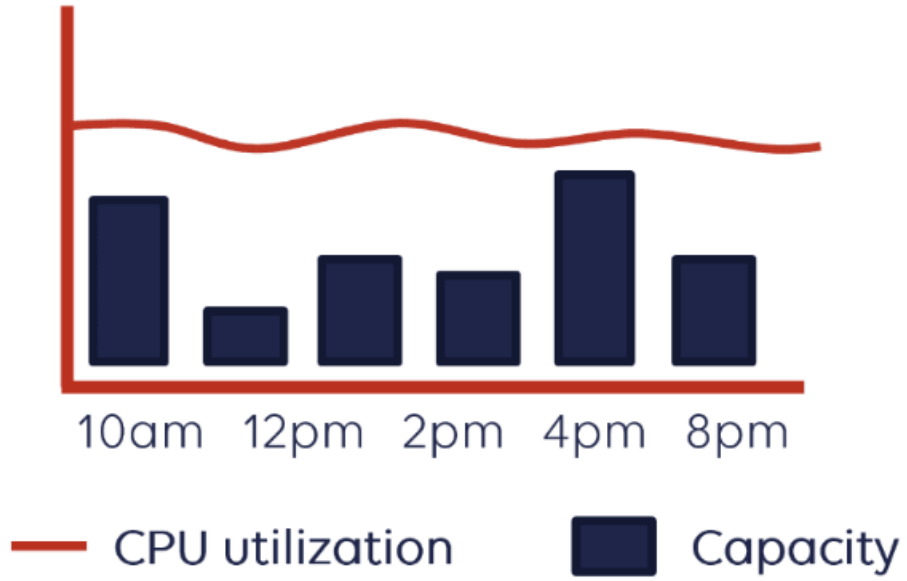


Como lidar com um consumo dinâmico da minha solução 🤔

sem escala dinâmica



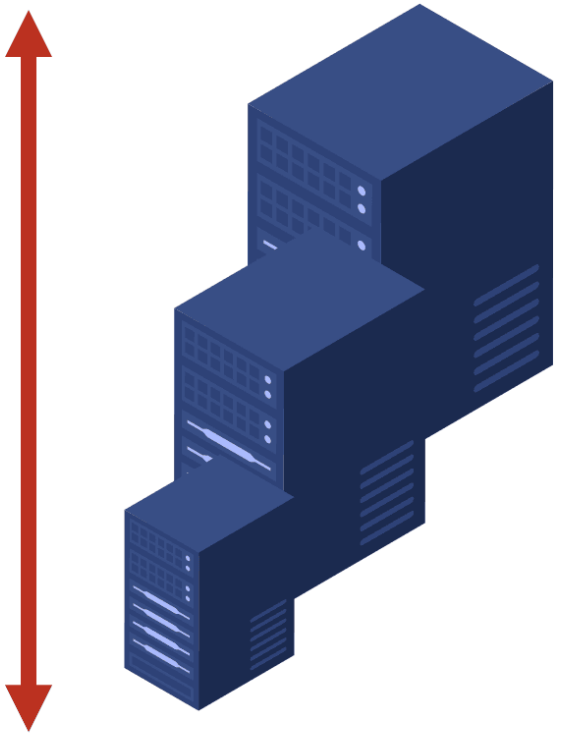
com escala dinâmica



Como escalar minha infraestrutura de forma adequada 🤔

Vertical Scaling

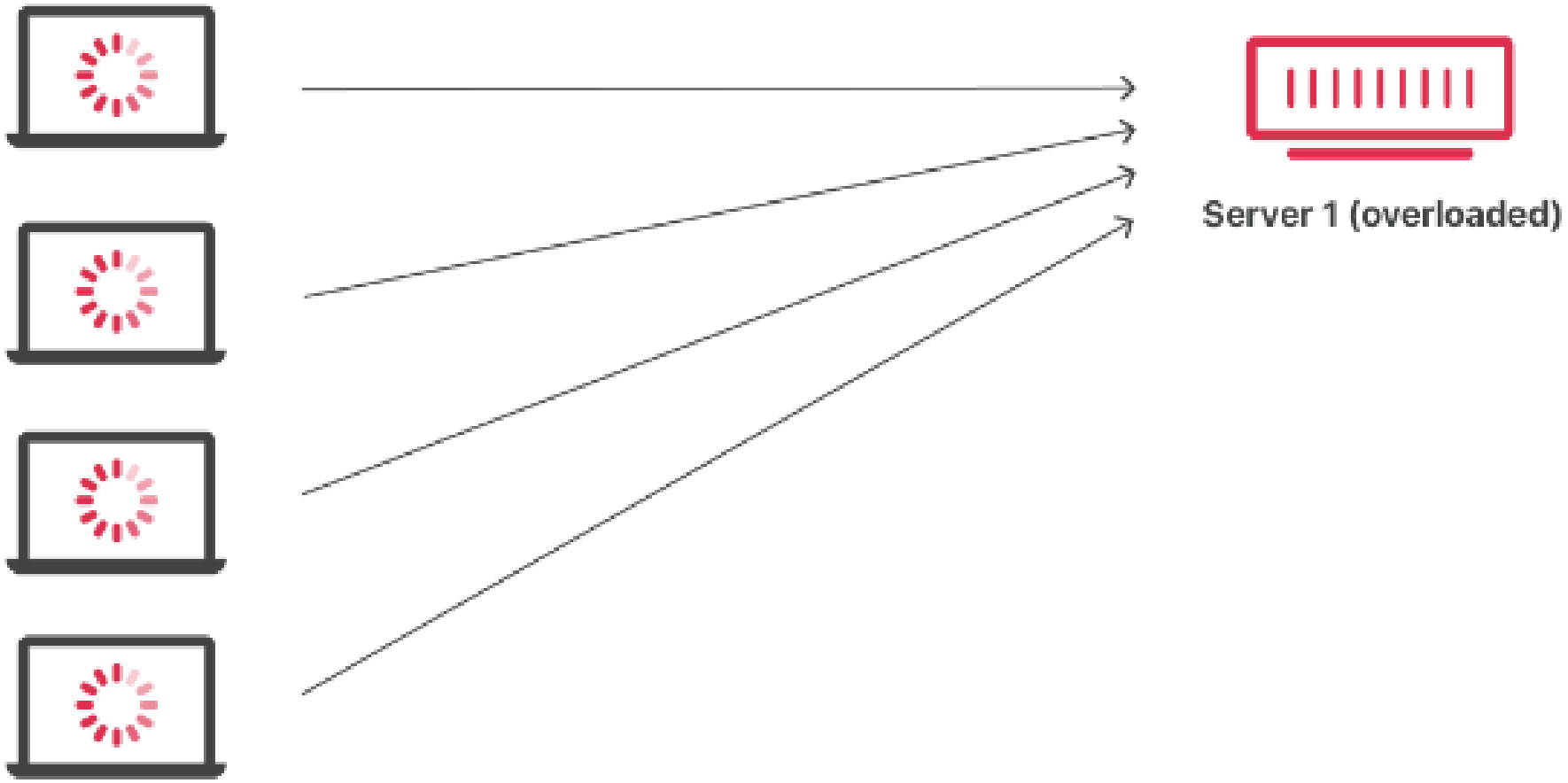
Increase or decrease the capacity of existing services/instances.



