Introdução a Computação em Nuvem

Conceitos teóricos e práticos, evolução e novas possibilidades

Daniel Cordeiro
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

CESUPA – Belém/PA – maio de 2013

Olá! :)

Um pouco sobre mim

- Pós-doutorando no IME/USP
- *Docteur en Mathématiques et en Informatique* – Université de Grenoble, França, 2012
- Mestre em Computação – Universidade de São Paulo, 2006

Para mais informações:

`http://www.ime.usp.br/~danielc/`

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todos já devem ter ouvido algo sobre Cloud Computing

Ou ao menos algumas dessas ideias:

- "Computação em nuvem finalmente tornou realidade o sonho da computação utilitária"
- "Desenvolvedores não precisam mais se preocupar em conseguir grandes somas de dinheiro antes de colocar uma nova aplicação web no ar"
- "Adeus aos problemas de provisionamento de servidores (Elasticidade dos recursos)"
- Software como um serviço
- Plataforma como um serviço
- Infraestrutura como um serviço

Todo mundo fala sobre Computação em Nuvem, mas o que é isso?

The interesting thing about Cloud Computing is that we've redefined Cloud Computing to include everything that we already do (...) I don't understand what we would do differently in the light of Cloud Computing other than change the wording of some of our ads.

Larry Ellison (CEO da Oracle), The Wall Street Journal, 26 de setembro de 2008

Todo mundo fala sobre Computação em Nuvem, mas o que é isso?

A lot of people are jumping on the [cloud] bandwagon, but I have not heard two people say the same thing about it. There are multiple definitions out there of "the cloud."

Andy Isherwood (vice-presidente de vendas da HP na Europa), ZDnet News, 11 de dezembro de 2008

Todo mundo fala sobre Computação em Nuvem, mas o que é isso?

It's stupidity. It's worse than stupidity: it's a marketing hype campaign. Somebody is saying this is inevitable — and whenever you hear somebody saying that, it's very likely to be a set of businesses campaigning to make it true.

Richard Stallman (Free Software Foundation), The Guardian, 29 de setembro de 2008

Histórico e motivações

Quatro problemas que (ainda) requerem constante inovação tecnológica

- Problemas "em escala da web"
- **Grandes** *data centers*
- Computação paralela e distribuída
- Aplicações web interativas

Problemas em escala da web

- Características
  - Definitivamente *data-intensive*
  - Mas podem também ser *processing-intensive*
- Exemplos:
  - *Crawling*, indexação, busca, mineração de dados da web
  - Pesquisa em biologia computacional na era "pós-genômica"
  - Processamento de dados científicos (física, astronomia, etc.)
  - Redes de sensores
  - Aplicações Web 2.0
  - etc.

De qual volume de dados estamos falando?

Problemas da ordem de petabytes!

$$
\begin{aligned}
1 \text{ PB} &= 1.000.000.000.000.000 \text{ B} \\
&= 1.000^5 \text{ B} \\
&= 10^{15} \text{ B} \\
&= 1 \text{ milhão de gigabytes} \\
&= 1 \text{ mil terabytes}
\end{aligned}
$$

De qual volume de dados estamos falando?

Muitos, mas **muitos** dados

- O Google processa cerca de **20 petabytes** de dados por dia (2008)
- O Wayback Machine tem cerca de **3 petabytes + 100 terabytes/dia** (mar/2009)
- O Facebook tem cerca de **2,5 petabytes de dados de usuários + 15 terabytes/dia** (abr/2009)
- O site eBay tem cerca de **6,5 petabytes de dados dos usuários + 50 terabytes/dia** (mai/2009)
- O Grande Colisor de Hádrons do CERN irá gerar cerca de **15 petabytes/ano**

De qual volume de dados estamos falando?

Muitos, mas **muitos** dados

- O Google processa cerca de **20 petabytes** de dados por dia (2008)
- O Wayback Machine tem cerca de **3 petabytes + 100 terabytes/dia** (mar/2009)
- O Facebook tem cerca de **2,5 petabytes de dados de usuários + 15 terabytes/dia** (abr/2009)
- O site eBay tem cerca de **6,5 petabytes de dados dos usuários + 50 terabytes/dia** (mai/2009)
- O Grande Colisor de Hádrons do CERN irá gerar cerca de **15 petabytes/ano**

De qual volume de dados estamos falando?

Muitos, mas **muitos** dados

- O Google processa cerca de **20 petabytes** de dados por dia (2008)
- O Wayback Machine tem cerca de **3 petabytes + 100 terabytes/dia** (mar/2009)
- O Facebook tem cerca de **2,5 petabytes de dados de usuários + 15 terabytes/dia** (abr/2009)
- O site eBay tem cerca de **6,5 petabytes de dados dos usuários + 50 terabytes/dia** (mai/2009)
- O Grande Colisor de Hádrons do CERN irá gerar cerca de **15 petabytes/ano**

De qual volume de dados estamos falando?

Muitos, mas **muitos** dados

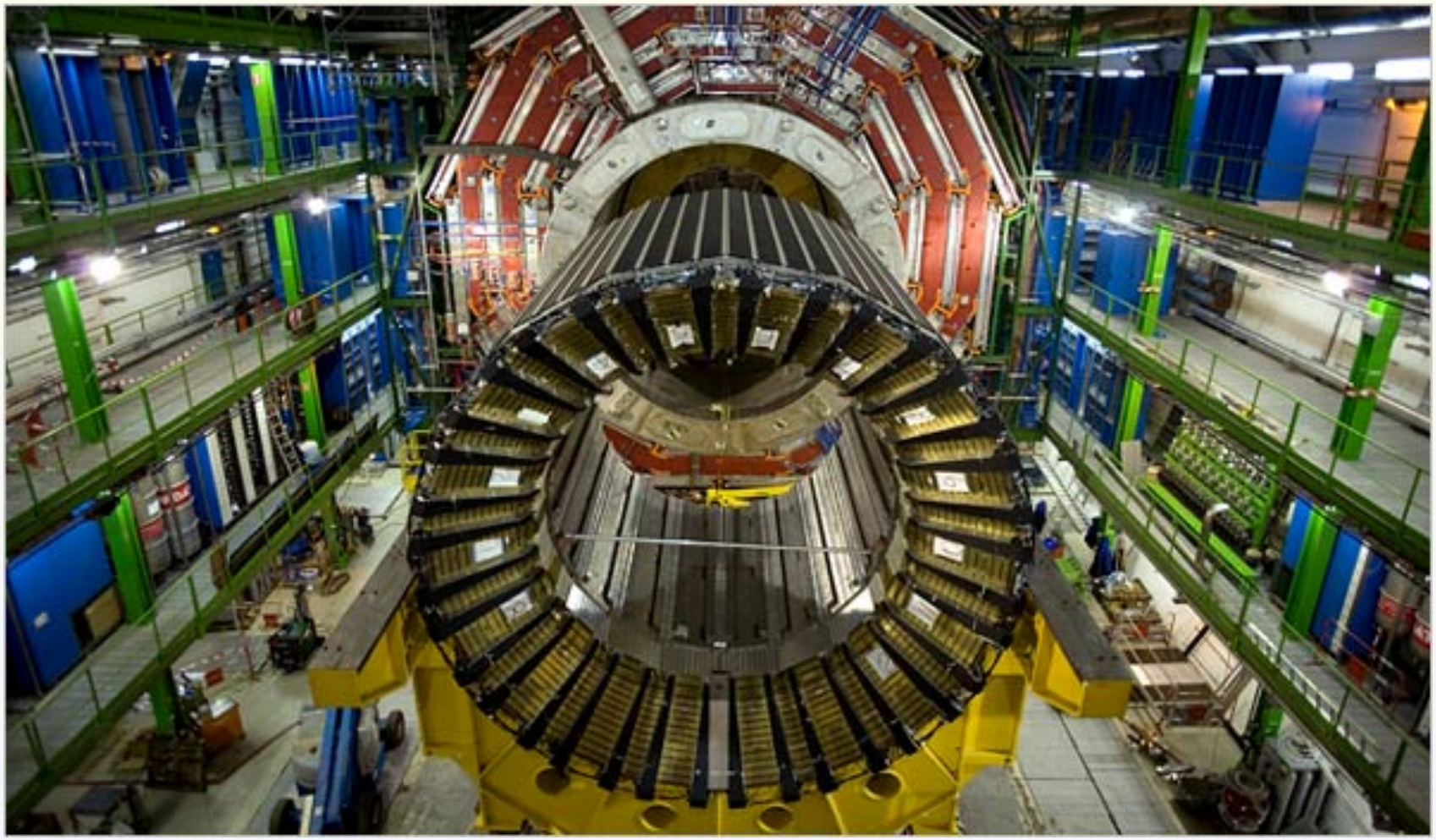
- O Google processa cerca de **20 petabytes** de dados por dia (2008)
- O Wayback Machine tem cerca de **3 petabytes + 100 terabytes/dia** (mar/2009)
- O Facebook tem cerca de **2,5 petabytes de dados de usuários + 15 terabytes/dia** (abr/2009)
- O site eBay tem cerca de **6,5 petabytes de dados dos usuários + 50 terabytes/dia** (mai/2009)
- O Grande Colisor de Hádrons do CERN irá gerar cerca de **15 petabytes/ano**

De qual volume de dados estamos falando?

