Cloud Computing

Introduction to Cloud Computing

Dell Zhang
Birkbeck, University of London
2014/15
What is Cloud Computing?
Origin of the “Cloud” Metaphor
The best thing since sliced bread?

• Before Clouds ...
  – Supercomputers
  – Grids

• Cloud Computing means many different things:
  – Big-Data Processing
  – Rebranding of Web 2.0
  – Utility Computing
  – Everything as a Service
Big-Data Processing

• Lots of User-Generated Content
  – Examples: Youtube, Instagram, ...

• Even More User-Behaviour Data
  – Examples: Google search logs, Google ad clicks, Facebook friend suggestions, ...

• Data Science
  – Gather as much data as you need and run machine learning / data mining / predictive analytics algorithms to generate insights
Rebranding of Web 2.0

• Rich, interactive web applications
  – Clouds refer to the servers that run them
  – AJAX as the de facto standard (for better or worse)
  – Examples: Facebook, YouTube, Gmail, ...

• “The network is the computer”: take two
  – User data is stored “in the cloud”
  – Rise of netbooks, tablets, smartphones, etc.
  – Browser is the OS
Rebranding of Web 2.0

Mozilla 1993
Netscape 1994
Internet Explorer 1995
Opera 1996
Safari 2003
Firefox 2004
Chrome 2008
Rebranding of Web 2.0
Utility Computing

• What?
  – Computing resources as a metered service (i.e., “pay as you go”)
  – Ability to dynamically provision virtual machines
Utility Computing

On demand

Pay as you go

Uniform

Available
Utility Computing

On demand

Pay as you go

Infrastructure

Uniform

Available
Utility Computing

API

Infrastructure
Utility Computing

• Why?
  – Cost: capital vs. operating expenses
  – Scalability: “infinite” capacity
  – Elasticity: scale up or down on demand

• Does it make sense?
  – Benefits to cloud users
  – Business case for cloud providers
Utility Computing

• Who cares?
  – Provision of Hadoop clusters on-demand in the cloud
  – Lower barrier to entry for tackling big-data problems
  – Commoditization and democratization of big-data capabilities
Everything as a Service
The History of Cloud Computing

• It’s a long way

I think there is a world market for maybe five computers.
<table>
<thead>
<tr>
<th>Year</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1943</td>
<td>&quot;I think there is a world market for maybe five computers.&quot;</td>
</tr>
<tr>
<td>late 1970s</td>
<td>&quot;The mainframe will always be the prevalent computing platform. The minicomputer is a toy.&quot;</td>
</tr>
<tr>
<td>early 1980s</td>
<td>&quot;The PC will never be successful. People do not need their own personal computers.&quot;</td>
</tr>
<tr>
<td>mid-1980s</td>
<td>&quot;The minicomputer will prevail. PC and networked computers are merely toys.&quot;</td>
</tr>
<tr>
<td>early 1990s</td>
<td>&quot;The Internet has no real future as a computing platform. Too unreliable. Too hard to use. Could never support millions.&quot;</td>
</tr>
<tr>
<td>mid-1990s</td>
<td>&quot;Electronic commerce is a joke. The Web is just a way to provide marketing information.&quot;</td>
</tr>
<tr>
<td>late 1990s</td>
<td>&quot;There is no business model giving software away for free. The concept of collecting 'eyeballs' will never make money.&quot;</td>
</tr>
</tbody>
</table>
1960s-1980s: Time-sharing

Origin of “virtual machine” concept
Mainframe computing costly, so one user’s idle time used to service other users

1990s: Client-server

IT gets re-invented
Client-server model splits tasks between client systems initiating requests and server systems responding over a computer network

2000s: Grids & SaaS

Early concepts of “utility” computing
Large arrays of commodity hardware harnessed for big compute tasks.
Complex applications begin to be accessed over the internet via web browsers

2005+: The cloud

Utility model reborn
Utility data center resources made available in on-demand model as a service accessible via a browser on the internet
Where does it all start?

- 2004-2005: Google GFS/MapReduce/BigTable
- 2006: Amazon Web Services (AWS)
- 2010: Microsoft Windows Azure
How did Amazon...