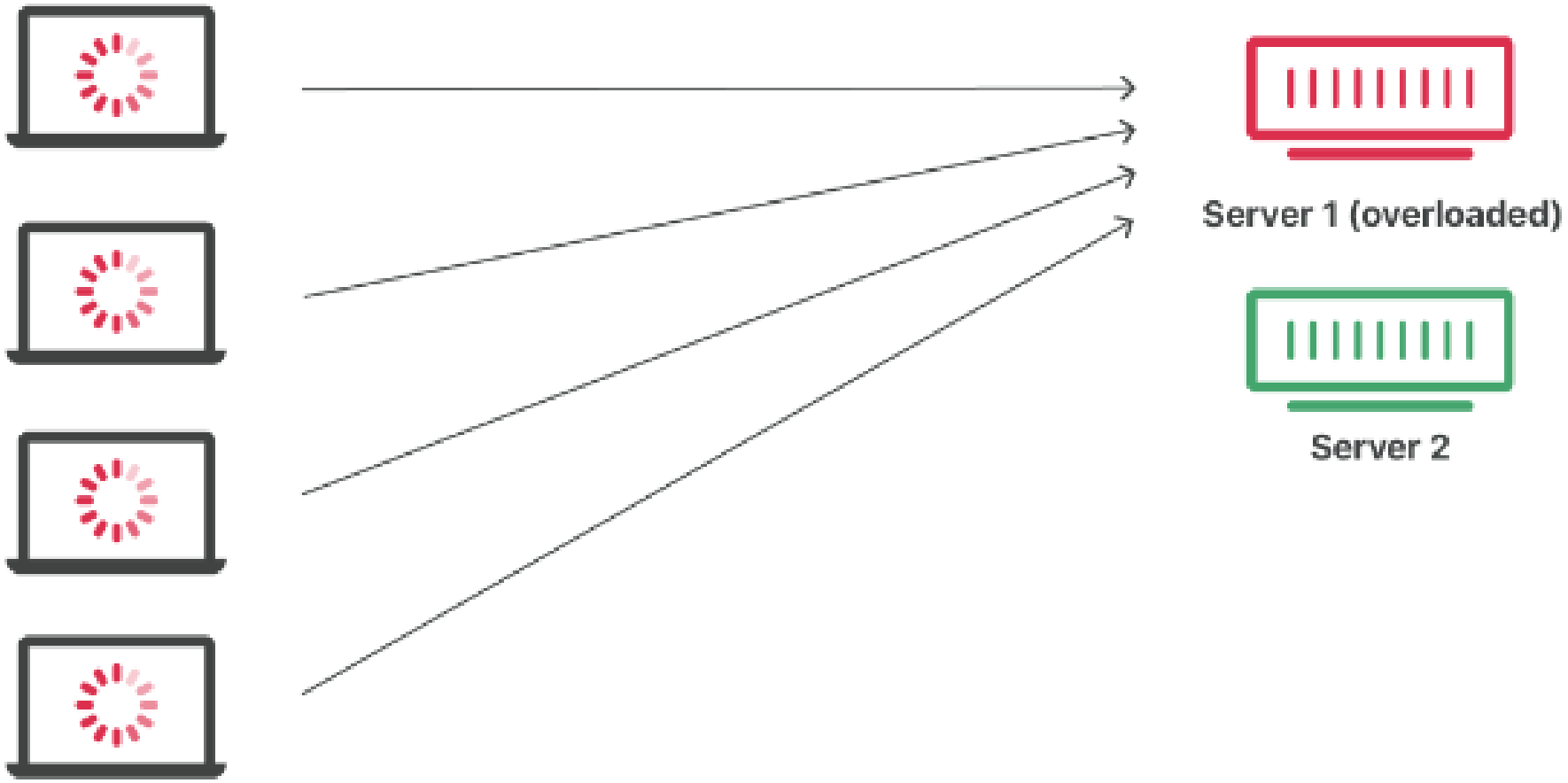
Horizontal Scaling

Add more resources like virtual machines to your system to spread out the workload across them.