Muitos, mas **muitos** dados

- O Google processa cerca de **20 petabytes** de dados por dia (2008)
- O Wayback Machine tem cerca de **3 petabytes + 100 terabytes/dia** (mar/2009)
- O Facebook tem cerca de **2,5 petabytes de dados de usuários + 15 terabytes/dia** (abr/2009)
- O site eBay tem cerca de **6,5 petabytes de dados dos usuários + 50 terabytes/dia** (mai/2009)
- O Grande Colisor de Hádrons do CERN irá gerar cerca de **15 petabytes/ano**

De qual volume de dados estamos falando?



De qual volume de dados estamos falando?

$$
\begin{aligned}
1 \text{ PB} &= 1.000.000.000.000.000 \text{ B} \\
&= 1.000^5 \text{ B} \\
&= 10^{15} \text{ B} \\
&= 1 \text{ milhão de gigabytes} \\
&= 1 \text{ mil terabytes}
\end{aligned}
$$

Ou seja, os 15 petabytes que o CERN irá gerar por ano equivalem a 15 milhões de gigabytes. Seriam necessários 1,7 milhão de DVDs *dual-layer* para armazenar tanta informação!

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - PESSOA nasceu em DATA
    - que, por sua vez, permitem encontrar novos fatos

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - os fatos: "Nascimento-de(Mozart, 1756)" e "Nascimento-de(Einstein, 1879)"
    - levam aos dados: "Wolfgang Amadeus Mozart (1756–1791)" e "Einstein nasceu em 1879"
    - que levam a diferentes padrões: "PESSOA (DATA –" e "PESSOA nasceu em DATA"
    - que, por sua vez, permitem encontrar novos fatos

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - os fatos: "Nascimento-de(Mozart, 1756)" e "Nascimento-de(Einstein, 1879)"
    - levam aos dados: "Wolfgang Amadeus Mozart (1756–1791)" e "Einstein nasceu em 1879"
    - que levam a diferentes padrões: "PESSOA (DATA –" e "PESSOA nasceu em DATA"
    - que, por sua vez, permitem encontrar novos fatos

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - os fatos: "Nascimento-de(Mozart, 1756)" e "Nascimento-de(Einstein, 1879)"
    - levam aos dados: "Wolfgang Amadeus Mozart (1756–1791)" e "Einstein nasceu em 1879"
    - que levam a diferentes padrões: "PESSOA (DATA –" e "PESSOA nasceu em DATA"
    - que, por sua vez, permitem encontrar novos fatos

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - os fatos: "Nascimento-de(Mozart, 1756)" e "Nascimento-de(Einstein, 1879)"
    - levam aos dados: "Wolfgang Amadeus Mozart (1756–1791)" e "Einstein nasceu em 1879"
    - que levam a diferentes padrões: "PESSOA (DATA –" e "PESSOA nasceu em DATA"
    - que, por sua vez, permitem encontrar novos fatos

O que se faz com tantos dados?

- Encontram informações sobre novos fatos
  - Casamento de padrões com informações da web
  - ex: quem matou John Lennon?
- Procuram por novas relações entre os dados
  - Alguns padrões levam a novas relações:
    - os fatos: "Nascimento-de(Mozart, 1756)" e "Nascimento-de(Einstein, 1879)"
    - levam aos dados: "Wolfgang Amadeus Mozart (1756–1791)" e "Einstein nasceu em 1879"
    - que levam a diferentes padrões: "PESSOA (DATA –" e "PESSOA nasceu em DATA"
    - que, por sua vez, permitem encontrar novos fatos

Como resolver problemas tão grandes?

Estratégia simples (mas de difícil execução):

- Dividir para conquistar
- Usar mais recursos computacionais a medida que mais dados aparecerem

Grandes *data centers*

Pergunta:

Quão grandes são os *data centers* que fazem sistemas que afetam a vida de quase todo mundo que se conecta a Internet (como os do Google, Facebook, etc.) funcionarem?

Grandes *data centers*

Fonte: http://www.google.com/intl/pt-BR/about/datacenters/

Grandes *data centers*

Fonte: http://www.google.com/intl/pt-BR/about/datacenters/

Grandes *data centers*

Fonte: http://www.google.com/intl/pt-BR/about/datacenters/

Grandes *data centers*

Fonte: http://www.google.com/intl/pt-BR/about/datacenters/

Grandes *data centers*

Só o Google tem treze desses espalhados pelo mundo!

Américas

- Berkeley County, Carolina do Sul
- Council Bluffs, Iowa
- Douglas County, Georgia
- Mayes County, Oklahoma
- Lenoir, Carolina do Norte
- The Dalles, Oregon
- Quilicura, Chile

Grandes *data centers*

Só o Google tem treze desses espalhados pelo mundo!

Ásia

- Hong Kong
- Cingapura
- Taiwan

Grandes *data centers*

Só o Google tem treze desses espalhados pelo mundo!

Europa

- Hamina, Finlândia
- St Ghislain, Bélgica
- Dublin, Irlanda

Como isso era feito até então?

Evolução da computação

- anos 50: computadores eram grandes calculadoras programadas com cartões perfurados; início da computação paralela
- final dos anos 60: ARPANET (computadores começavam a serem interconectados; noção de computação *como um serviço*)
- anos 70: surgem os primeiros microprocessadores
- anos 80: popularização dos computadores pessoais
- anos 90: popularização da Internet

Evolução da computação

- anos 50: computadores eram grandes calculadoras programadas com cartões perfurados; início da computação paralela
- final dos anos 60: ARPANET (computadores começavam a serem interconectados; noção de computação *como um serviço*)
- anos 70: surgem os primeiros microprocessadores
- anos 80: popularização dos computadores pessoais
- anos 90: popularização da Internet

Evolução da computação

- anos 50: computadores eram grandes calculadoras programadas com cartões perfurados; início da computação paralela
- final dos anos 60: ARPANET (computadores começavam a serem interconectados; noção de computação *como um serviço*)
- anos 70: surgem os primeiros microprocessadores
- anos 80: popularização dos computadores pessoais
- anos 90: popularização da Internet

Evolução da computação

- anos 50: computadores eram grandes calculadoras programadas com cartões perfurados; início da computação paralela
- final dos anos 60: ARPANET (computadores começavam a serem interconectados; noção de computação *como um serviço*)
- anos 70: surgem os primeiros microprocessadores
- anos 80: popularização dos computadores pessoais
- anos 90: popularização da Internet

Evolução da computação

- anos 50: computadores eram grandes calculadoras programadas com cartões perfurados; início da computação paralela
- final dos anos 60: ARPANET (computadores começavam a serem interconectados; noção de computação *como um serviço*)
- anos 70: surgem os primeiros microprocessadores
- anos 80: popularização dos computadores pessoais
- anos 90: popularização da Internet

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

Paradigmas de computação

- Computadores Pessoais
- Computadores Paralelos
- Aglomerados de Computadores (*clusters*)
- Computação Voluntária (*Volunteer Computing*):
  - The Great Internet Mersenne Prime Search (1996): busca por primos de Mersenne (primos da forma $2^n - 1, n \in \mathbb{N}$)
  - distributed.net (1997): decriptografia por força-bruta
  - SETI@Home (1999): análise de sinais de rádio vindos do espaço em busca de evidência de vida extra-terreste
- Computação em Grade (*Grid Computing*)

E como é feito agora?

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Grandes data centers

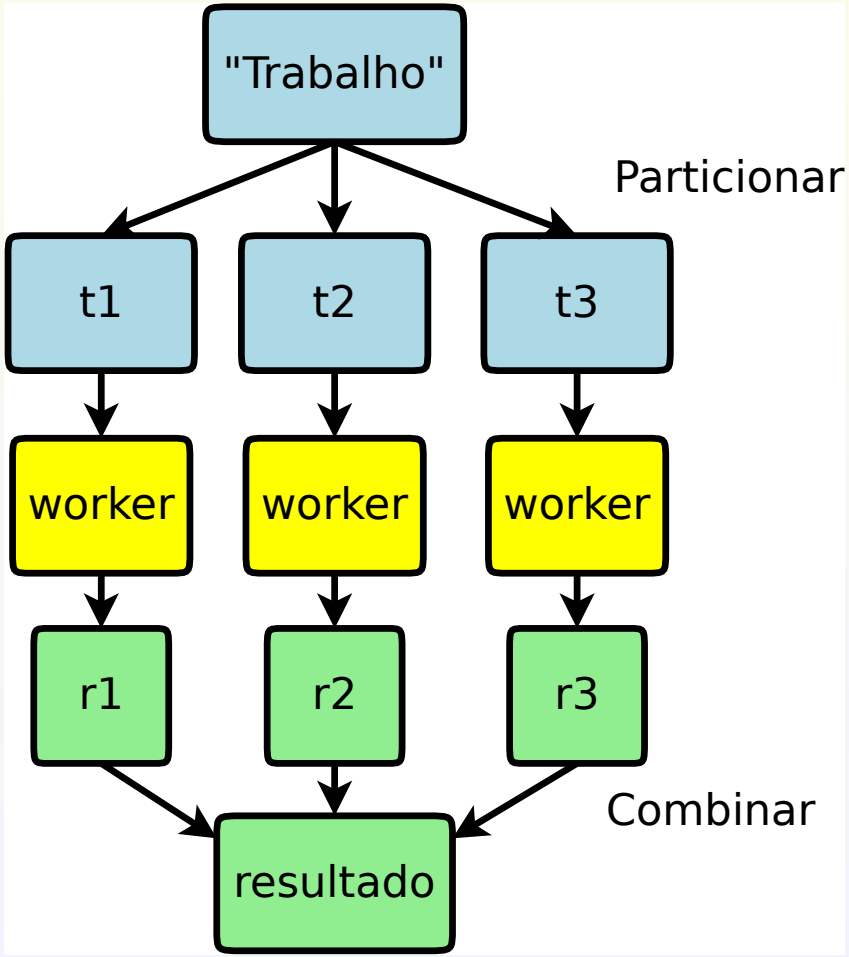
- Seu problema aumenta na mesma escala da web?
  Fácil: basta adicionar mais máquinas
- Tendência: centralização dos recursos computacionais em grandes data centers
  - O que os *fiordes noruegueses*, a *Islândia*, o estado americano do *Oregon* e *minas abandonadas* tem em comum?
- Problemas a serem resolvidos:
  - Redundância
  - Eficiência
  - Utilização
  - Gerenciamento

