



Sistemas de Informação Distribuídos

Professores:

Fernando Antonio Mota Trinta

Paulo Antonio Leal Rego



Desenvolvimento de Software para Nuvem

Professores:

Fernando Antonio Mota Trinta

Paulo Antonio Leal Rego



Tópicos Especiais em Redes e Sistemas Distribuídos II

Professores:

Fernando Antonio Mota Trinta

Paulo Antonio Leal Rego



Elasticidade

Professores:

Fernando Antonio Mota Trinta

Paulo Antonio Leal Rego

Relembrando definição de Cloud Computing

- A Nuvem é um grande repositório de recursos virtualizados facilmente utilizáveis e acessíveis (como hardware, plataformas de desenvolvimento e/ou serviços). Esses recursos podem ser dinamicamente reconfigurados para ajustar a carga (escala) variável do sistema, permitindo também um uso ótimo dos recursos. Esse reservatório de recursos é geralmente explorado por um modelo pay-per-use (pagar para usar) no qual as garantias são oferecidas por um Provedor de Infraestrutura por meio de SLAs (Service Level Agreements - Acordo de Nível de Serviço)

Vaquero, L.M. and Roderio-Merino, L. and Caceres, J. and Lindner, M. "A break in the clouds: towards a cloud definition" em ACM SIGCOMM Computer Communication Review, 2008.

Relembrando definição de Cloud Computing

- A Nuvem é um grande repositório de recursos virtualizados facilmente utilizáveis e acessíveis (como hardware, plataformas de desenvolvimento e/ou serviços). Esses recursos podem ser dinamicamente reconfigurados **para ajustar a carga (escala) variável do sistema, permitindo também um uso ótimo dos recursos**. Esse reservatório de recursos é geralmente explorado por um modelo pay-per-use (pagar para usar) no qual as garantias são oferecidas por um Provedor de Infraestrutura por meio de SLAs (Service Level Agreements - Acordo de Nível de Serviço)

Vaquero, L.M. and Roderio-Merino, L. and Caceres, J. and Lindner, M. "A break in the clouds: towards a cloud definition" em ACM SIGCOMM Computer Communication Review, 2008.

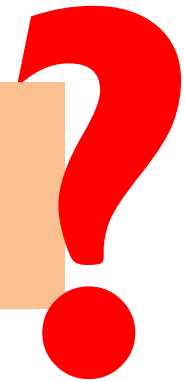
Características Essenciais da Nuvem

- Auto-serviço sob demanda
- Acesso amplo a serviços
- Pooling de recursos
- Serviço medido
- Elasticidade rápida

Características Essenciais da Nuvem

- Auto-serviço sob demanda
- Acesso amplo a serviços
- Pooling de recursos
- Serviço medido
- Elasticidade rápida

“Mas o que seria formalmente elasticidade”



Elasticidade em Outros Domínios

- Na Física
 - Ramo da física que estuda o comportamento de corpos materiais que se deformam ao serem submetidos a ações externas (forças devidas ao contato com outros corpos, ação gravitacional agindo sobre sua massa, etc.), retornando à sua forma original quando a ação externa é removida.
 - Propriedade que um objeto ou material voltar à sua forma natural depois de ter sido comprimida ou esticada

