01

Balanceando a carga

Transcrição

Escalamos nossa aplicação criando mais uma máquina. Com duas máquinas garantirmos maior disponibilidade do nosso projeto. Contudo, temos duas URL's. E na prática para nosso cliente, como fica?

Não podemos simplesmente dizer para ele: *"Olha meu camarada, quando o sistema tiver fora em umas das URL's, basta você acessar outra".* Precisamos de algo mais transparente.

Load Balancer

Uma solução é, em vez de acessarmos diretamente essas URL's, podemos acessar um cara intermediário. É ele que decidirá qual máquina (URL) deve ser acessada. Algumas decisões podem ser tomadas, por exemplo, acessar a máquina que estive menos sobrecarregada, ou a única disponível, que ainda esta de pé. Essa solução consiste no uso de um **load balancer**.

Existem várias implementações de um load balancer (HAProxy, Nginx, Serviço da AWS). Em nosso caso, utilizaremos um próprio serviço da Amazon que cria o load balancer para nós.