...get into cloud computing?
<table>
<thead>
<tr>
<th>Consumer Business</th>
<th>Seller Business</th>
<th>IT Infrastructure Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tens of millions of active customer accounts</td>
<td>Sell on Amazon websites</td>
<td>Cloud computing infrastructure for hosting web-scale solutions</td>
</tr>
<tr>
<td></td>
<td>Use Amazon technology for your own retail website</td>
<td>Hundreds of thousands of registered customers in over 190 countries</td>
</tr>
<tr>
<td>Eight countries: US, UK, Germany, Japan, France, Canada, China, Italy</td>
<td>Leverage Amazon’s massive fulfillment center network</td>
<td></td>
</tr>
</tbody>
</table>
1 Trillion

650k+ peak transactions per second
### Cloud Computing Activities by Age Cohorts

Internet users in each age group who do the following online activities (%)

<table>
<thead>
<tr>
<th>Activity</th>
<th>18-29</th>
<th>30-49</th>
<th>50-64</th>
<th>65+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use webmail services such as Hotmail, Gmail, or Yahoo! mail</td>
<td>77%</td>
<td>58%</td>
<td>44%</td>
<td>27%</td>
</tr>
<tr>
<td>Store personal photos</td>
<td>50%</td>
<td>34%</td>
<td>26%</td>
<td>19%</td>
</tr>
<tr>
<td>Use online applications such as Google Documents or Adobe Photoshop Express</td>
<td>39%</td>
<td>28%</td>
<td>25%</td>
<td>19%</td>
</tr>
<tr>
<td>Store personal videos</td>
<td>14%</td>
<td>6%</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>Pay to store computer files online</td>
<td>9%</td>
<td>4%</td>
<td>5%</td>
<td>3%</td>
</tr>
<tr>
<td>Back up hard drive to an online site</td>
<td>7%</td>
<td>5%</td>
<td>5%</td>
<td>4%</td>
</tr>
<tr>
<td>Have done at least <strong>one</strong> activity</td>
<td><strong>87%</strong></td>
<td><strong>71%</strong></td>
<td><strong>59%</strong></td>
<td><strong>46%</strong></td>
</tr>
<tr>
<td>Have done at least <strong>two</strong> activities</td>
<td><strong>59%</strong></td>
<td><strong>39%</strong></td>
<td><strong>31%</strong></td>
<td><strong>21%</strong></td>
</tr>
</tbody>
</table>

Source: Pew Internet & American Life Project April-May 2008 Survey. N=1,553 Internet users. Margin of error is ±3%.
The Principles of Cloud Computing

• (1) Pooled Resources
• (2) Virtualization
• (3) Elasticity
• (4) Automation
• (5) Metered Billing
(1) Pooled Resources

- Available to any subscribing users
COOLING: High-efficiency water-based cooling systems—less energy-intensive than traditional chillers—circulate cold water through the containers to remove heat, eliminating the need for air-conditioned rooms.

STRUCTURE: A 24,000-square-meter facility houses 400 containers. Delivered by trucks, the containers attach to a spine infrastructure that feeds network connectivity, power, and water. The data center has no conventional raised floors.

POWER: Two power substations feed a total of 300 megawatts to the data center, with 200 MW used for computing equipment and 100 MW for cooling and electrical losses. Batteries and generators provide backup power.

CONTAINER: Each 67.5-cubic-meter container houses 2,500 servers about 10 times as many as conventional data centers pack in the same space. Each container integrates computing, networking, power, and cooling systems.
(2) Virtualization

• High utilization of hardware assets

15%-

=> 65%+
<table>
<thead>
<tr>
<th>Benefit</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decouples users from implementation</td>
<td>The concept of a virtual server forces users to not worry about the physical servers or their location. Instead, they focus on service-level agreements and their applications.</td>
</tr>
<tr>
<td>Decreases server provisioning from months to minutes</td>
<td>Getting a (physical) server requisitioned, installed, configured, and deployed takes larger organizations 60–90 days and some 120 days. In the virtual server model, it’s literally minutes or hours from request to fully ready for application deployment, depending on how much automation has been put in place.</td>
</tr>
<tr>
<td>Breaks software pricing and licensing</td>
<td>No longer can the data center charge for an entire server or every server the software runs on. Instead, they have to charge for actual usage—a whole new model for IT.</td>
</tr>
</tbody>
</table>
(3) Elasticity

