Datacenter is new “server”
(Luíz Barroso, Google Distinguished Engineer, early 2007)

- Google program == Web search, Gmail,…
- Google computer == 1000’s computers, storage, network
- Warehouse-sized facilities and workloads
- New datacenter ideas (2007-2008): truck container (Sun), floating (Google), datacenter-in-a-tent (Microsoft)
- *How to enable innovation in new services without first building a Google-sized company?*
Enable **1 person** to develop, deploy, operate next-generation Internet application

- Key enabling technology: Statistical machine learning
  - debugging, power management, performance prediction, ...
- Highly interdisciplinary faculty & students
  - PI’s: Patterson/Fox/Katz (systems/networks), Jordan (machine learning), Stoica (networks & P2P), Joseph (systems/security), Franklin (databases)
  - 2 postdocs, ~30 PhD students, ~5 undergrads

Examples

- Predict performance of complex software system when demand is scaled up
- Automatically add/drop servers to fit demand, without violating SLA
- Distill millions of lines of log messages into an operator-friendly “decision tree” that pinpoints “unusual” incidents/conditions
- **Recurring theme:** cutting-edge SML methods work where simpler methods have failed
- **Sponsor feedback:** Need to demonstrate applicability on at least 1000’s of machines!
Utility Computing Arrives

- Amazon Elastic Compute Cloud (EC2)
- “Compute unit” rental: $0.10-0.80/hr.
  - 1 CU ≈ 1.0-1.2 GHz 2007 AMD Opteron/Xeon core

<table>
<thead>
<tr>
<th>“Instances”</th>
<th>Platform</th>
<th>Cores</th>
<th>Memory</th>
<th>Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small - $0.10 / hr</td>
<td>32-bit</td>
<td>1</td>
<td>1.7 GB</td>
<td>160 GB</td>
</tr>
<tr>
<td>Large - $0.40 / hr</td>
<td>64-bit</td>
<td>4</td>
<td>7.5 GB</td>
<td>850 GB – 2 spindles</td>
</tr>
<tr>
<td>XLarge - $0.80 / hr</td>
<td>64-bit</td>
<td>8</td>
<td>15.0 GB</td>
<td>1690 GB – 3 spindles</td>
</tr>
</tbody>
</table>

- No up-front cost, no contract, no minimum
- Billing rounded to nearest hour; pay-as-you-go storage also available
- A new paradigm (!) for deploying services?

But...

What is cloud computing, exactly?
“It’s nothing new”

“...we’ve redefined Cloud Computing to include everything that we already do... I don’t understand what we would do differently ... other than change the wording of some of our ads.”

Larry Ellison, CEO, Oracle (Wall Street Journal, Sept. 26, 2008)

“It’s a trap”

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

Above the Clouds:
A Berkeley View of Cloud Computing

abovetheclouds.cs.berkeley.edu

• White paper by RAD Lab PI’s and students
  – Clarify terminology around Cloud Computing
  – Quantify comparison with conventional computing
  – Identify Cloud Computing challenges & opportunities

• Why can we offer new perspective?
  – Strong engagement with industry
  – Users of cloud computing in our own research and teaching in last 12 months

• This talk: introduce some key points of report, persuade you to visit RAD Lab later today

What is it? What’s new?

• Old idea: Software as a Service (SaaS)
  – Software hosted in the infrastructure vs. installed on local servers or desktops
  – Recently: “[Hardware, Infrastructure, Platform] as a service”? Poorly defined, so we avoid

• New: pay-as-you-go utility computing
  – Illusion of infinite resources on demand
  – Fine-grained billing: release == don’t pay
  – Earlier examples: Sun, Intel Computing Services—longer commitment, more $$$/hour
Why Now (not then)?