Ideias principais

- Escalabilidade horizontal, não vertical
  - Existem limites para máquinas SMP e arquiteturas de memória compartilhada
- Mova o processamento para perto dos dados
  - a banda de rede é limitada
- Processe os dados sequencialmente, evite padrões de acesso aleatórios
  - *seeks* são custosos, mas a vazão (*throughput*) do disco é razoável

Como programar aplicações escaláveis?

Divisão e conquista



Desafios de paralelização

- Como repartir as unidades de trabalho entre os *workers*?
- O que fazer quando temos mais trabalho do que *workers*?
- E se os *workers* precisarem compartilhar resultados intermediários entre si?
- Como agregar os resultados parciais?
- O que fazer se um *worker* parar de funcionar?
- Como saber se todos os *workers* terminaram seus trabalhos?

Problema recorrente

- Problemas de paralelização surgem por causa de:
  - comunicação entre os *workers*
  - acesso a recursos compartilhados (por exemplo, dados)
- Portanto, precisamos de algum mecanismo de sincronização

Gerenciar múltiplos *workers*

- É difícil, pois:
  - Não sabemos em que ordem cada *worker* será executado
  - Não sabemos quando um *worker* irá interromper outro *worker*
  - Não sabemos em qual ordem os *workers* irão acessar os dados compartilhados
- Por tanto, nós precisamos de:
  - Semáforos (`lock`, `unlock`)
  - Variáveis condicionais (`wait`, `notify`, `broadcast`)
  - Barreiras de sincronização
- Ainda assim, restam problemas como:
  - Deadlock, starvation, race coditions, ...

Ferramentas atuais

- Modelos de programação:
  - Memória compartilhada (pthreads)
  - Passagem de mensagens (MPI)
- Padrões arquiteturais:
  - Mestre-escravo
  - Produtor-consumidor
  - Filas de trabalho compartilhadas

Moral da história

- Tudo se resume ao nível mais adequado de abstração
- Esconda os detalhes do sistema dos desenvolvedores
  - Evita os problemas com race conditions, contenção em locks, etc.
- Separe o "quê" do "como":
  - O desenvolvedor especifica apenas o que deve ser computado
  - O arcabouço deve se encarregar de como realizar a execução

O data center *é* o computador!

Ruptura na indústria de TI

Computação em Nuvem é uma ideia antiga que finalmente pode ser colocada em prática graças a combinação de várias tecnologias recentes:

- Tecnologias de aplicações web (AJAX, REST, SOA, etc.)
- Virtualização
- Computação Utilitária

Tecnologias web

AJAX *Asynchronous JavaScript and XML*, permitiu a criação de clientes interativos para aplicações web. "Front-end" de computação em nuvem.

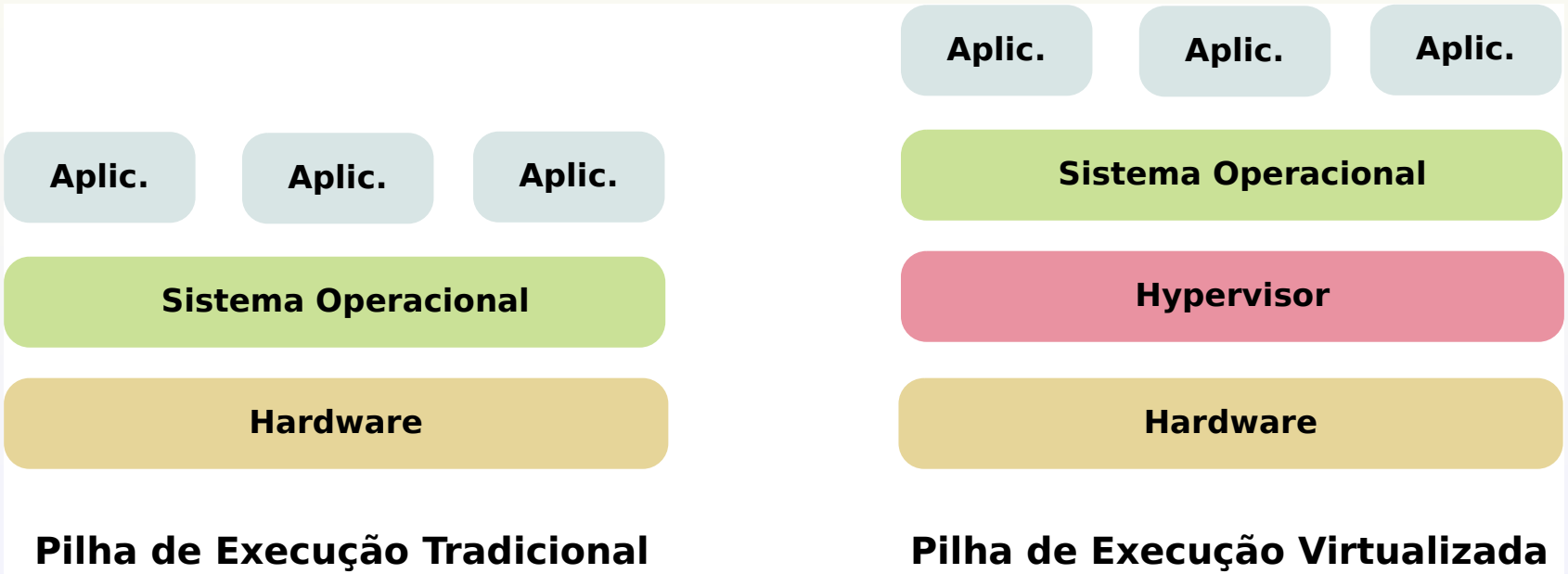
REST *REpresentational State Transfer*, definiu um padrão arquitetural comum para aplicações web

SOA *Service-oriented architecture*, definiu uma série de princípios e metodologias que tornaram os serviços interoperáveis

Virtualização

Definição

Virtualização é a criação de uma versão *virtual* de recursos como um sistema operacional, um servidor, um dispositivo de armazenamento, recursos de rede, etc.



Tipos principais de virtualização de hardware:

- **Virtualização completa**: simulação quase completa do hardware, permite a execução de um sistema operacional hóspede (*guest*) sem que esse precise ser modificado. Exemplos: Parallels Workstation, VirtualBox, Oracle VM, Virtual PC, Virtual Server, VMware Workstation, QEMU, etc.
- **Virtualização assistida pelo hardware**: o hardware provê funcionalidades que facilitam a execução de um monitor de máquinas virtuais e permite a execução isolada de SOs hóspedes. Exemplos: Linux KVM, VMware Workstation, Microsoft Virtual PC, Xen, Oracle VM Server for SPARC, VirtualBox and Parallels Workstation.

Tipos principais de virtualização de hardware:

- **Virtualização parcial**: máquinas virtuais simulam múltiplas instâncias do hardware através de espaços de endereçamento de memória simulados.
- **Para-virtualização**: a máquina virtual não necessariamente simula o hardware, apenas provê uma API que pode ser usado por um SO hóspede (modificado) para notificar mudanças que podem alterar o estado do hardware. Exemplos: Xen, IBM LPARs, Sun's Logical Domains, z/VM, and TRANGO.

Usos de virtualização

- Consolidação de servidores (reduz CapEx/OpEx)
- Alta disponibilidade / recuperação de desastres
- Otimização de infraestrutura (permite planejamento preditivo de recursos)
- Mobilidade (migrações) e segurança (isolamento)
- Infraestrutura inteligente (recursos sob demanda)
- Aplicações "prontas para executar" (*deploy* facilitado)
- etc.

"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility (...) The computer utility could become the basis of a new and important industry."