• Dynamic scale without CAPEX

503
Service Temporarily Unavailable

The server is temporarily unable to service your request due to maintenance downtime or capacity problems. Please try again later.
Call CreateLoadBalancer with the following parameters:

```
AvailabilityZones = us-east-1a
LoadBalancerName = MyLoadBalancer
Listeners = lb-port=80,instance-port=8080,protocol=HTTP
```

Call CreateLaunchConfiguration with the following parameters:

```
ImageId = myAMI
LaunchConfigurationName = MyLaunchConfiguration
InstanceType = ml.small
```

Call CreateAutoScalingGroup with the following parameters:

```
AutoScalingGroupName = MyAutoScalingGroup
AvailabilityZones = us-east-1a
LaunchConfigurationName = MyLaunchConfiguration
LoadBalancerNames = MyLoadBalancer
MaxSize = 20
MinSize = 2
```

Call CreateOrUpdateScalingTrigger with the following parameters:

```
AutoScalingGroupName = MyAutoScalingGroup
MeasureName = CPUUtilization
Statistic = Average
TriggerName = MyTrigger1a
Namespace = AWS/EC2
Period = 60
LowerThreshold = 40
LowerBreachScaleIncrement = -1
UpperThreshold = 80
UpperBreachScaleIncrement = 1
BreachDuration = 600
```
(4) Automation

• Building, deploying, configuring, provisioning, and moving, all without manual intervention
<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>An Amazon Machine Image is an encrypted and signed machine image suitable to run in a virtual server environment. For example, it may contain Linux, Apache, MySQL, or PHP, as well as the application of the AMI’s owner. AMIs can be public (provided by Amazon), private (custom designed by its creator), paid (purchased from a third party), or shared (created by the community for free). AMIs can be stored in Amazon’s Simple Storage Service (S3).</td>
</tr>
<tr>
<td>Instance</td>
<td>The result of launching an AMI is a running system called an instance. When an instance terminates, the data on that instance vanishes. For all intents and purposes, an Instance is identical to a traditional host computer.</td>
</tr>
</tbody>
</table>
| Standard flow   | 1. Use a standard AMI by customizing an existing one.  
2. Bundle the AMI, and get an AMI ID to enable launching as many instances of the AMI as needed.  
3. Launch one or more instances of this AMI.  
4. Administer and use the running instance(s). |
| Connecting      | From a web browser, go to http://<hostname>, where <hostname> is your instance’s public hostname. If you want to connect to a just-launched public AMI that hasn’t been modified, run the `ec2-get-console-output` command. The result in either case enables you to log in as root and exercise full control over this instance, just like any host computer you could walk up to in a data center. |
(5) Metered Billing

• Pay for what you use
The Benefits of Cloud Computing

- (1) **Economic** benefits of the change from capital expenses (CAPEX) to operational expenses (OPEX)
- (2) **Agility** benefits from not having to procure and provision servers
- (3) **Efficiency** benefits that may lead to competitive advantages
- (4) **Security** stronger and better in the cloud
Cloud

- Faster time to market
- Higher availability
- No need for CapEx
- Large cost savings
- Focus on core competency
On-Premise Infrastructure

Your Business

30%

Managing All of the "Undifferentiated Heavy Lifting"

70%

Cloud-Based Infrastructure

More Time to Focus on Your Business

70%

Configuring Your Cloud Assets

30%
The Economics of Cloud Computing

Application deployment models

- **Own data center**
  - CAPEX: $$$
  - OPEX: $$$

- **Colocation**
  - CAPEX: $$
  - OPEX: $$

- **Managed hosting**
  - CAPEX: 0
  - OPEX: $$$

- **Cloud computing**
  - CAPEX: 0
  - OPEX: $$
A small e-commerce configuration

• 3 years

• 2 firewalls: 2 x $1,500 = $3,000
• 2 load-balancers: 2 x $5,000 = $10,000
• 6 commodity servers: 6 x $3,000 = $18,000
## Internal IT Deployment

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two firewalls</td>
<td>$3,000</td>
</tr>
<tr>
<td>Two load-balancers</td>
<td>$10,000</td>
</tr>
<tr>
<td>Six servers</td>
<td>$18,000</td>
</tr>
</tbody>
</table>

Total CAPEX cost of hardware: $31,000

Depreciated over three years (36 months): $861 per month
## Colocation Deployment