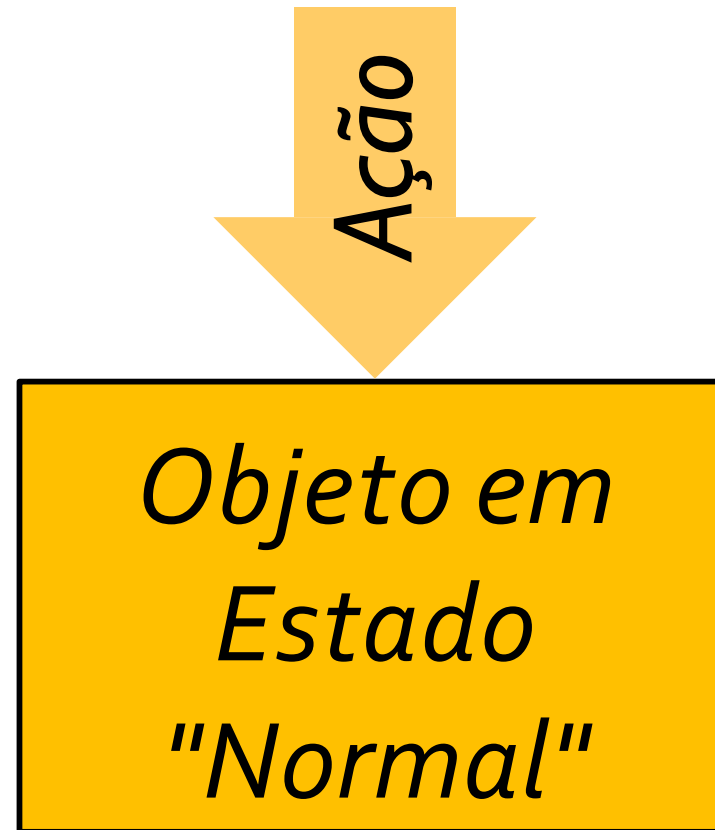
Elasticidade em Outros Domínios

- Na Economia
 - O tamanho do impacto que a alteração em uma variável (ex.: preço) exerce sobre outra variável (ex.: demanda);
 - Em sentido genérico, é a alteração percentual de uma variável, dada a alteração percentual em outra;
 - Sinônimo de sensibilidade;
 - Resposta, reação de uma variável, em face de mudanças em outras variáveis.

Em comum...

*Objeto em
Estado
"Normal"*

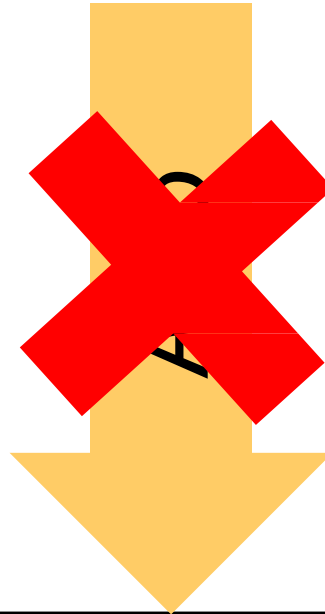
Em comum...



Em comum...



Em comum...

