HY559
Infrastructure Technologies for Large-Scale Service-Oriented Systems

Kostas Magoutis
magoutis@ics.forth.gr
http://www.ics.forth.gr/~magoutis
Cloud Computing

• Applications delivered as a service over the Internet
  – Long referred to as “Software as a Service”

• Hardware and systems-software in the datacenters that provide above services
  – The “Cloud”

• There are different flavors of Cloud Computing
  – Available in a “pay as you go” manner to general public
    • Service being sold is Utility Computing
  – Internal data centers of businesses or other organizations
    • Private Clouds (as opposed to Public Clouds defined above)
Users and providers of Cloud Computing

Diagram:
- SaaS User
  - Web applications
- SaaS Provider / Cloud User
  - Utility computing
- Cloud Provider
New aspects of Cloud Computing

• Illusion of infinite computing resources available on demand
  – No need to plan far ahead for provisioning

• Elimination of upfront commitment by Cloud users
  – Possible to start small and grow resources as needed

• Ability to pay for use of computing resources on short-term basis, as needed, release them as needed
  – Reward conservation
When is Cloud Computing a viable business

- Construction and operation of extremely large-scale, commodity-computer datacenters at low-cost locations uncovers large economies of scale
  - Factors of 5-7 decrease in cost of electricity, network bandwidth, operations, software, hardware

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost in Medium-sized DC</th>
<th>Cost in Very Large DC</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbit/sec/month</td>
<td>$13 per Mbit/sec/month</td>
<td>7.1</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per GByte / month</td>
<td>$0.40 per GByte / month</td>
<td>5.7</td>
</tr>
<tr>
<td>Administration</td>
<td>≈140 Servers / Administrator</td>
<td>&gt;1000 Servers / Administrator</td>
<td>7.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Price per KWH</th>
<th>Where</th>
<th>Possible Reasons Why</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.6¢</td>
<td>Idaho</td>
<td>Hydroelectric power; not sent long distance</td>
</tr>
<tr>
<td>10.0¢</td>
<td>California</td>
<td>Electricity transmitted long distance over the grid; limited transmission lines in Bay Area; no coal fired electricity allowed in California.</td>
</tr>
<tr>
<td>18.0¢</td>
<td>Hawaii</td>
<td>Must ship fuel to generate electricity</td>
</tr>
</tbody>
</table>
When is Cloud Computing a viable business

• Statistical multiplexing increases utilization compared to private Cloud
  – Possible to offer service at low cost, make good profit

• Leverage existing investment
  – Example: Amazon

• Defend a franchise
  – Example: Microsoft applications
When is Cloud Computing a viable business

- Leverage customer relationships
  - Example: IBM Global Services

- Become a platform
  - Example: Facebook
Cloud client benefits

- Transfers risk of misestimating load to Cloud provider
  - Peak vs. average expected load

HY-559 Spring 2013
Cloud client benefits

- No penalty from unexpectedly scaling down
- Track variable rates of hardware cost reduction
- Avoid practical limits on utilization of purchased equipment, operational costs
Key technologies (1)

• Highly-scalable software systems
  – Initially built to support internal needs (Google, Amazon)

• Statistical multiplexing by virtualizing resources
  – Computation, storage, communication

• Ubiquitous remote access over Web services APIs
  – SOAP, REST
Key technologies (2)

• Fine-grain monitoring and charging for resource use
  – Must be built ground-up

• Multi-tenancy for fine-grain co-location of clients

• Service-level agreements (SLAs)
What are the different offerings?

Distinguished based on level of abstraction and level of resource management:

- **Amazon EC2**
  - Low level, physical hardware view (VMs)

- **Microsoft Azure**
  - .NET libraries, Common Language runtime

- **Google AppEngine**
  - Traditional Web application style
Top obstacles to cloud computing

1. Availability of service
2. Data lock-in
3. Data confidentiality
4. Data transfer bottlenecks
5. Performance unpredictability
6. Scalable storage
7. Bugs in large-scale distributed systems
8. Scaling quickly
9. Reputation fate sharing
10. Software licensing