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$31,000 Total cost of HW</td>
<td>10 Mbit contract</td>
</tr>
<tr>
<td>$31,000 ÷ 36 Months</td>
<td></td>
</tr>
<tr>
<td>$861 per month</td>
<td>+ $1,000 per month</td>
</tr>
<tr>
<td>= $1,861 per month</td>
<td></td>
</tr>
</tbody>
</table>
## Managed-Service Deployment

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$300/month Firewalls</td>
<td>500 GB/server included</td>
</tr>
<tr>
<td>$1,500/month Load balancers</td>
<td>$0 Additional charge</td>
</tr>
<tr>
<td>$6,000/month Six servers</td>
<td>$7,800 ÷ 36 Installation (once)</td>
</tr>
<tr>
<td>$7,800 per month</td>
<td>$217 per month</td>
</tr>
<tr>
<td></td>
<td><strong>= $8,017 per month</strong></td>
</tr>
</tbody>
</table>

**Total Annual Cost:**

$8,017 per month × 12 months = $96,204 per year
<table>
<thead>
<tr>
<th>Size</th>
<th>Memory</th>
<th>Num EC2 compute units (1 = 1.0-1.2 GHz 2007 Opteron CPU)</th>
<th>Storage</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>1.7 GB</td>
<td>1</td>
<td>160 GB</td>
<td>32-bit</td>
</tr>
<tr>
<td>Large</td>
<td>7.5 GB</td>
<td>4</td>
<td>850 GB</td>
<td>64-bit</td>
</tr>
<tr>
<td>Extra large</td>
<td>15 GB</td>
<td>8</td>
<td>1690 GB</td>
<td>64-bit</td>
</tr>
</tbody>
</table>
Cloud Deployment

<table>
<thead>
<tr>
<th>Hardware + storage</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$216/month Virtual VPN</td>
<td>10 TB max outbound</td>
</tr>
<tr>
<td>$25/month Load-balancing</td>
<td></td>
</tr>
<tr>
<td>$300/month Storage</td>
<td></td>
</tr>
<tr>
<td>$759/month Six large</td>
<td>$135/month @ $0.17/GB</td>
</tr>
<tr>
<td>instances</td>
<td></td>
</tr>
<tr>
<td>$1,300/month</td>
<td>$135/month</td>
</tr>
</tbody>
</table>

= $1,435 per month
A Small E-Commerce Configuration

• What if we only need it for 6 months?
  – Internal IT deployment?
  – Cloud deployment?
Where Does the Cloud Make Sense?

• (1) Limited lifetime requirement or short-term need
  – e.g., ...?

• (2) Scale variability or volatility
  – e.g., ...?

• (3) Nonstrategic applications or low organizational value
  – e.g., ...?
On and Off

Fast Growth

Variable peaks

Predictable peaks

CUSTOMER DISSATISFACTION

WASTE
Your IT needs

Time

Capacity

Traditional IT capacity

Elastic cloud capacity
Case Study

Time: +72h

<10 cores

- 0
- 500
- 1000
- 1500

Mar 29, 2010 10:00 PM Mar 31, 2010 5:00 PM Apr 2, 2010 12:00 PM Apr 4, 2010 7:00 AM Apr 6, 2010 2:00 AM Apr 7, 2010 9:00 PM Apr 9, 2010 4:00 PM Apr 11, 2010 11:00 AM
40 servers to 5000 in 3 days

- EC2 scaled to peak of 5000 instances
- "Techcrunched"
- Launch of Facebook modification
- Steady state of ~40 instances

Case Study
Daily United States People
09/02/09 - 02/28/10

- Directly Measured
- Rough Estimate

target.com

US: 1.2M  Max: 3.0M  11/29/09
GLOBAL

global stats not yet available for estimated data

Rough Estimate ⚠️

© 2010 Quantcast Corp.
Daily Reach at Leading Celebrity Gossip Sites

(Unique Visitors on the domain as a % of all Unique visitors online that day. June, 5 – July, 5.)

- tmz.com
- perezhilton.com
- eonline.com
- ew.com
- people.com
Where Does the Cloud Not Make Sense?

• (1) Legacy systems
  – e.g., ...?

• (2) Applications involving real-time or mission-critical scenarios
  – e.g., ...?

• (3) Applications dealing with confidential data
  – e.g., ...?
Take Home Messages

• What is Cloud Computing
• The History of Cloud Computing
• The Principles of Cloud Computing
• The Benefits of Cloud Computing
• The Economics of Cloud Computing
• Where Does the Cloud (Not) Make Sense