— John McCarthy, discurso no MIT Centennial em **1961**

Computação Utilitária

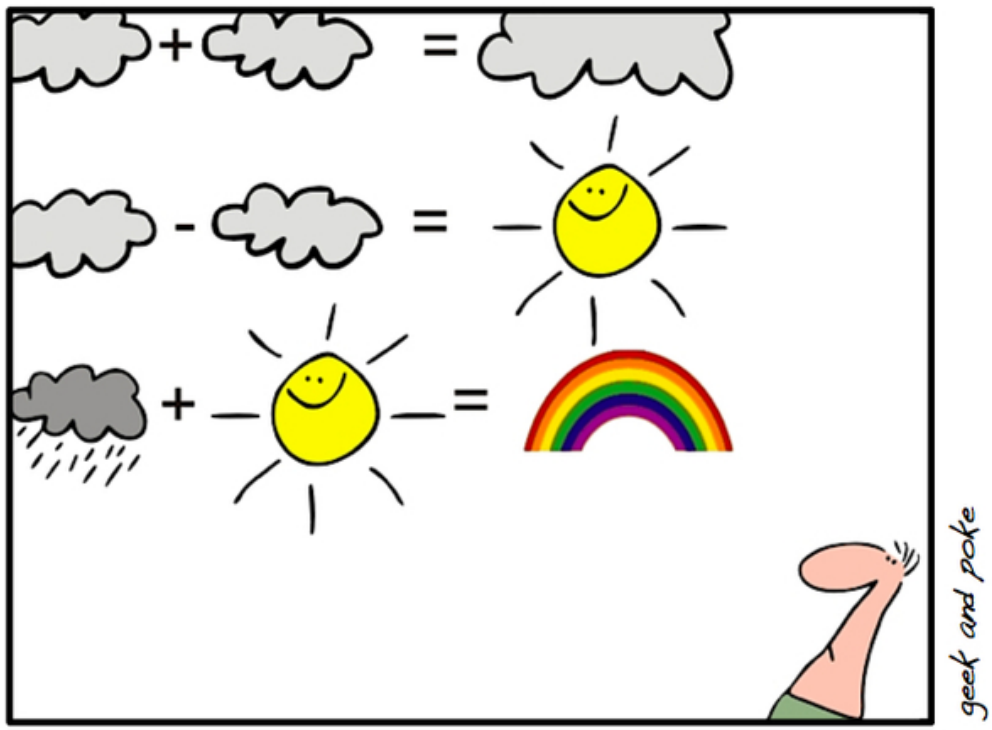
O que é?

- Recursos de computação oferecidos como um serviço onde você paga pelo o que usa
- Habilidade de aprovisionar dinamicamente máquinas (virtuais)

Por quê?

- Custo: despesas de capital vs. despesas de operação
- Escalabilidade: capacidade "infinita"
- Elasticidade: aumentar ou diminuir o poder de processamento

O que é Computação em Nuvem?



http://geek-and-poke.com/2008/05/simply-explaine.html

Origem do termo "Computação em Nuvem"

Segundo a Wikipedia:

A origem do termo Computação em Nuvem é obscura, o termo parece ter sido derivado do uso de uma nuvem estilizada em diagramas de redes de computadores e sistemas de comunicação. A palavra "nuvem" é usada como uma metáfora para a Internet (...)

Mas o que é Computação em Nuvem?

O termo refere-se ao mesmo a dois conceitos distintos:

- aplicações disponibilizadas como serviços via Internet
- hardware e sistemas de software em data centers que proveem esses serviços

Classificação em serviços

A forma como as aplicações e hardware são disponibilizadas para os usuários é utilizada para classificar as plataformas de Computação em Nuvem como:

- Software as a Service (SaaS)
  - as aplicações que rodam no navegador são oferecidas como um serviço
- Platform as a Service (PaaS)
  - a plataforma de desenvolvimento e execução para a criação de aplicações é oferecida como um serviço
- Infrastructure as a Service (IaaS)
  - A infraestrutura de hardware é oferecida como um serviço
- **X**aaS, a lista não para de crescer

Software como um serviço (SaaS)

- Surgiu por volta de 1999 (Salesforce.com)
- Aplicações são licenciadas como um serviço sob demanda
- Modelo de distribuição do serviço:
  - o aplicativo roda diretamente nos servidores do fornecedor
  - o aplicativo é instalado em um dispositivo do cliente e desabilitado quando o contrato termina

Software como um serviço (SaaS)

- Configuração e personalização: a mesma aplicação geralmente funcionalidades e *look-and-feel* diferentes para cada usuário
- *Multi-tenant efficient*: uso de um mesmo servidor para vários "locatários" (*tenants*), sem diferenças aparentes para os usuários
- Escalabilidade: basta fazer o balanceamento de carga entre as instâncias dos usuários

Software como um serviço (SaaS)

- Atualizações frequentes: aplicações SaaS são atualizadas com mais frequência:
  - as aplicações são hospedadas em um único local, o que evita que os clientes tenham que instalar as novas versões
  - a aplicação roda sempre em um ambiente conhecido, o que facilita a fase de testes
  - o fornecedor tem acesso a todos os dados dos clientes, o que acelera os testes de regressão
  - o fornecedor tem acesso a dados de usabilidade das aplicações (via *web analytics*), o que permite detectar as funcionalidades que precisam de melhorias

Exemplos de aplicações oferecidas como SaaS:

- CRM (Customer Relationship Management)
- e-mail
- desktops virtuais
- jogos
- etc.

Exemplos de provedores de SaaS:

- Google Apps (GMail, Drive, Calendar, Talk, etc.)
- iCloud
- salesforce.com
- Basecamp
- Microsoft Office 365
- etc.

Plataforma como um serviço (PaaS)

- Plataforma de computação integrada (para desenvolver / testar / implantar novas aplicações) a um conjunto de soluções, disponibilizada como um serviço
- Permite que uma aplicação seja implantada e distribuída sem que seja preciso se preocupar com as camadas de hardware e software necessárias
- Tipicamente inclui serviços de armazenamento de dados, *middleware*, desenvolvimento, monitoramento, segurança, etc.

Plataforma como um serviço (PaaS)

Vantagens:

- capacidade de provisionamento de novos servidores quase que em tempo real
- ambiente de execução otimizado para a plataforma
- modelo arquitetural padronizado para as aplicações

Plataforma como um serviço (PaaS)

Exemplos serviços fornecidos como plataformas:

- Arcabouços de execução
- Gerenciadores de bancos de dados
- Servidores Web
- Ferramentas de desenvolvimento
- etc.

Plataforma como um serviço (PaaS)

Provedores de PaaS:

- Google AppEngine
- Heroku
- EngineYard
- Force.com
- Windows Azure Cloud Services
- Oracle Platform as a Service

Infraestrutura como um serviço (IaaS)

- Oferecimento de infraestrutura computacional (tipicamente através de um ambiente virtualizado) como um serviço
- Recursos disponibilizados:
  - Servidores
  - Software
  - Espaço no data center
  - Equipamentos de rede

Infraestrutura como um serviço (IaaS)

- Virtualização é a tecnologia fundamental que permitiu a criação de provedores de IaaS
- Graças a virtualização é possível:
  - garantir altas taxas de utilização dos servidores do data center
  - permitir a execução de qualquer sistema operacional hóspede
  - a criação de novas instâncias de servidores pré-configurados (a partir de uma imagem de máquina virtual). Uma nova instância pode ser adicionada em poucos minutos

Exemplos:

- Amazon EC2
- Windows Azure
- Rackspace Cloud
- Google Compute Engine

Por que usar Computação em Nuvem?

- Aplicações em grande escala para processamento de muitos dados
- Flexibilidade
- Escalabilidade
- Adequação as necessidades atuais:
  - hardware
  - software
- Consequências:
  - custos reduzidos
  - menos tempo de manutenção
  - alta disponibilidade
  - menos emissões de carbono

Por que usar Computação em Nuvem?

Flexibilidade

- Software: permite que seu software seja usado a partir de qualquer plataforma
- Acesso: permite acesso aos recursos a partir de qualquer computador conectado a Internet
- Infraestrutura de implantação adaptável:
  - Software controla a infraestrutura

Por que usar Computação em Nuvem?