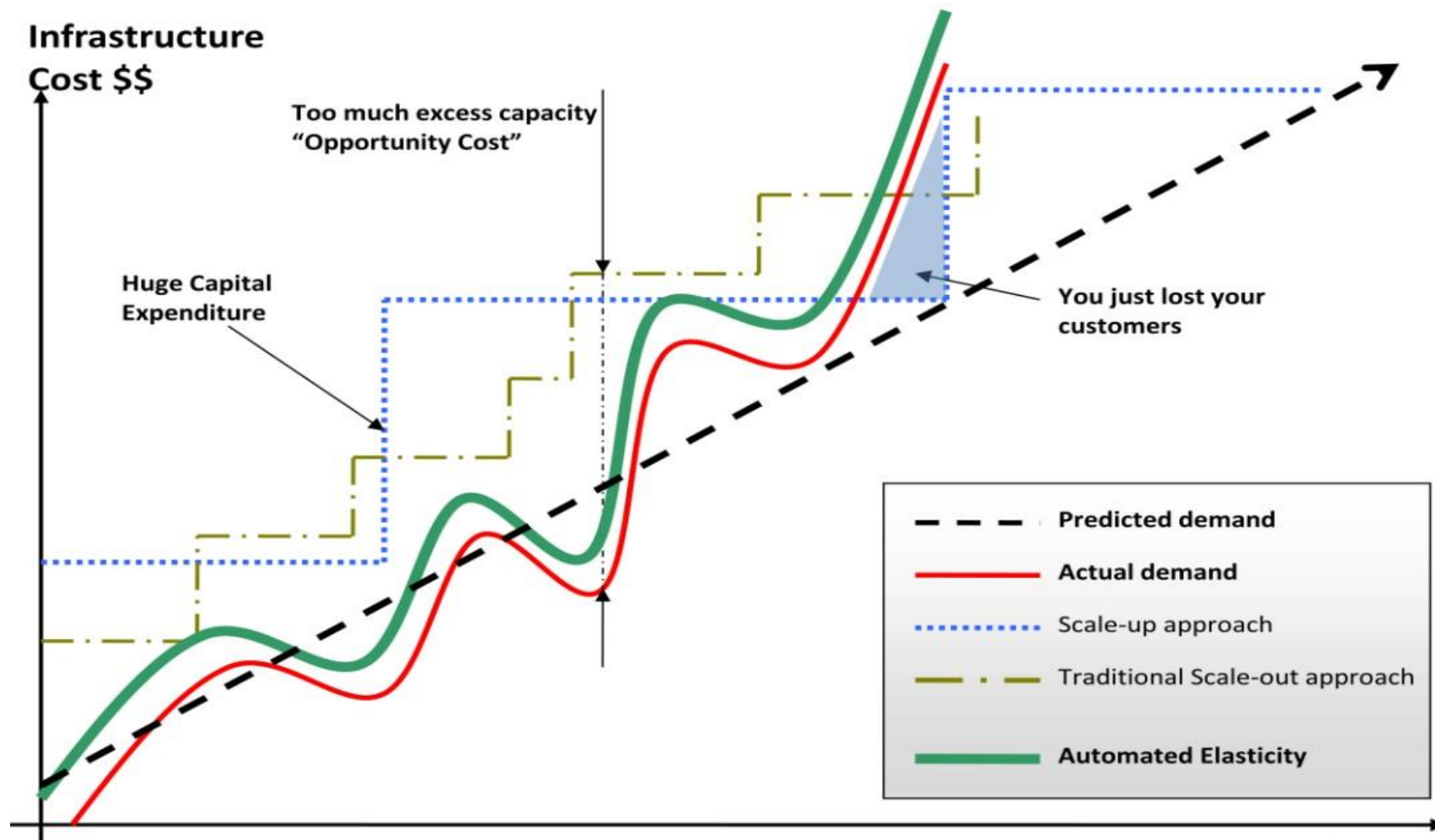


*Objeto em Estado
"Alterado"*

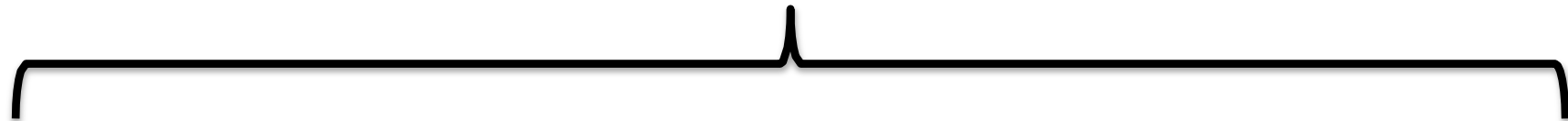
Em comum...

*Objeto em
Estado
"Normal"*

Ações e Deformidades para TI



Elasticidade em TI

Elasticidade*Implementar serviços
com QoS para nuvem*

Permite adicionar ou remover recursos, sem interrupções e em tempo de execução para lidar com a variação da carga



Recursos podem ser adquiridos de forma rápida e liberados Recursos disponíveis para uso parecem ser ilimitados Podem ser adquiridos em qualquer quantidade e a qualquer momento

*Aumento e
retração da
demanda*

Definição, Segundo NIST

- “...habilidade de rápido provisionamento e desprovisionamento, com capacidade de recursos virtuais praticamente infinita e quantidade adquirível sem restrição a qualquer momento...”

MELL, P.; GRANCE, T. The NIST definition of cloud computing. National Institute of Standards and Technology, NIST, v. 53, n. 6, p. 50, 2009.

Muitas Outras Definições

Cooper et al. (2010)	Capacidade de adicionar novas instâncias e distribuir a carga de trabalho para estas instâncias
Fito, Goiri e Guitart (2010)	Habilidade de adquirir e liberar recursos de granularidades variadas de acordo com a carga da trabalho em um curto intervalo de tempo
Aisopos, Tserpes e Varvarigou (2011)	Capacidade do provedor alterar dinamicamente a quantidade de recursos de CPU, memória e espaço em disco para uma determinada tarefa
Espadas et al. (2011)	Habilidade de criar um número variável de instâncias de máquinas virtuais que dependem da demanda da aplicação
Li et al. (2011)	Habilidade do sistema de se adaptar à mudanças repentinas na carga de trabalho