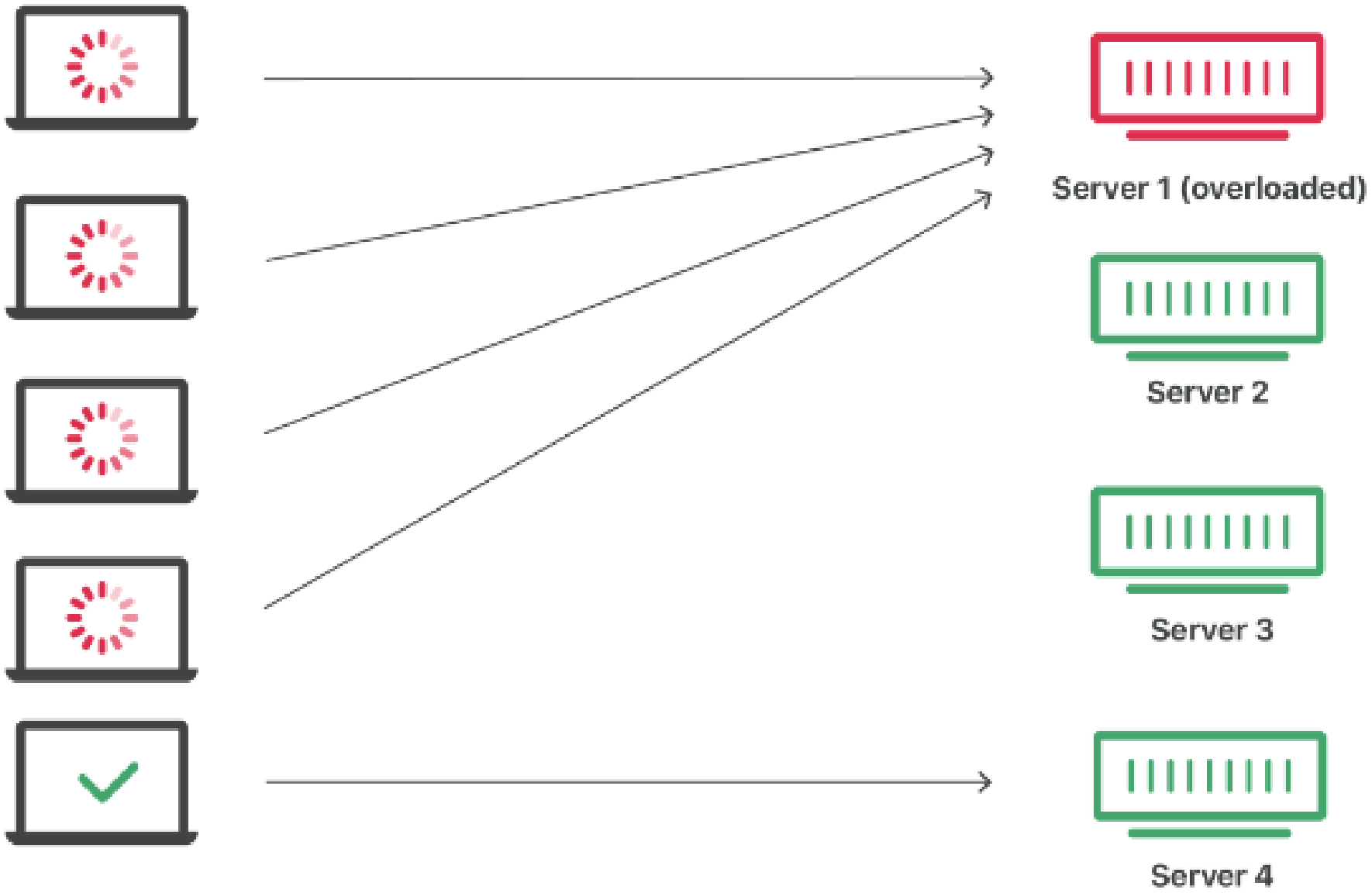
O meu servidor está sobrecarregado, e agora... 🤔