Escalabilidade

- **Instantânea**
- Controle via software:
  - Adiciona / remove / reconstrói recursos instantaneamente (**elasticidade**)
- Eliminação do comprometimento inicial com o número de recursos necessários ($$$): permite que empresas comecem com um número modesto de recursos e aumente conforme necessário
- Ilusão de um número infinito de recursos computacionais

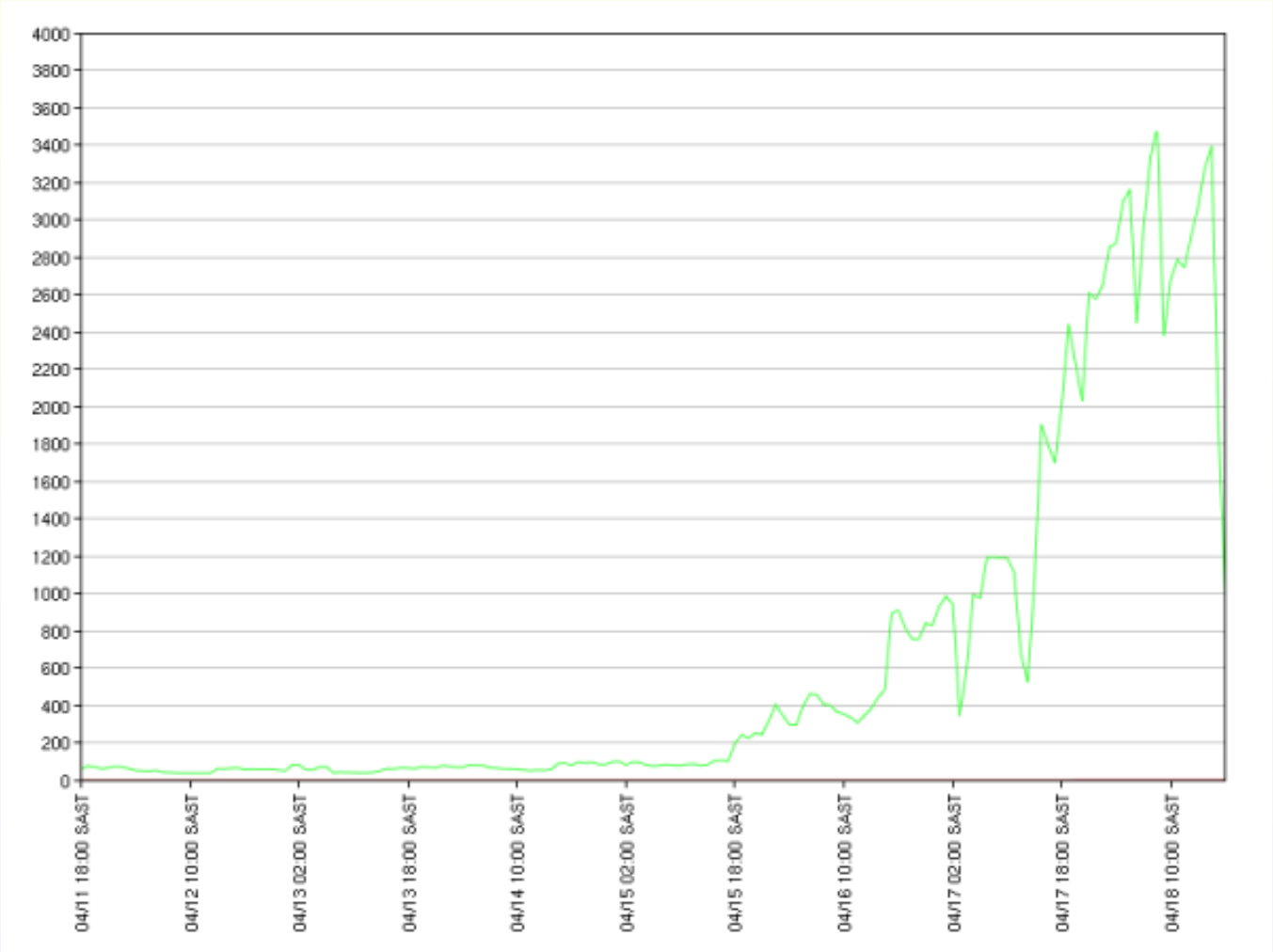
Causo

Quando a *Animoto*[1] tornou seu serviço disponível no Facebook, houve uma explosão na demanda que exigiu que o número de servidores fosse aumentado de 50 para 3.500 em três dias. Após esse pico de utilização, o tráfego caiu para um nível **muito** menor do que o pico.

- Se fosse uma companhia tradicional, o que teria acontecido?
- Com Computação em Nuvem: pague mais durante os picos, devolva os recursos desnecessários depois

[1] A Animoto é uma *startup* que oferece uma aplicação web que produz vídeos a partir de fotos, videoclipes e música.

Evolução do número de instâncias EC2 usadas pela Animoto



Por que usar Computação em Nuvem?

Personalização

- Plataforma de software
- Armazenamento
- Largura de banda de rede
- Velocidade
- ...

Por que usar Computação em Nuvem?

Custo

- Pague a medida que usar
- Pequenas/médias empresas podem competir com a infraestrutura de grandes corporações
  - *time to service / market*
  - sem custo inicial
- Permite reduzir o tamanho do departamento de TI do cliente

Por que usar Computação em Nuvem?

Manutenção

- É inteira responsabilidade do provedor do serviço
- Essas responsabilidades incluem:
  - Atualizações de software
  - Atualizações de segurança
  - Monitoramento do estado do sistema
  - Backup do sistema
  - etc.

Por que usar Computação em Nuvem?

Utilização

- Consolidação de uma grande quantidade de recursos:
  - ciclos de CPU
  - armazenamento
  - largura de banda de rede

Por que usar Computação em Nuvem?

Disponibilidade

- Acesso ao software, plataforma, infraestrutura de qualquer lugar, a qualquer hora
- Basta acesso a Internet

Confiabilidade

- A tolerância a falha do sistema é gerenciada pelo provedor de Computação em Nuvem e os usuários não precisam se preocupar com isso

Por que usar Computação em Nuvem?

Emissão de CO_2

- Consolidação dos servidores
- Maiores taxas de utilização
- Consumo de energia reduzido

Desafios e oportunidades de Computação em Nuvem

- **Disponibilidade do serviço**: use múltiplos provedores; use a elasticidade para se proteger de ataques DDOS
- *Data lock-in*: APIs padronizadas
- **Confidencialidade dos dados e auditabilidade**: empregar criptografia, VLANs, firewalls, armazenamento de dados em diferentes localidades
- **Gargalos na transferência dos dados**: envio de discos pelos Correios; backup/arquivamento
- **Imprevisibilidade do desempenho**: melhorias na tecnologia de máquinas virtuais; uso de memória *flash*; melhorias no escalonamento das máquinas virtuais

Desafios e oportunidades de Computação em Nuvem

- **Disponibilidade do serviço**: use múltiplos provedores; use a elasticidade para se proteger de ataques DDOS
- ***Data lock-in***: APIs padronizadas
- **Confidencialidade dos dados e auditabilidade**: empregar criptografia, VLANs, firewalls, armazenamento de dados em diferentes localidades
- **Gargalos na transferência dos dados**: envio de discos pelos Correios; backup/arquivamento
- **Imprevisibilidade do desempenho**: melhorias na tecnologia de máquinas virtuais; uso de memória *flash*; melhorias no escalonamento das máquinas virtuais

Desafios e oportunidades de Computação em Nuvem

- **Disponibilidade do serviço**: use múltiplos provedores; use a elasticidade para se proteger de ataques DDOS
- ***Data lock-in***: APIs padronizadas
- **Confidencialidade dos dados e auditabilidade**: empregar criptografia, VLANs, firewalls, armazenamento de dados em diferentes localidades
- **Gargalos na transferência dos dados**: envio de discos pelos Correios; backup/arquivamento
- **Imprevisibilidade do desempenho**: melhorias na tecnologia de máquinas virtuais; uso de memória *flash*; melhorias no escalonamento das máquinas virtuais

Desafios e oportunidades de Computação em Nuvem

- **Disponibilidade do serviço**: use múltiplos provedores; use a elasticidade para se proteger de ataques DDOS
- ***Data lock-in***: APIs padronizadas
- **Confidencialidade dos dados e auditabilidade**: empregar criptografia, VLANs, firewalls, armazenamento de dados em diferentes localidades
- **Gargalos na transferência dos dados**: envio de discos pelos Correios; backup/arquivamento
- **Imprevisibilidade do desempenho**: melhorias na tecnologia de máquinas virtuais; uso de memória *flash*; melhorias no escalonamento das máquinas virtuais

Desafios e oportunidades de Computação em Nuvem

- **Disponibilidade do serviço**: use múltiplos provedores; use a elasticidade para se proteger de ataques DDOS
- ***Data lock-in***: APIs padronizadas
- **Confidencialidade dos dados e auditabilidade**: empregar criptografia, VLANs, firewalls, armazenamento de dados em diferentes localidades
- **Gargalos na transferência dos dados**: envio de discos pelos Correios; backup/arquivamento
- **Imprevisibilidade do desempenho**: melhorias na tecnologia de máquinas virtuais; uso de memória *flash*; melhorias no escalonamento das máquinas virtuais

Desafios e oportunidades de Computação em Nuvem

- **Armazenamento escalável**: inventar uma tecnologia de armazenamento escalável
- **Bugs em sistemas distribuídos grandes**: inventar um depurador que utiliza as máquinas virtuais distribuídas
- **Escalabilidade mais rápida**: inventar um escalonador automático baseado em aprendizado computacional
- **Licenciamento de software**: licenças que cobram pelo uso

Desafios e oportunidades de Computação em Nuvem

- **Armazenamento escalável**: inventar uma tecnologia de armazenamento escalável
- **Bugs em sistemas distribuídos grandes**: inventar um depurador que utiliza as máquinas virtuais distribuídas
- **Escalabilidade mais rápida**: inventar um escalonador automático baseado em aprendizado computacional
- **Licenciamento de software**: licenças que cobram pelo uso

Desafios e oportunidades de Computação em Nuvem

- **Armazenamento escalável**: inventar uma tecnologia de armazenamento escalável
- **Bugs em sistemas distribuídos grandes**: inventar um depurador que utiliza as máquinas virtuais distribuídas
- **Escalabilidade mais rápida**: inventar um escalonador automático baseado em aprendizado computacional
- **Licenciamento de software**: licenças que cobram pelo uso

Desafios e oportunidades de Computação em Nuvem

- **Armazenamento escalável**: inventar uma tecnologia de armazenamento escalável
- **Bugs em sistemas distribuídos grandes**: inventar um depurador que utiliza as máquinas virtuais distribuídas
- **Escalabilidade mais rápida**: inventar um escalonador automático baseado em aprendizado computacional
- **Licenciamento de software**: licenças que cobram pelo uso

Modelos de implantação

Nuvem Privada a infraestrutura é provisionada para uso exclusivo de uma única organização com vários usuários

Nuvem Comunitária a infraestrutura é provisionada para um grupo de organizações com interesses em comum

Nuvem Pública a infraestrutura é provisionada para uso pelo público em geral. O provedor pode ser uma empresa, universidade, organização governamental, etc.