Muitas Outras Definições

Perez-Sorrosal et al. (2011)	Capacidade de adicionar novas instâncias e distribuir a carga de trabalho para estas instâncias
Garg, Versteeg e Buyya (2011, 2012)	Habilidade de adquirir e liberar recursos de granularidades variadas de acordo com a carga da trabalho em um curto intervalo de tempo
Han et al. (2012)	Capacidade do provedor alterar dinamicamente a quantidade de recursos de CPU, memória e espaço em disco para uma determinada tarefa
Islam et al. (2012)	Habilidade de criar um número variável de instâncias de máquinas virtuais que dependem da demanda da aplicação
Pandey et al. (2012)	Habilidade do sistema de se adaptar à mudanças repentinas na carga de trabalho

Uma Definição Mais Moderna

- Grau no qual um sistema é capaz de se adaptar à variações na carga de trabalho pelo provisionamento e desprovisionamento de recursos de maneira autônômica, de modo que em cada ponto no tempo os recursos disponíveis combinem com a demanda da carga de trabalho o mais próximo possível.

Nikolas Roman Herbst, Samuel Kounev, and Ralf Reussner. Elasticity in Cloud Computing: What It Is, and What It Is Not. ICAC'2013

Elasticidade vs Escalabilidade

- Escalabilidade
 - Habilidade que um sistema computacional tem de sustentar (ou suportar) o aumento na carga de trabalho com um desempenho adequado, a partir da adição de recursos
- Certa confusão entre os termos
 - Certos aspectos não são considerados ao falar sobre escalabilidade
 - Resposta automática? Tempo de adaptação?

Dimensões e Aspectos Importantes

- Elasticidade é um processo adaptativo
 - Mudança de estado
- Aspectos importantes
 - Velocidade
 - Definido pelo tempo necessário para sair de um estado “de falta de recursos” para um estado ótimo, ou de “overprovisioned”
 - Precisão
 - Definido com o desvio absoluto da quantidade atual de recursos alocados da demanda atual de recursos.

Elasticidade

- Elasticidade = Escalabilidade + **alta velocidade** + **alta precisão**
- Em outras palavras
 - Escalabilidade é um requisito para elasticidade
 - Não leva em consideração aspectos temporais
 - Não leva em consideração se os recursos utilizados para suporte a aumento de cargas de trabalho é otimizado

Benefícios da Elasticidade

- Mantem sistema em execução mesmo quando o volume de acesso aumenta de forma inesperada
- Diminuição de custos, evitando manter servidores a mais ou mais potentes que o necessário

Como Conseguir Elasticidade?

- Existem diferentes abordagens para se conseguir escalabilidade/elasticidade
- Duas abordagens principais: escala horizontal e/ou escala vertical

