Above the Clouds

A Berkeley View of Cloud Computing

UC Berkeley RAD Lab

Presentation at RPI, September 2011

Outline

- What is it?
- Why now?
- Cloud killer apps
- Economics for users
- Economics for providers
- Challenges and opportunities
- Implications

Cloud computing is "hot"…

"A new term for the long-held dream of computing as a utility [D. Parkhill, *The Challenge of the Computer Utility*, Addison Wesley, 1966]"

"...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." *Sept. 2008*

Larry Ellison, Oracle's CEO,
quoted in Wall Street Journal, September 26, 2008

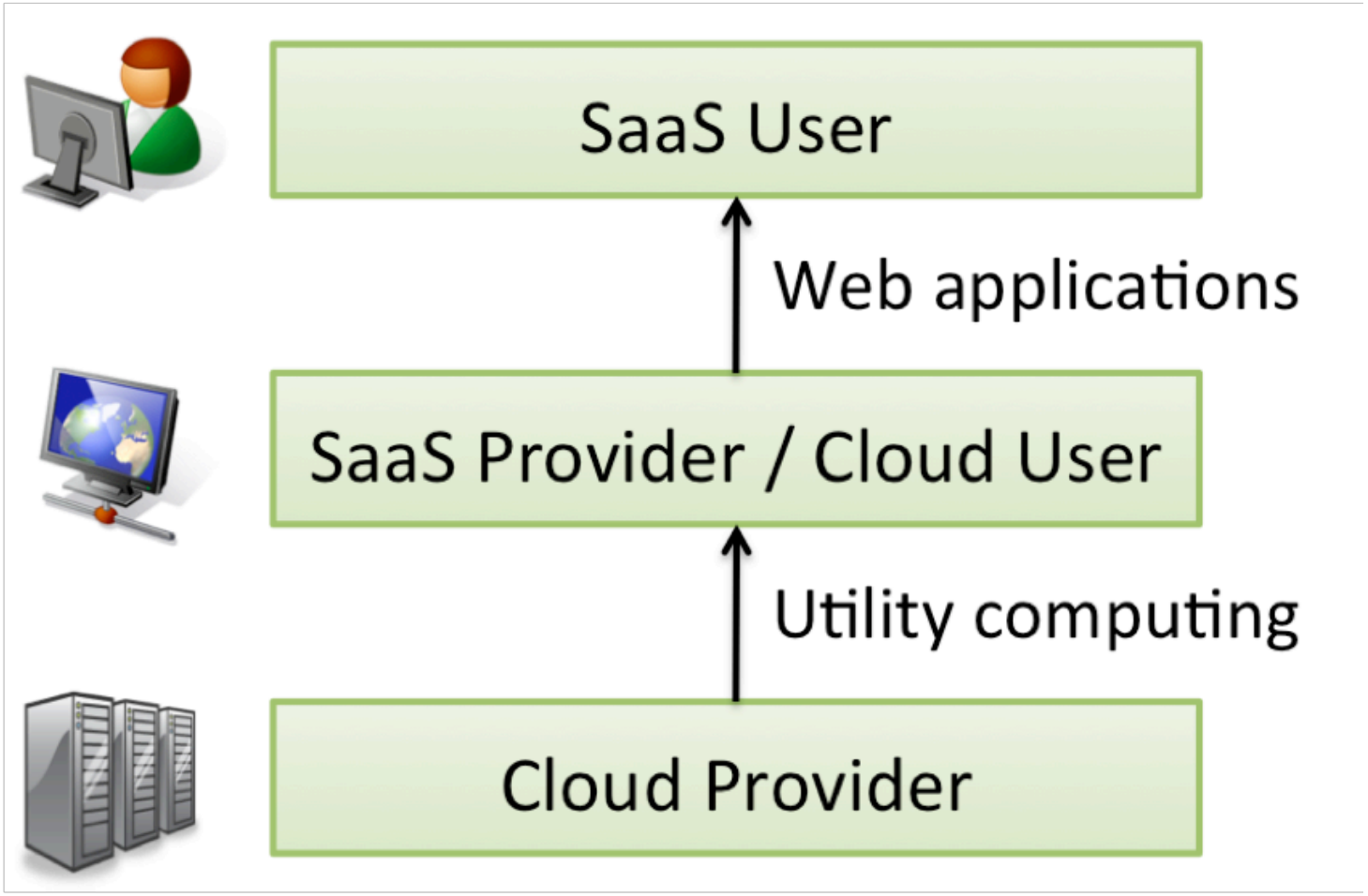
What is Cloud Computing?