Mas as requisições não estão chegando no novo servidor... 🤔



Adiciona outro servidor 😎



Adiciona outros servidores 😎

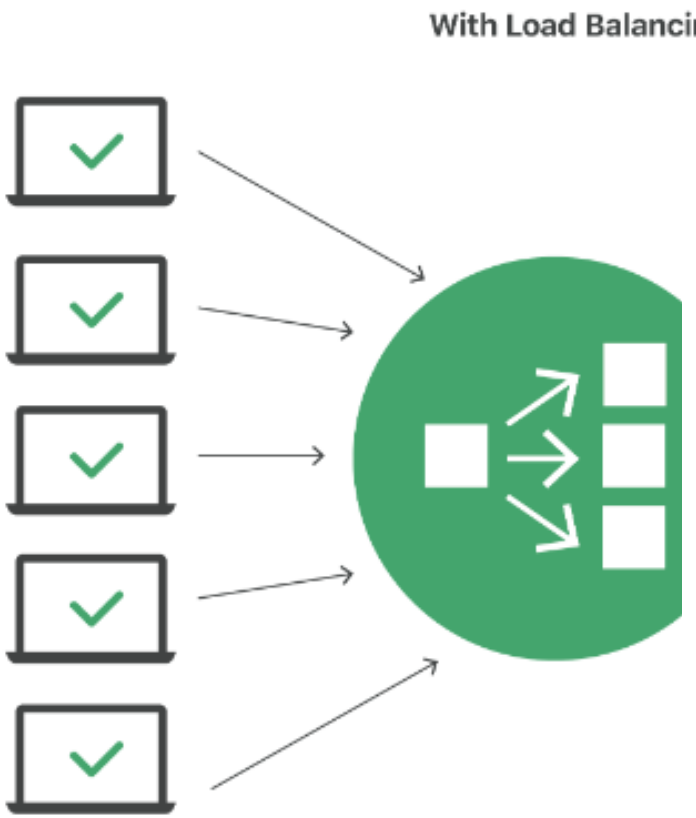
Definição

Balanceamento de carga (LB) é o **processo de distribuição eficiente do tráfego de rede entre vários servidores** para otimizar a **disponibilidade** de aplicativos e garantir uma experiência positiva para o usuário final.

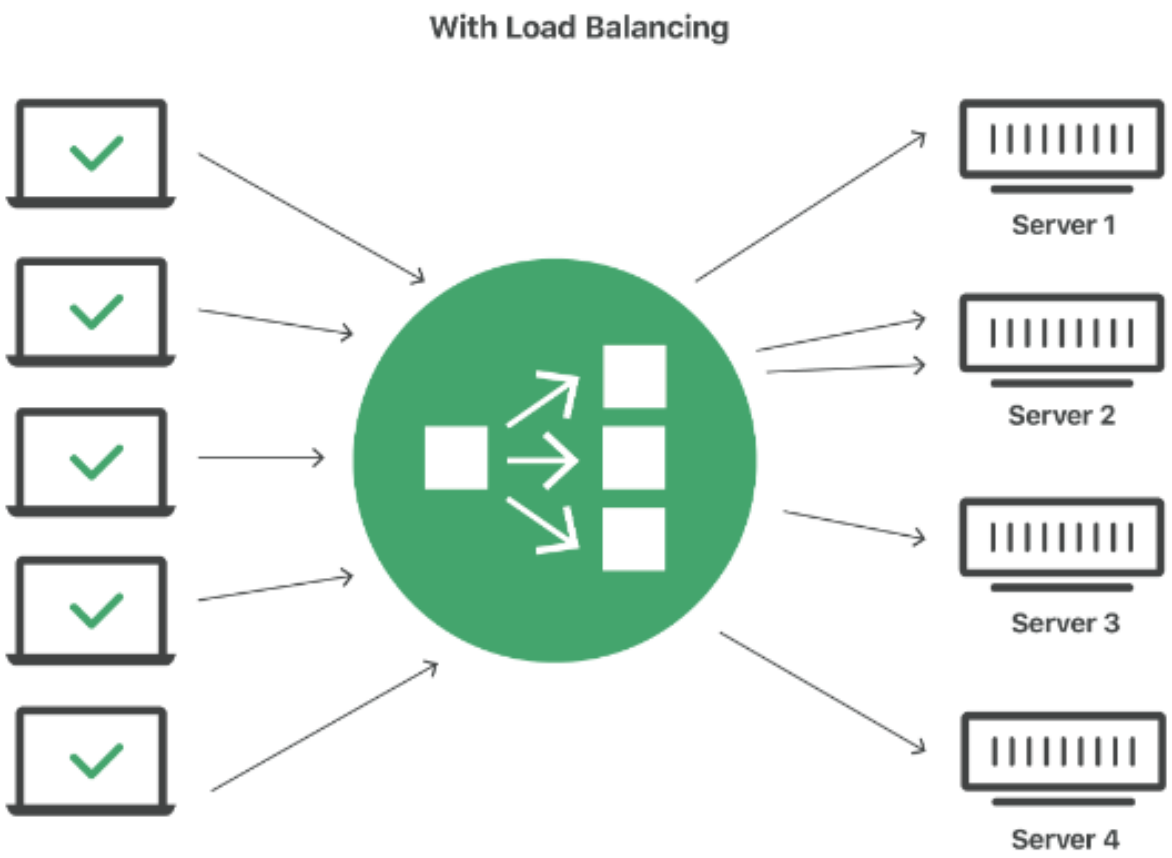
Para lidar com **volumes altos de tráfego**, a maioria das aplicações tem muitos servidores de recursos com dados duplicados entre eles.