Nuvem Híbrida a infraestrutura é uma composição de dois ou mais tipos infraestrutura (privada, pública ou comunitária)

Modelos de implantação

Nuvem Privada a infraestrutura é provisionada para uso exclusivo de uma única organização com vários usuários

Nuvem Comunitária a infraestrutura é provisionada para um grupo de organizações com interesses em comum

Nuvem Pública a infraestrutura é provisionada para uso pelo público em geral. O provedor pode ser uma empresa, universidade, organização governamental, etc.

Nuvem Híbrida a infraestrutura é uma composição de dois ou mais tipos infraestrutura (privada, pública ou comunitária)

Modelos de implantação

Nuvem Privada a infraestrutura é provisionada para uso exclusivo de uma única organização com vários usuários

Nuvem Comunitária a infraestrutura é provisionada para um grupo de organizações com interesses em comum

Nuvem Pública a infraestrutura é provisionada para uso pelo público em geral. O provedor pode ser uma empresa, universidade, organização governamental, etc.

Nuvem Híbrida a infraestrutura é uma composição de dois ou mais tipos infraestrutura (privada, pública ou comunitária)

Modelos de implantação

Nuvem Privada a infraestrutura é provisionada para uso exclusivo de uma única organização com vários usuários

Nuvem Comunitária a infraestrutura é provisionada para um grupo de organizações com interesses em comum

Nuvem Pública a infraestrutura é provisionada para uso pelo público em geral. O provedor pode ser uma empresa, universidade, organização governamental, etc.

Nuvem Híbrida a infraestrutura é uma composição de dois ou mais tipos infraestrutura (privada, pública ou comunitária)

Com tantas vantagens

Por que agora é o momento da Computação em Nuvem?

Vantagens para os provedores

- Dinheiro
  - grandes compras ($>$ 10.000 unidades) permitem negociar preços de equipamento 5–7 vezes mais baratos que em uma compra média (100–1.000)
  - multiplexação de recursos
- Alavancar investimentos em outros negócios
  - companhias podem já possuir nuvens privadas para outros fins
- Defender uma marca
  - migrar clientes atuais para uma plataforma de Computação em Nuvem

Vantagens para os provedores

- Posicionamento estratégico (atacar o "inimigo")
  - Google vs. Microsoft
- Melhorar a relação com os clientes
  - Ex: IBM
  - preservar relações ao oferecer um serviço de Computação em Nuvem *de marca*
- Se tornar uma nova plataforma: mais clientes = mais $

E para os usuários, por que agora?

- Novo modelo de negócios: pague a medida que usar
  - em 2000–2001 a Intel lançou o Intel Computing Service, que exigia um contrato de longa duração. Falhou.
  - clientes não gostam de se comprometer
- Novas aplicações
  - Aplicações móveis interativas
  - Processamento paralelo em *batch*: **muitos** dados
  - *Business analytics*
  - Aplicações desktop com computação intensa (ex: Matlab e Mathematica)

Provedores de plataformas de Computação em Nuvem

Alguns exemplos de plataformas disponíveis no mercado

- Amazon Web Services (EC2, S3, etc.)
  `http://aws.amazon.com`
- Google Cloud Platform (App Engine, Compute Engine, CloudStorage, etc.)
  `http://cloud.google.com`
- Windows Azure
  `http://www.windowsazure.com/`
- UOL Cloud
  `http://uol.com.br/cloud`
- Cloud Locaweb
  `http://locaweb.com.br/nasnuvens`

Amazon Web Services

Amazon Web Services (AWS)

- Amazon.com? Aquela loja que vende livros?
- 2002: Amazon começou a oferecer serviços para outros sites através de protocols como HTTP, REST e SOAP. Os serviços eram cobrados pelo uso
- 2004: alguns engenheiros da Amazon apresentaram um artigo onde explicavam como usar a infraestrutura da loja Amazon.com. Nasce o Amazon EC2
- 2007: Amazon divulga que mais de 330.000 desenvolvedores se inscreveram para usar a Amazon Web Services

Amazon Web Services (AWS)

- Amazon.com? Aquela loja que vende livros?
- 2002: Amazon começou a oferecer serviços para outros sites através de protocols como HTTP, REST e SOAP. Os serviços eram cobrados pelo uso
- 2004: alguns engenheiros da Amazon apresentaram um artigo onde explicavam como usar a infraestrutura da loja Amazon.com. Nasce o Amazon EC2
- 2007: Amazon divulga que mais de 330.000 desenvolvedores se inscreveram para usar a Amazon Web Services

Amazon Web Services (AWS)

- Amazon.com? Aquela loja que vende livros?
- 2002: Amazon começou a oferecer serviços para outros sites através de protocols como HTTP, REST e SOAP. Os serviços eram cobrados pelo uso
- 2004: alguns engenheiros da Amazon apresentaram um artigo onde explicavam como usar a infraestrutura da loja Amazon.com. Nasce o Amazon EC2
- 2007: Amazon divulga que mais de 330.000 desenvolvedores se inscreveram para usar a Amazon Web Services

Amazon Web Services (AWS)

- Amazon.com? Aquela loja que vende livros?
- 2002: Amazon começou a oferecer serviços para outros sites através de protocols como HTTP, REST e SOAP. Os serviços eram cobrados pelo uso
- 2004: alguns engenheiros da Amazon apresentaram um artigo onde explicavam como usar a infraestrutura da loja Amazon.com. Nasce o Amazon EC2
- 2007: Amazon divulga que mais de 330.000 desenvolvedores se inscreveram para usar a Amazon Web Services

Alguns produtos que usam o AWS

- Instagram
- foursquare
- Netflix
- Dropbox
- Heroku
- Pinterest
- tumblr
- etc.

Algumas empresas que confiam no AWS

- Samsung
- Shell
- The New York Times
- Ticketmaster
- Nasa
- Unilever
- Nasdaq
- etc.

Muitos brasileiros já usam

O acesso às plataformas de Computação em Nuvem não são exclusividade das empresas estrangeiras.

- Portal Terra
- SulAmérica
- Grupo Pão de Açúcar
- Gol
- Peixe Urbano
- R7
- Caelum
- etc.

Infraestrutura Global do AWS

Em dez/2011 a Amazon disponibilizou o primeiro data center na América Latina, aqui no Brasil (em São Paulo).

Regiões e zonas de disponibilidade

Regiões

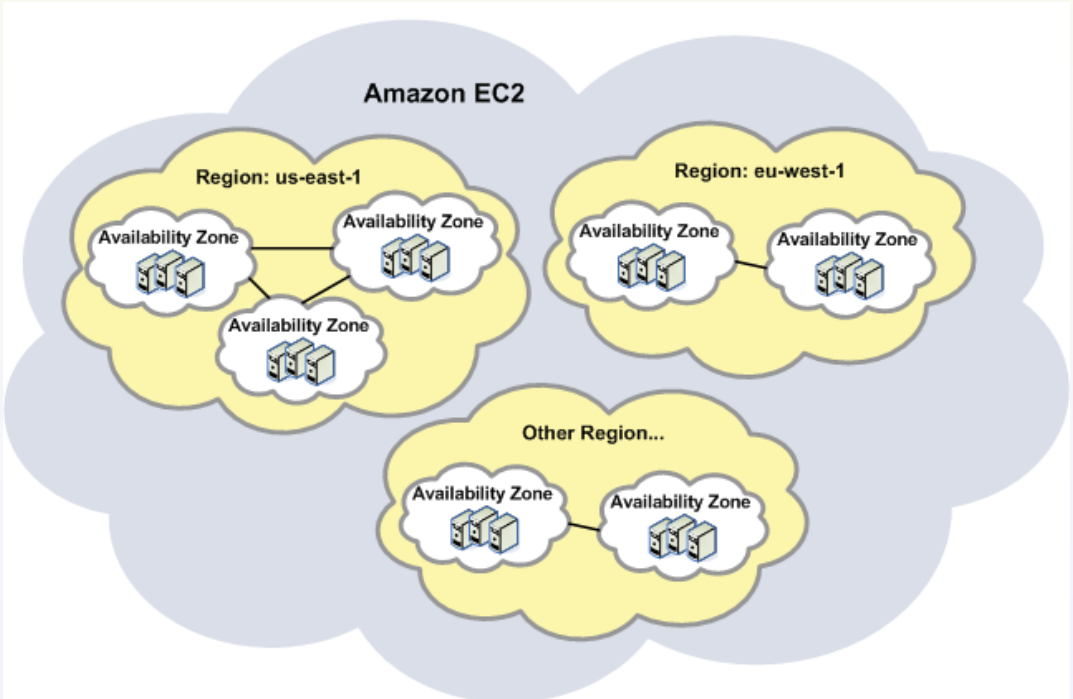
Data centers espalhados geograficamente em diversas localidades:

- EUA Leste (Virginia)
- EUA Oeste (Oregon)
- EUA Oeste (Califórnia)
- Europa (Irlanda)
- Ásia (Singapura)
- Ásia (Tóquio)
- América do Sul (São Paulo)

Regiões e zonas de disponibilidade

Zonas de disponibilidade

Zonas de disponibilidade são áreas separadas dentro de uma região que foram arquitetadas para que falhas em uma zona não afetem os serviços de outra zona. A conexão entre duas zonas de uma mesma região é rápida.