- Old idea: Software as a Service (SaaS)
  - Def: delivering applications over the Internet
- Recently: "[Hardware, Infrastructure, Platform] as a service"
  - Poorly defined so we avoid *all* "X as a service"

except SaaS…

- Utility Computing: pay-as-you-go computing
  - Illusion of infinite resources
  - No up-front cost
  - Fine-grained billing (e.g. hourly)

SaaS and Cloud: Users and Providers

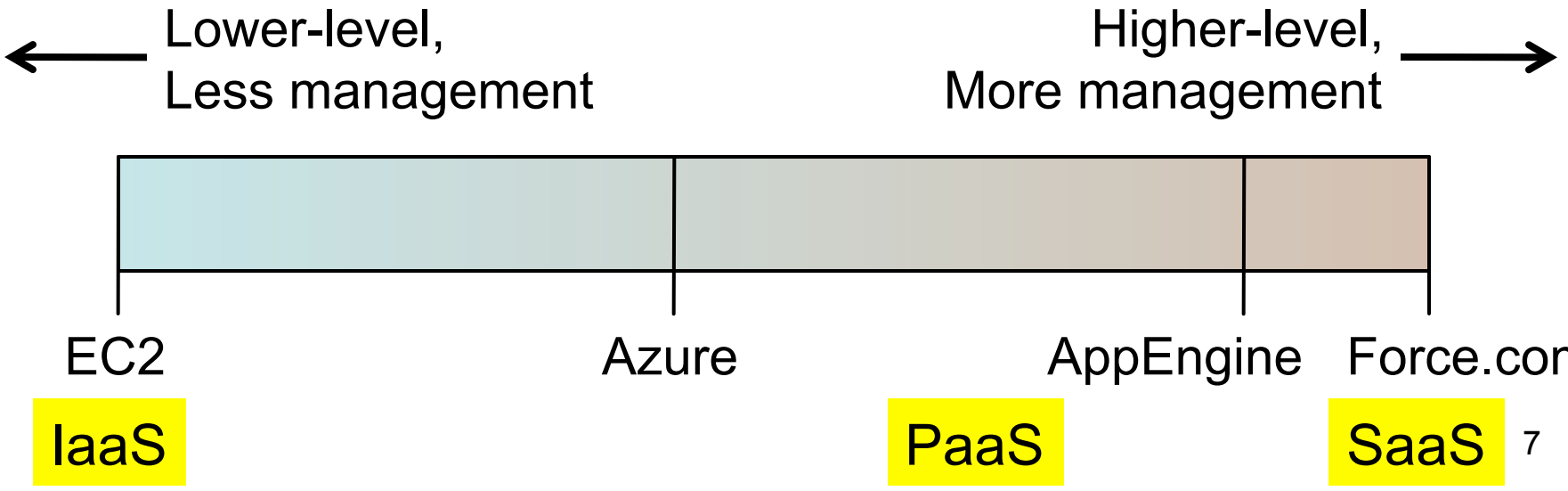


Why Now?

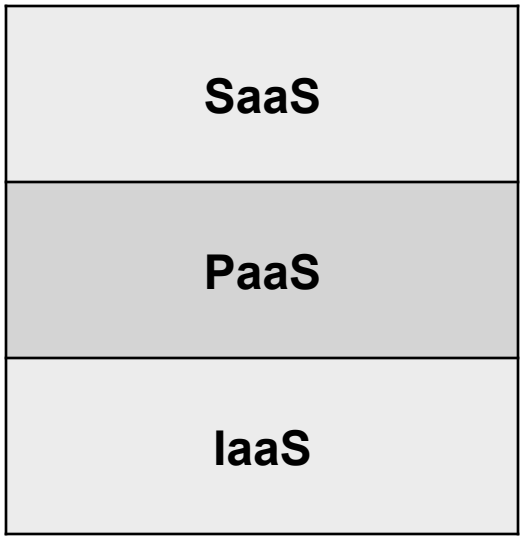
- Experience with very large datacenters
  - Unprecedented economies of scale
- Other factors
  - Pervasive broadband Internet
  - Fast x86 virtualization
  - Pay-as-you-go billing model
  - Standard software stack

Spectrum of Clouds

- Instruction Set VM (Amazon EC2, 3Tera)
- Bytecode VM (Microsoft Azure)
- Framework VM
  - Google AppEngine, Force.com