Um balanceador de carga é um dispositivo/serviço que fica entre o usuário e o **grupo de servidores** e atua como um **facilitador invisível**, garantindo que todos os **servidores de recursos sejam usados igualmente.**

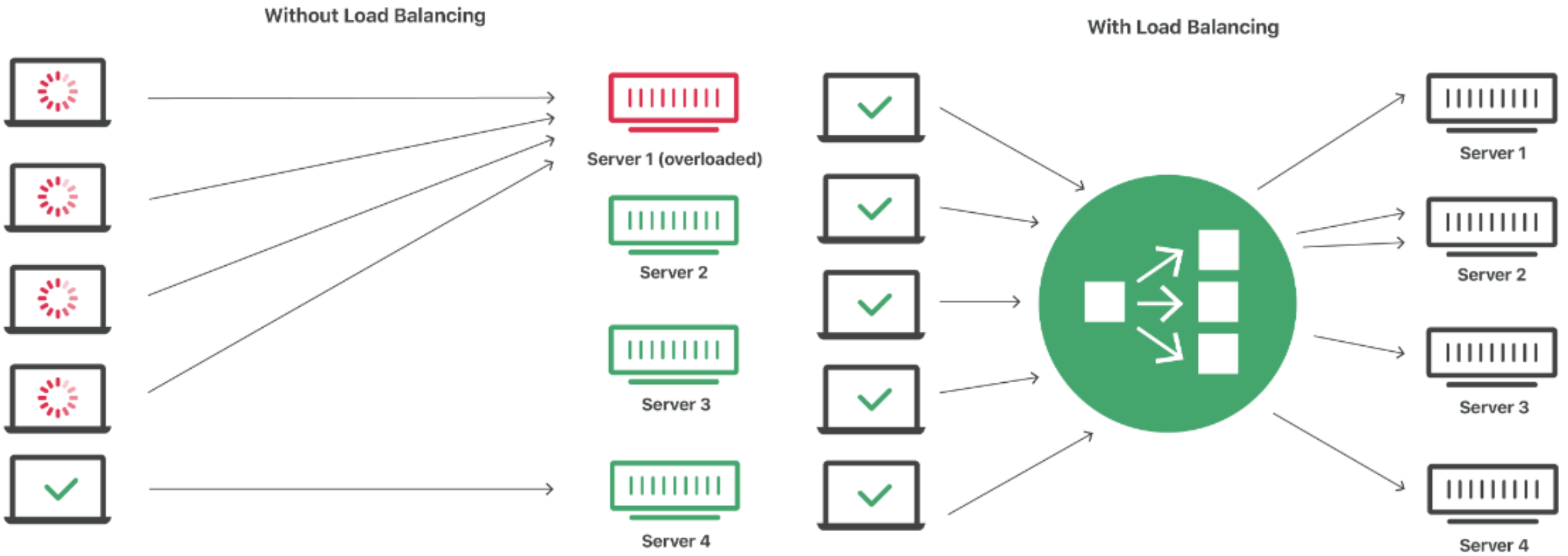
Com load balance



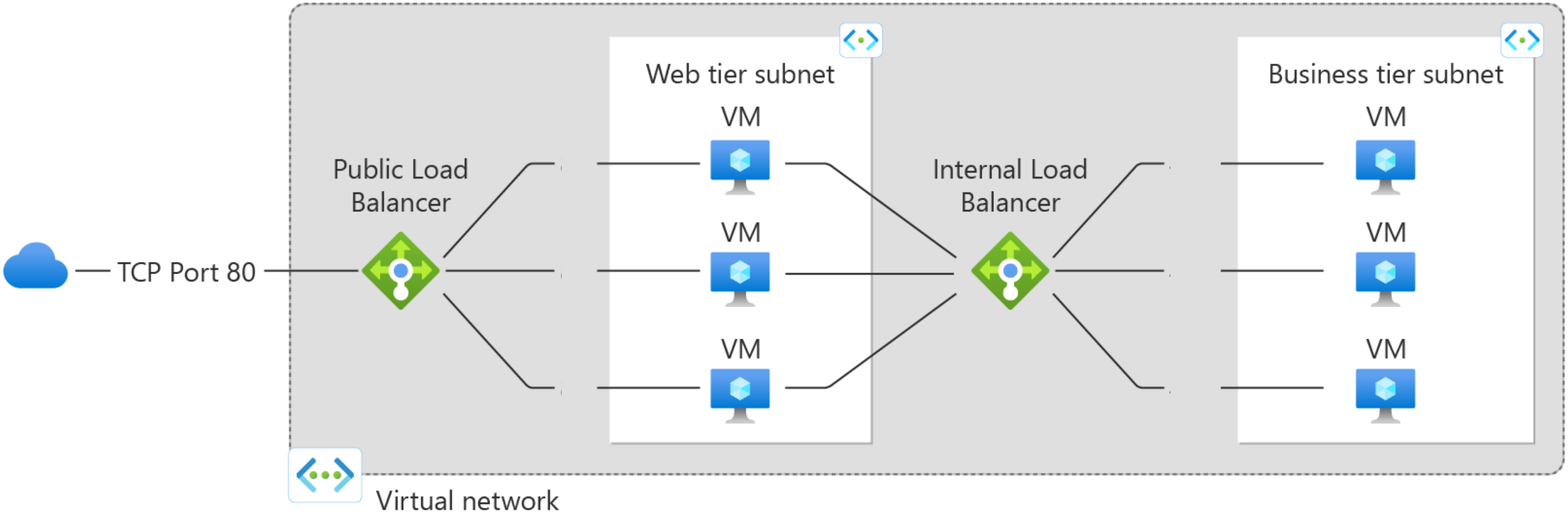
Com load balance



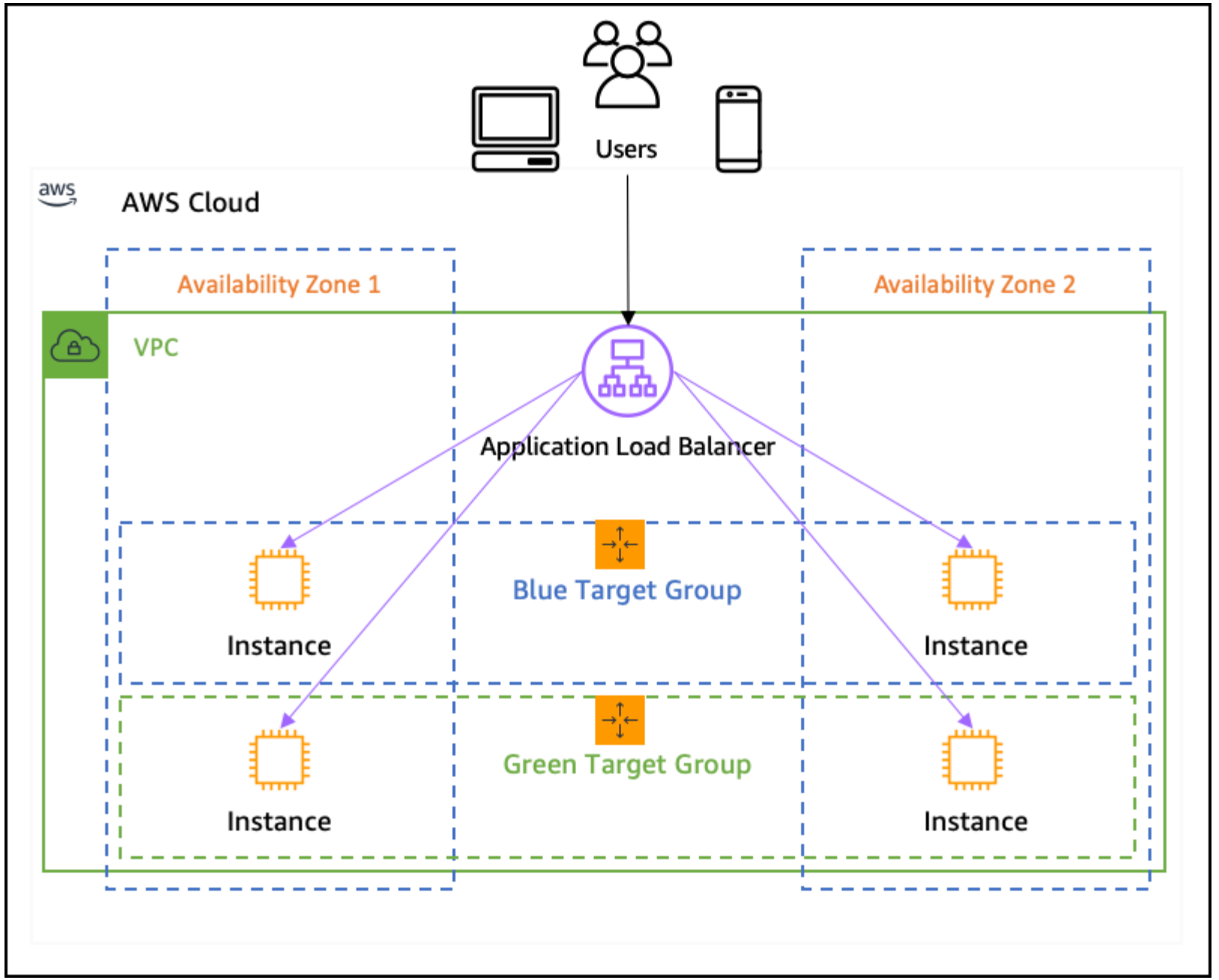
Sem load balance / Com load balance



EXEMPLO 1


Web tier subnet
Business tier subnet
VM
VM
VM
VM
VM
VM
Public Load Balancer
Internal Load Balancer
TCP Port 80
Virtual network

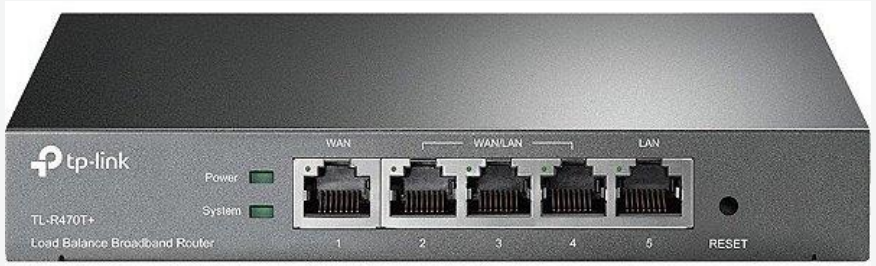
EXEMPLO 2


Users
aws
AWS Cloud
Availability Zone 1
Availability Zone 2
VPC
Application Load Balancer
Blue Target Group
Instance
Instance
Green Target Group
Instance
Instance

Como funciona o balanceamento?

O balanceamento de carga pode ser implementado de algumas maneiras.

Balanceadores de carga de hardware são dispositivos físicos instalados e mantidos no local.



Balanceadores de carga de software são aplicativos instalados em servidores ou como um serviço de nuvem gerenciada (balanceamento de carga na nuvem).

AWS ELB
Elastic Load Balancer

Load Balancers

Como funciona o balanceamento?

Os balanceadores de carga trabalham mediando as **solicitações de clientes recebidos** em tempo real e **determinando quais servidores** de frontend e backend podem processar essas solicitações da melhor maneira possível.

Para evitar que um único servidor seja **sobrecarregado**, o balanceador de carga **encaminha as solicitações** para vários **servidores disponíveis** nas instalações ou hospedados em data centers em nuvem.