Que tipos de aplicação essas empresas rodam usando Computação em Nuvem?

Aplicações de negócio Shell, Nasdaq, Gol, etc.

Aplicações web Grupo Pão de Açúcar, Samsung, etc.

BigData e HPC Unilever, Ticketmaster, etc.

Recuperação de desastres Hitachi, Amazon.com, etc.

O que os analistas dizem?

the 451 group	AWS é o líder de market share de IaaS (59%)
Gartner	líder em 2011 do Gartner IaaS Magic Quadrant
Forrester	líder em 2011 do Forrester Hadoop Wave

Visão geral dos serviços

Amazon Simple Storage Service (S3)

- Armazenamento para a web. Sempre online, acesso via HTTP
- Armazena e distribui qualquer quantidade de dados, a qualquer momento, para qualquer lugar da web
- Escalável, confiável, rápido e durável. Planejado para durabilidade de 99,999999999% durante um ano (ou seja, 0,000000001% dos objetos podem ser perdidos durante um ano devido a falhas)

Crescimento do Amazon S3

Aumento no número de objetos armazenados na Amazon S3 ao longo dos anos:

2006	2,9 bilhões
2007	14 bilhões
2008	40 bilhões
2009	102 bilhões
2010	262 bilhões
2011	762 bilhões
jun/2012	1 trilhão

Amazon Elastic Block Service (EBS)

- Volumes para uso com instâncias do EC2
- Você o cria e anexa a instância como um disco rígido
- Persiste independente da instância
- *Snapshots* para backup podem ser armazenados no S3

AWS Import/Export

- Acelera a transferência de dados para dentro e para fora do Amazon S3 ou do Amazon EBS
- Transfere dados diretamente de dispositivos de armazenamento físicos
- Usa a rede de alta velocidade interna da Amazon

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Amazon Elastic Compute Cloud (EC2)

- Solução de IaaS da Amazon
- Capacidade computacional de tamanho ajustável
- Controle completo de seus servidores
- Reduz o tempo necessário para obter novos servidores para ordem de minutos
- Escala a capacidade de acordo com demanda automaticamente (se você assim quiser)
- Pague apenas pela capacidade que utilizar

Tipos de instâncias disponíveis

- A Amazon disponibiliza diferentes tipos de computadores
- Cada tipo provê uma capacidade computacional diferente
- As instâncias são cobradas por hora de execução
- Cabe ao desenvolvedor escolher qual instância oferece o melhor custo-benefício para a aplicação
- O poder computacional é definido em termos de *unidades computacionais EC2*: 1 unidade $\simeq$ 1,0 – 1,2 GHz.

Tipos de instâncias disponíveis

Instâncias padrão

Instância pequena:

- Memória de 1,7 GB
- 1 unidade computacional EC2 (1 núcleo virtual com 1 unidade computacional EC2)
- Armazenamento de instâncias de 160 GB
- Plataforma de 32 ou 64 bits
- Desempenho de E/S: moderado
- Otimizada para EBS disponível: não
- Nome da API: m1.small

Tipos de instâncias disponíveis

Instâncias padrão

Instância média:

- Memória de 3,75 GB
- 2 unidades de processamento EC2 (2 núcleos virtuais com 1 unidade de processamento EC2)
- Armazenamento de instâncias de 410 GB
- Plataforma de 32 ou 64 bits
- Desempenho de E/S: moderado
- Otimizada para EBS disponível: não
- Nome da API: m1.medium

Tipos de instâncias disponíveis

Instâncias padrão

Instância grande:

- Memória de 7,5 GB
- 4 unidades de processamento EC2 (2 núcleos virtuais com 2 unidades de processamento EC2 cada)
- Armazenamento de instâncias de 850 GB
- Plataforma de 64 bits
- Desempenho de E/S: alto
- Otimizada para EBS disponível: 500 Mbps
- Nome da API: m1.large

Tipos de instâncias disponíveis

Instâncias padrão

Instância extra-grande:

- Memória de 15 GB
- 8 unidades de processamento EC2 (4 núcleos virtuais com 2 unidades de processamento EC2 cada)
- Armazenamento de instâncias de 1.690 GB
- Plataforma de 64 bits
- Desempenho de E/S: alto
- Otimizada para EBS disponível: 1000 Mbps
- Nome da API: m1.xlarge

Outros tipos de instâncias

Micro pequena quantidade de recursos (até 2 unidades EC2 e memória de 613 MB)

de mais memória grandes tamanhos de memória para aplicações de alta taxa de transferência, incluindo banco de dados e aplicativos de cache de memória. (17, 34 e 68 GB de memória)

de CPUs de alto desempenho 5 a 20 unidades de processamento

de Clusters de Computação CPU de alto desempenho interligadas por rede de alto desempenho

de Cluster GPU CPU de alto desempenho com 2 x GPUs NVIDIA Tesla "Fermi" M2050

Preço do EC2 em SP (outubro/2012)

Região: América do Sul (São Paulo)

	Uso do Linux/UNIX	Uso do Windows
Instâncias On Demand padrão		
Pequeno (padrão)	$0.115 por hora	$0.150 por hora
Médio	$0.230 por hora	$0.300 por hora
Grande	$0.460 por hora	$0.600 por hora
Extragrande	$0.920 por hora	$1.200 por hora
Instâncias On Demand Micro		
Micro	$0.027 por hora	$0.037 por hora
Instâncias sob demanda de memória de alta performance		
Extragrande	$0.680 por hora	$0.800 por hora
Dupla Extragrande	$1.360 por hora	$1.600 por hora
Quádrupla Extragrande	$2.720 por hora	$3.200 por hora
Instâncias sob demanda de CPU de alta performance		
Médio	$0.230 por hora	$0.350 por hora
Extragrande	$0.920 por hora	$1.400 por hora
Instâncias de Cluster Compute		
Quádrupla Extragrande	N/A*	N/A*
Instâncias de Cluster GPU		
Quádrupla Extragrande	N/A*	N/A*
* Cluster and High I/O Instances are not available in all regions		

Preços do EC2 nos EUA (outubro/2012)

Região: Leste dos EUA (Virgínia)

	Uso do Linux/UNIX	Uso do Windows
Instâncias On Demand padrão		
Pequeno (padrão)	$0.080 por hora	$0.115 por hora
Médio	$0.160 por hora	$0.230 por hora
Grande	$0.320 por hora	$0.460 por hora
Extragrande	$0.640 por hora	$0.920 por hora
Instâncias On Demand Micro		
Micro	$0.020 por hora	$0.020 por hora
Instâncias sob demanda de memória de alta performance		
Extragrande	$0.450 por hora	$0.570 por hora
Dupla Extragrande	$0.900 por hora	$1.140 por hora
Quádrupla Extragrande	$1.800 por hora	$2.280 por hora
Instâncias sob demanda de CPU de alta performance		
Médio	$0.165 por hora	$0.285 por hora
Extragrande	$0.660 por hora	$1.140 por hora
Instâncias de Cluster Compute		
Quádrupla Extragrande	$1.300 por hora	$1.610 por hora
Óctupla extra grande	$2.400 por hora	$2.970 por hora
Instâncias de Cluster GPU		
Quádrupla Extragrande	$2.100 por hora	$2.600 por hora
Instâncias com elevada E/S on demand		
Quádrupla Extragrande	$3.100 por hora	$3.580 por hora

Instâncias Spot

- Permitem negociar a parte não usada da infraestrutura do EC2
- Você diz o quanto quer pagar no máximo por uma instância de determinado tipo
- Preços das instâncias são redefinidos periodicamente, variando com a oferta e demanda por capacidade
- Se o preço atual for menor que o preço máximo definido, você recebe uma instância e paga só pelo preço atual (normalmente menor do que o preço máximo)
- Você usa a instância até que opte por terminá-la ou até que o preço atual ultrapasse o preço máximo definido

Como preparar sua instância?

- Criando sua própria imagem de um sistema operacional (mais difícil, porém mais flexível)
- Usando uma *Amazon Machine Image*, imagens de SOs pré-configuradas e prontas para uso
  - Ubuntu
  - Red Hat
  - Windows
  - Cent OS
  - Oracle Linux
  - OpenSolaris
  - Fedora
  - Gentoo
  - OpenSUSE
  - Debian
  - etc.