- Build-out of extremely large datacenters (1,000’s to 10,000’s of commodity computers)
  - Economy of scale: 5-7x cheaper than provisioning a medium-sized (100’s machines) facility
  - Build-out driven by demand growth (more users)
  - Infrastructure software: eg Google FileSystem
  - Operational expertise: failover, DDoS, firewalls...

- Other factors
  - More pervasive broadband Internet
  - x86 as universal ISA, fast virtualization
  - Standard software stack, largely open source (LAMP)

Cloud Economics 101

- Static provisioning for peak: wasteful, but necessary for SLA

```
<table>
<thead>
<tr>
<th>Time</th>
<th>Demand</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Statically provisioned” data center

```

```
<table>
<thead>
<tr>
<th>Time</th>
<th>Demand</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Virtual” data center in the cloud

Unused resources
```
Risk of underutilization

- Underutilization results if “peak” predictions are too optimistic

Static data center

Risks of underprovisioning

Lost revenue

Lost users
New Scenarios Enabled by “Risk Transfer”

• “Cost associativity”: 1,000 computers for 1 hour same price as 1 computer for 1,000 hours
  – Washington Post converted Hillary Clinton’s travel documents to post on WWW \(<1\text{ day}\) after released
  – RAD Lab graduate students demonstrate improved Hadoop (batch job) scheduler—on 1,000 servers

• Major enabler for SaaS startups
  – Animoto traffic doubled every 12 hours for 3 days when released as Facebook plug-in
  – Scaled from 50 to >3500 servers
  – ...then scaled back down

Classifying Clouds

• Instruction Set VM (Amazon EC2, 3Tera)
• Managed runtime VM (Microsoft Azure)
• Framework VM (Google AppEngine, Force.com)
• Tradeoff: flexibility/portability vs. “built in” functionality

Lower-level, Less managed       Higher-level, More managed

EC2  Azure  AppEngine  Force.com
Challenges & Opportunities

• Challenges to adoption, growth, & business/policy models
• Both technical and nontechnical
• Most translate to 1 or more opportunities
• Complete list in paper; I’ll give an example of each
• Paper also provides worked examples to quantify tradeoffs (“Should I move my service to the cloud?”)

Adoption Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability / business continuity</td>
<td>Multiple providers &amp; DCs</td>
</tr>
<tr>
<td>Data lock-in</td>
<td>Standardization</td>
</tr>
<tr>
<td>Data Confidentiality and Auditability</td>
<td>Encryption, VLANs, Firewalls; Geographical Data Storage</td>
</tr>
</tbody>
</table>
### Growth Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data transfer bottlenecks</td>
<td>FedEx-ing disks, Data Backup/Archival</td>
</tr>
<tr>
<td>Performance unpredictability</td>
<td>Improved VM support, flash memory, scheduling VMs</td>
</tr>
<tr>
<td>Scalable structured storage</td>
<td>Major research opportunity</td>
</tr>
<tr>
<td>Bugs in large distributed systems</td>
<td>Invent Debugger that relies on Distributed VMs</td>
</tr>
<tr>
<td>Scaling quickly</td>
<td>Invent Auto-Scaler that relies on ML; Snapshots</td>
</tr>
</tbody>
</table>

### Policy and Business Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reputation Fate Sharing</td>
<td>Offer reputation-guarding services like those for email</td>
</tr>
<tr>
<td>Software Licensing</td>
<td>Pay-as-you-go licenses; Bulk licenses</td>
</tr>
</tbody>
</table>

*Breaking news (2/11/09):* IBM WebSphere™ and other service-delivery software will be available on Amazon AWS with *pay-as-you-go* pricing
Visit & give us feedback

• RAD Lab Open House, 2PM, 465 Soda
  – posters, research, students, faculty
  – meet authors: Michael Armbrust, Rean Griffith, Anthony Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ari Rabkin, Ion Stoica, and Matei Zaharia

• abovetheclouds.cs.berkeley.edu
  – Paper, executive summary, slides
  – “Above the Clouds” blog
  – Impromptu video interview with authors