Após o **servidor atribuído** receber a solicitação, ele responde ao cliente por meio do balanceador de carga.

Benefícios do balanceamento

Disponibilidade

São **realizadas verificações** de funcionamento nos servidores antes de encaminhar as solicitações para eles.

Se um servidor estiver prestes a falhar, ou estiver offline, a carga de trabalho é **redirecionada para um servidor em operação** para evitar interrupções de serviço e manter **alta disponibilidade**.

Escalabilidade

Com uma infraestrutura de alto desempenho, que pode **receber um alta cargas de tráfego** de rede.

Servidores físicos ou virtuais podem ser **adicionados ou removidos** conforme necessário, tornando a **escalabilidade simples e automatizada.**

Segurança

Podemos incluir recursos de **segurança**, como criptografia SSL, firewalls de aplicativos web (**WAF**) e autenticação multifatorial (MFA).

Ao rotear ou descarregar o tráfego de rede com segurança, o balanceamento de carga pode **ajudar a proteger contra riscos de segurança**, como ataques de distributed denial-of-service (DDoS).

Algoritmos de balanceamento

O **método** para rotear uma solicitação para um servidor específico é definido por um **algoritmo de balanceamento de carga**

- **Round-robin:** Usa o DNS (Domain Name System) para **atribuir sequencialmente solicitações** a cada servidor em uma rotação contínua. É o **método mais básico**, pois utiliza apenas o **nome ou IP de cada servidor** para determinar qual deles receberá a próxima solicitação recebida.

- **Round-robin ponderado:** Além de seu nome DNS, cada servidor nesse algoritmo também recebe um **"peso"**. O peso determina quais **servidores devem ter prioridade** sobre outros para lidar com as solicitações recebidas. Um administrador decide como cada servidor será ponderado com **base em sua capacidade e nas necessidades da rede.**

Algoritmos de balanceamento

- **Hash de IP:** Combina endereços de **IP de origem e de destino do tráfego de entrada** e usa uma função matemática para convertê-lo em um hash. Com base no hash, **a conexão é atribuída a um servidor específico.**

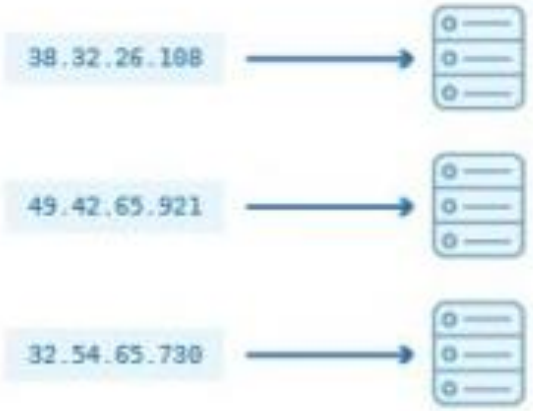
- **Menos conexões (Least connections):** Esse algoritmo dá **prioridade ao servidor com as menores conexões ativas** quando uma nova solicitação de cliente é recebida. Esse método ajuda a evitar que os servidores fiquem sobrecarregados com conexões e a manter uma carga consistente em todos os servidores o tempo todo.

Algoritmos de balanceamento

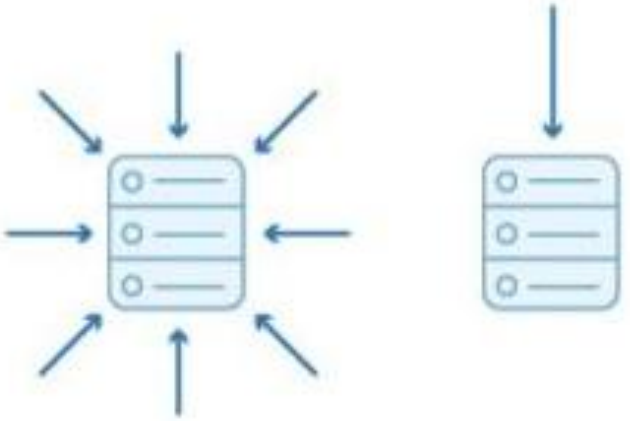
- **Menor tempo de resposta (Lest response time):** Este algoritmo **combina** o menor método de conexão com o menor tempo médio de resposta do servidor. Tanto o número de conexões quanto o tempo que leva para um servidor realizar solicitações e enviar uma resposta são avaliados. **O servidor mais rápido com menos conexões ativas receberá a solicitação recebida.**

- **Baseado em recursos:** os balanceadores de carga distribuem o tráfego **analisando a carga atual do servidor**. Um software especializado chamado **"agente" é executado em cada servidor** e calcula o uso de recursos do servidor, como sua capacidade de computação e memória. Em seguida, o **balanceador de carga verifica se há recursos livres suficientes** no agente antes de distribuir o tráfego para esse servidor.