Scale up – Nó mais poderoso

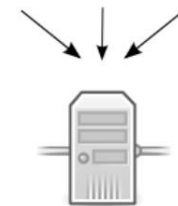
Service requests



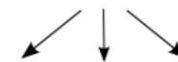
Server

Scale up – mais nós

Service requests

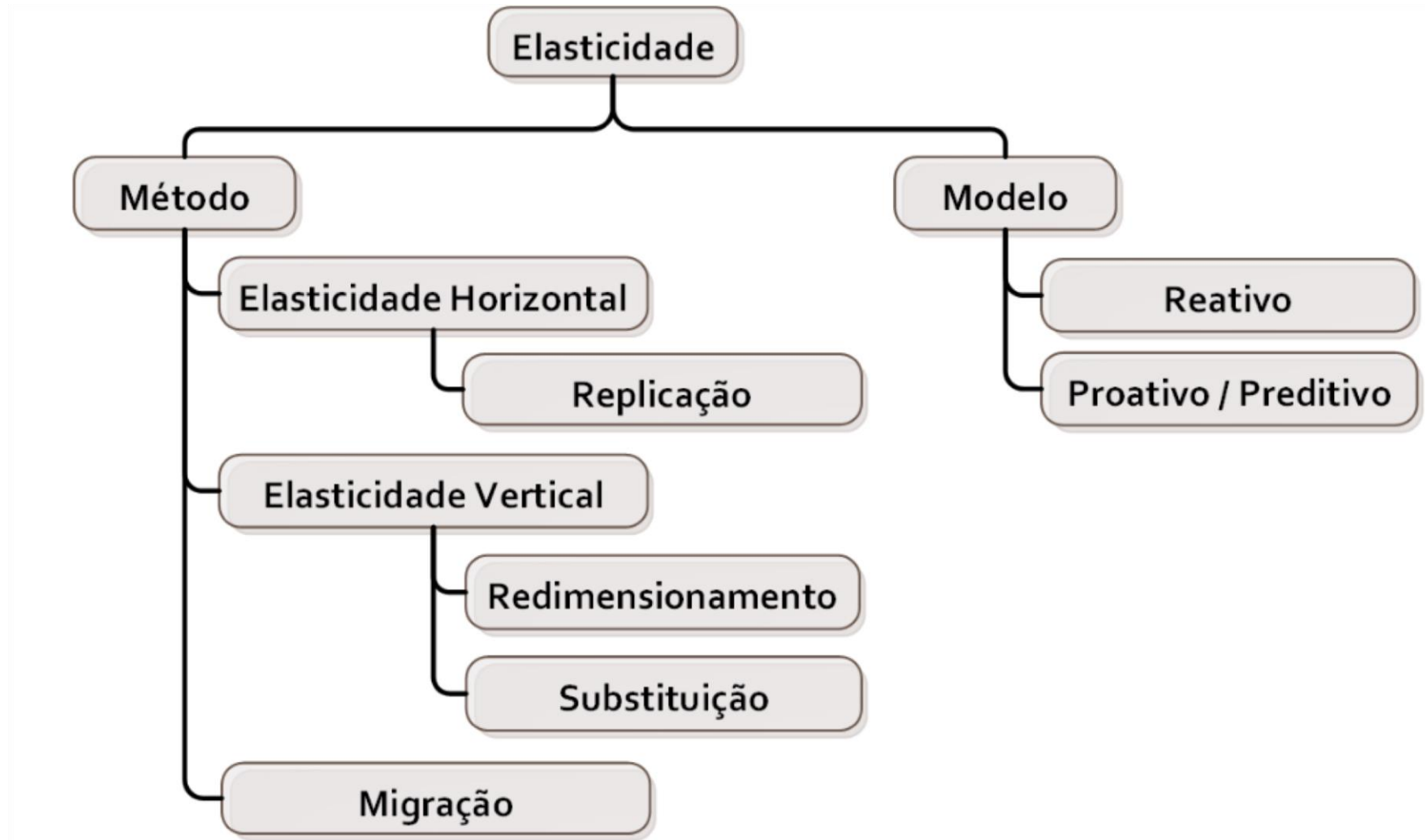


Load balancer



Server pool

Métodos e Modelos



Métodos (1/3)

- Elasticidade Horizontal (Replicação)
 - Adicionar/Remover instâncias
 - Aplicações, containers ou máquinas virtuais
 - Método mais utilizado para se alcançar elasticidade

Métodos (2/3)

- Elasticidade Vertical
 - Adicionar/Remover Recursos em uma máquina, virtual ou não
 - mais CPU, mais memória, mais armazenamento
 - Duas abordagens
 - Redimensionamento
 - Alteração em tempo de execução dos recursos associados à instância em execução
 - Substituição
 - servidores mais potentes na substituição de servidores menos potentes

Métodos (3/3)

- Migração

- transferência de máquinas virtuais ou aplicações que estão sendo executadas de um servidor físico para outro
- Objetivos: consolidação/balanceamento de carga

Modelos

- Reativo
 - Reage à carga de trabalho atual e utiliza limiares da utilização dos recursos ou violações de SLA para disparar a necessidade de capacidade adicional
 - O sistema reage à mudanças, mas não as antecipa

Modelos

- Proativo (preditivo)
 - técnicas de predição de cargas de trabalho para determinar quando a carga de trabalho futura irá superar a capacidade atual provisionada
 - Uso algoritmos de alocação adicional de servidores antes que sua capacidade seja excedida.
 - Uso de informações históricas

Métodos vs Modelos

- Várias combinações possíveis
 - Replicação Reativa
 - Dependência do monitoramento
 - Migração Reativa
 - Redimensionamento Reativo
 - Replicação Proativa
 - Migração Proativa

Dúvidas