Top 500

- A lista Top 500 elenca os 500 supercomputadores mais rápidos do mundo
- 1.064 instâncias do EC2 foram usadas para criar um supercomputador com 17.024 cores
- 240 teraflops de velocidade (240 trilhões de operações por segundo)
- Esse supercomputador é o 72º computador mais rápido do mundo na última lista do Top 500 (jun/2012)
- **Você** pode alugá-lo por menos de US$ 1.000/h

Top 500

- A lista Top 500 elenca os 500 supercomputadores mais rápidos do mundo
- 1.064 instâncias do EC2 foram usadas para criar um supercomputador com 17.024 cores
- 240 teraflops de velocidade (240 trilhões de operações por segundo)
- Esse supercomputador é o 72º computador mais rápido do mundo na última lista do Top 500 (jun/2012)
- **Você** pode alugá-lo por menos de US$ 1.000/h

Top 500

- A lista Top 500 elenca os 500 supercomputadores mais rápidos do mundo
- 1.064 instâncias do EC2 foram usadas para criar um supercomputador com 17.024 cores
- 240 teraflops de velocidade (240 trilhões de operações por segundo)
- Esse supercomputador é o 72º computador mais rápido do mundo na última lista do Top 500 (jun/2012)
- **Você** pode alugá-lo por menos de US$ 1.000/h

Top 500

- A lista Top 500 elenca os 500 supercomputadores mais rápidos do mundo
- 1.064 instâncias do EC2 foram usadas para criar um supercomputador com 17.024 cores
- 240 teraflops de velocidade (240 trilhões de operações por segundo)
- Esse supercomputador é o 72º computador mais rápido do mundo na última lista do Top 500 (jun/2012)
- Você pode alugá-lo por menos de US$ 1.000/h

Top 500

- A lista Top 500 elenca os 500 supercomputadores mais rápidos do mundo
- 1.064 instâncias do EC2 foram usadas para criar um supercomputador com 17.024 cores
- 240 teraflops de velocidade (240 trilhões de operações por segundo)
- Esse supercomputador é o 72º computador mais rápido do mundo na última lista do Top 500 (jun/2012)
- **Você** pode alugá-lo por menos de US$ 1.000/h

Amazon Elastic MapReduce (Amazon EMR)

- Permite processar um conjunto vasto de dados com uma boa relação custo-benefício
- Internamente utiliza o arcabouço Hadoop

Auto Scaling

- Permite ajudar a capacidade dos seus servidores EC2 automaticamente
- Útil para aplicações que possuem grande variabilidade no número de usuários
- Disponível sem custo adicional
- Muito útil, mas cuidado com imprevisibilidade dos custos (que também irão variar automaticamente)

Elastic Load Balancing

- Oferece roteamento e balanceamento de carga de conexões HTTP, HTTPS e TCP para instâncias EC2
- Verifica periodicamente a "saúde" das instâncias, para detectar e remover instâncias que estejam com problemas
- Aumenta e diminui dinamicamente o número de recursos baseados nos padrões de acesso
- Integrado com o auto-scaling para aumentar e diminuir o número de instâncias baseado nas medições de escalabilidade da aplicação
- Todos os recursos são acessados através de um único ponto de entrada (um único CNAME nas configurações do DNS)

Amazon DynamoDB

- Serviço de banco de dados NoSQL
- Não limita a quantidade de dados que pode ser armazenada
- Permite provisionar e definir qual a capacidade que cada tabela possui de atender consultas
- Foco no *throughput* e não na quantidade de armazenamento
- Integrado ao serviço de Elastic MapReduce

Amazon SimpleDB

- Implementa apenas as operações básicas de indexação de dados e consultas
- Não possui *schema*, indexação ocorre automaticamente
- Cria e gerencia várias réplicas distribuídas geograficamente
- Elimina a sobrecarga administrativa de modelagem dos dados, manutenção dos índices e ajustes de desempenho

Bancos de dados NoSQL

Um parênteses sobre NoSQL

Bancos de dados NoSQL (not only SQL) são uma classe de sistemas gerenciadores de bancos de dados que se distinguem por não seguirem o modelo de dados relacional.

Características gerais:

- Otimizados para operações de consultas e adição (*append*)
- As funcionalidades em geral se limitam ao armazenamento de registros (pares chave-valor)
- Não usam SQL para as consultas
- Não garantem as propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade)
- Proveem uma arquitetura distribuída tolerante a falhas

Amazon Relational Database Service (RDS)

- Provê acesso a instâncias de bancos de dados tradicionais: MySQL, Oracle ou SQL Server
- Esconde todo o trabalho de administração do banco de dados (configuração, atualizações de segurança, backup, etc.)
- Oferece bom custo-benefício e capacidade de redimensionamento
- Permite que você implante as aplicações que você já usa hoje, sem grandes dores de cabeça com o BD

Amazon ElastiCache

- Cluster para cache que segue a arquitetura do Memcached
- Gerencia tarefas de *patching*, detecção de falhas em nós de cache e recuperação desses nós
- Simples chamadas a uma API permitem aumentar ou diminuir o tamanho do cluster de cache
- Integra-se facilmente a instâncias do SimpleDB e do RDS

Amazon CloudFront

- Serviço web para distribuição de conteúdo
- Distribua conteúdo para usuários com baixa latência e alta velocidade de transferência de dados
- Distribui seu conteúdo usando uma rede global de nós
- Permite downloads, *streaming* e *live streaming* com o Adobe Flash Media Server.

Amazon Simple Workflow Service (SWF)

- Roda aplicações especificadas por *workflows* e processos de negócio no AWS
- Gerencia aplicações que estão na plataforma do SWF, aplicações móveis e mesmo aplicações in-loco (que acessam a Amazon para solicitar tarefas)
- Funciona com qualquer linguagem de programação

Amazon CloudSearch

- Serviço de busca gerenciado pela Amazon
- Permite integrar um serviço de busca rápido e escalável em qualquer aplicação
- Escalabilidade automática: se adapta ao aumento da quantidade de dados indexados ou ao aumento da quantidade de buscas
- AWS gerencia o provisionamento de recursos, particionamento dos dados e atualizações de software

Amazon Simple Notification Service (SNS)

- Facilita a configuração, a operação e o envio de notificações
- Publica mensagens geradas por uma aplicação e a envia para "assinantes" ou outras aplicações
- Mensagens podem ser enviadas usando diferentes protocolos (HTTP, e-mail, etc.).
- Usa um mecanismo "push" que elimina a necessidade de verificação periódica ou "poll" para novas informações e atualizações.

Amazon Simple Queue Service (SQS)

- Provê um serviço de filas escalável e confiável para o armazenamento de mensagens
- Move dados entre componentes distribuídos de uma aplicação

Amazon Simple Email Service (Amazon SES)

- Serviço de envio de e-mails (em lote) transacionais
- Elimina os problemas com o gerenciamento de servidores de e-mails, configurações de rede e padrões (rigorosos) dos servidores de acesso a Internet
- Provê um sistema de *feedback*, incluindo notificações sobre e-mails que não foram entregues, dados sobre as tentativas de entrega e reclamações sobre spam

AWS Elastic Beanstalk

- PaaS da Amazon
- Automaticamente gerencia a implantação, provisionamento, balanceamento de carga, auto-scaling e monitoramento do status do sistema
- Executa e gerencia aplicações escritas em PHP, .NET, Java e Python
- Controla todos os serviços do AWS necessários para a aplicação
- Porém, o usuário continua com o controle sobre a infraestrutura e sobre o software
- Possui ferramenta de desenvolvimento integrada ao Visual Studio e ao Eclipse
- Não há cobrança pelo uso do serviço, só pelos recursos utilizados

Amazon CloudWatch

- Permite visualizar a utilização dos recursos, desempenho operacional, e padrões de utilização dos serviços
- As métricas incluem uso de CPU, leituras e escritas em disco, tráfego de rede, etc.
- Permite também a definição de novas métricas específicas da aplicação
- Os dados são acessíveis a partir da interface de gerenciamento, mas também através de APIs, SDK ou CLI

AWS Identity and Access Management (IAM)

- Permite a criação de usuários e grupos com permissões específicas, tais como restrições de acesso a algumas APIs ou recursos do AWS
- Controla o acesso ao console de gerenciamento e a algumas chamadas às APIs, de acordo com as credenciais dos usuários
- Também possibilita a concessão de acesso a recursos da AWS para usuários gerenciados fora da AWS no seu diretório corporativo

Referências

- Above the Clouds: A Berkeley View of Cloud Computing
  http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html
- Cursos dos professores:
  - Majd F. Sakr (Carnegie Mellon Qatar) – http://www.qatar.cmu.edu/~msakr/15319-s12/
  - Jimmy Lin (University of Maryland) – http://www.umiacs.umd.edu/~jimmylin/cloud-2010-Spring/
- Amazon Web Services: http://aws.amazon.com/pt/
- The AWS Network on SlideShare:
  http://www.slideshare.net/AmazonWebServices,
  incluindo o canal da América Latina:
  http://www.slideshare.net/AmazonWebServicesLATAM/