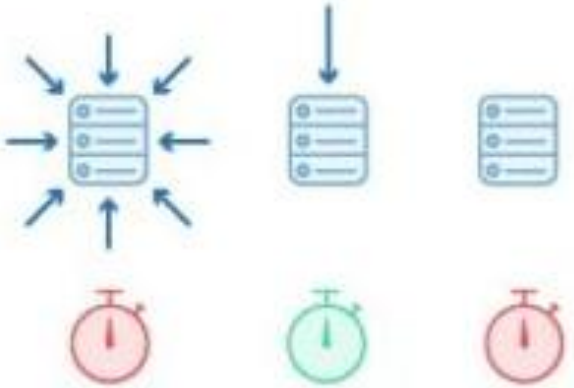
Round Robin



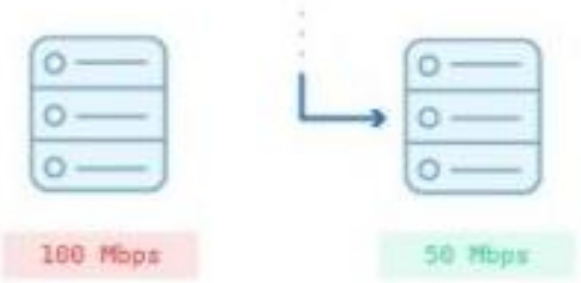
IP Hash



Least Connections



Least Response Time



Least Bandwidth

Tipos de balanceamento de carga

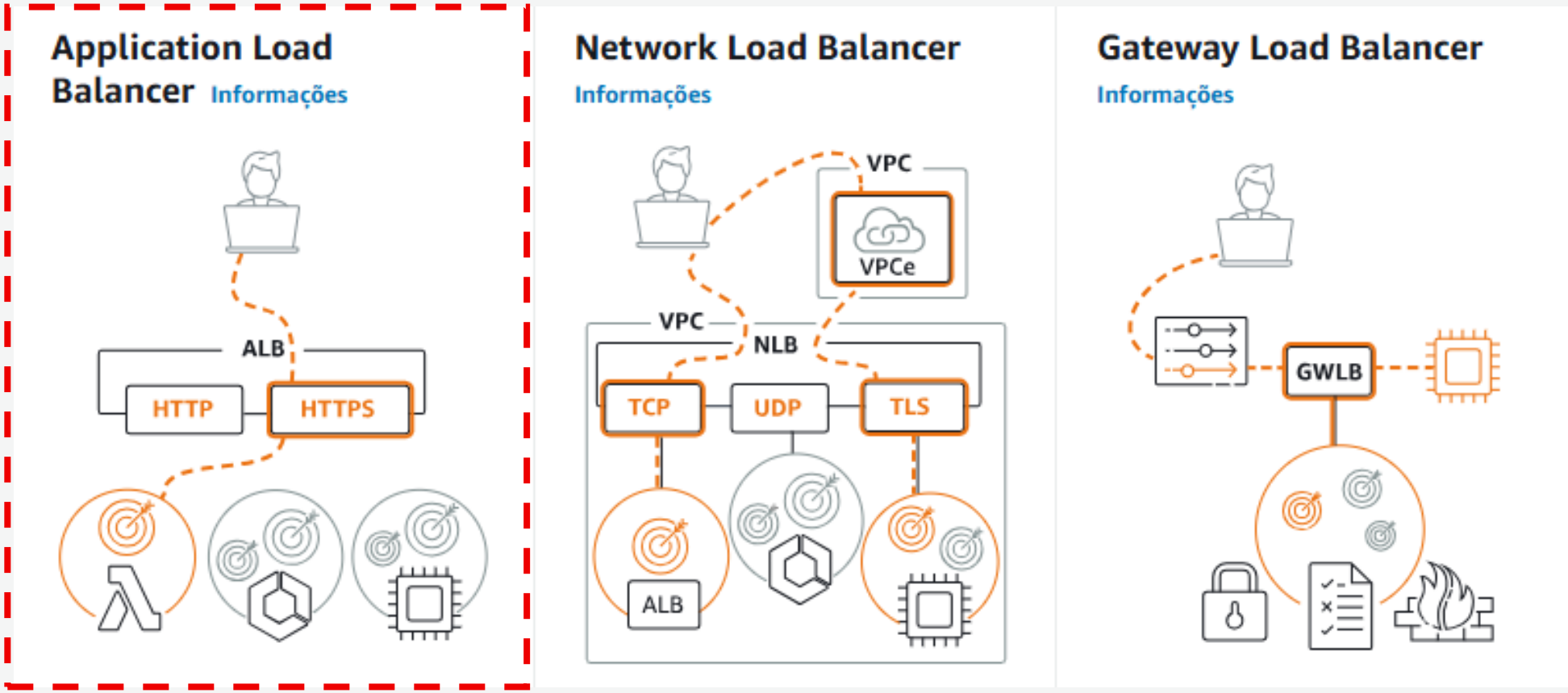
Podemos classificar o balanceamento de carga em **três categorias principais**, dependendo do que o balanceador de carga verifica na solicitação do cliente para redirecionar o tráfego.

- **Balanceamento de carga de aplicações:** eles examinam o **conteúdo da solicitação**, como cabeçalhos HTTP para redirecionar o tráfego.

- **Balanceamento de carga de rede:** eles **examinam endereços IP e outras informações de rede** para redirecionar o tráfego de maneira **ideal**.

- **Balanceamento de carga de DNS:** nele você configura seu domínio para rotear solicitações de rede em um **grupo de recursos no seu domínio**. Um domínio pode corresponder a um site, um sistema de correio, um servidor de impressão ou outro **serviço** acessível pela Internet.

Tipos de balanceamento de carga



7º Camada OSI (Aplicação)	4º Camada OSI (Transporte)	3º Camada OSI (Rede)

Agradeço
a sua atenção!

Marcio Santana

marcio.santana@sptech.school

SÃO PAULO TECH SCHOOL