Composite Clouds



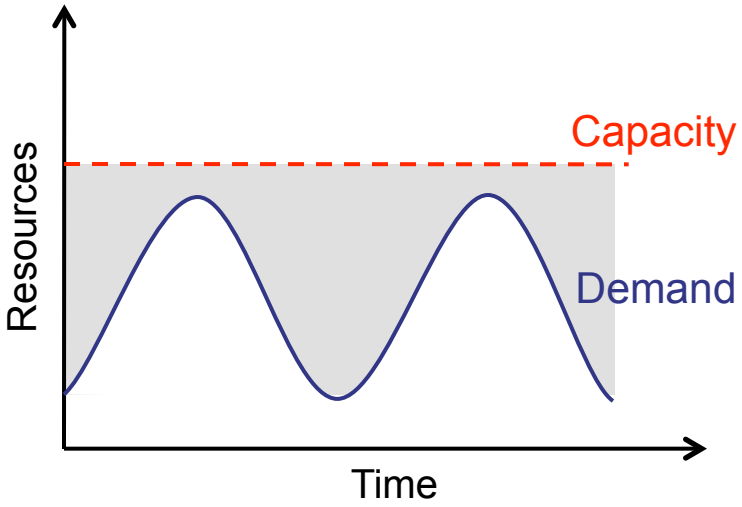
It is possible to stack/layer services, so that, e.g., Gmail (SaaS) uses the Google Apps Engine (PaaS) over virtual machines provided by Amazon (IaaS). Notice that layering hides SaaS user from back-end infrastructure.

Cloud Killer Apps

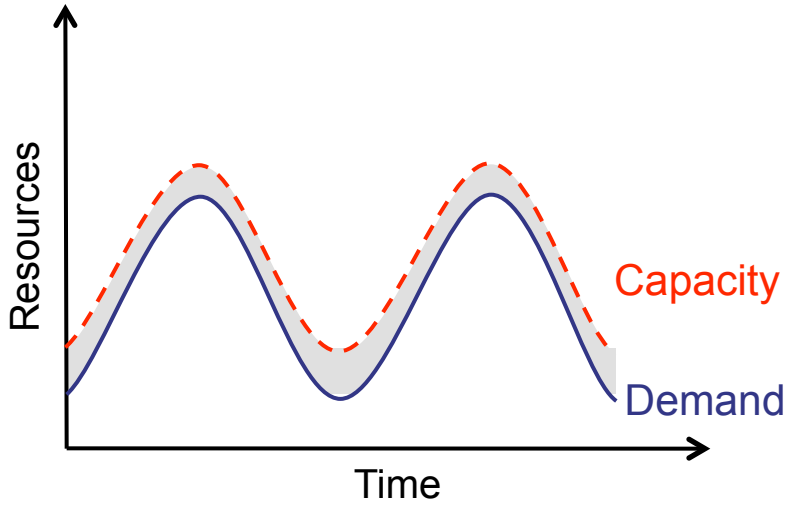
- Mobile and web applications
- Extensions of desktop software
  - Matlab, Mathematica
- Batch processing / MapReduce
  - Oracle at Harvard, Hadoop at NY Times

Economics of Cloud Users

- Pay by use instead of provisioning for peak



Static data center

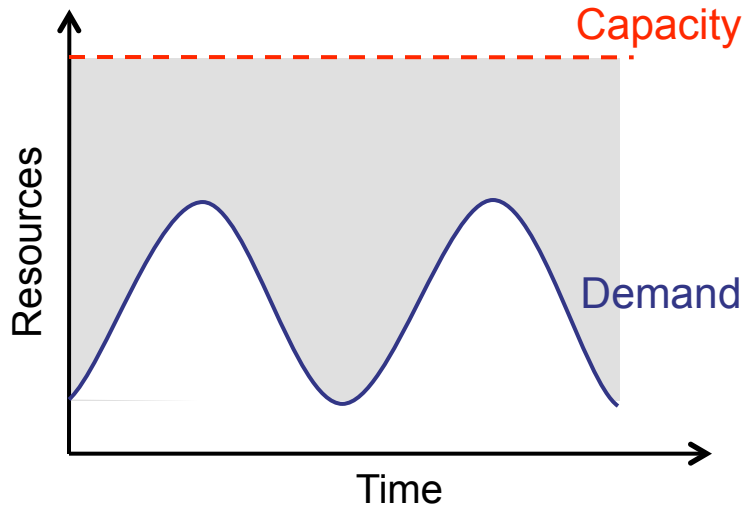


Data center in the cloud

Unused resources

Economics of Cloud Users

- Risk of over-provisioning: underutilization

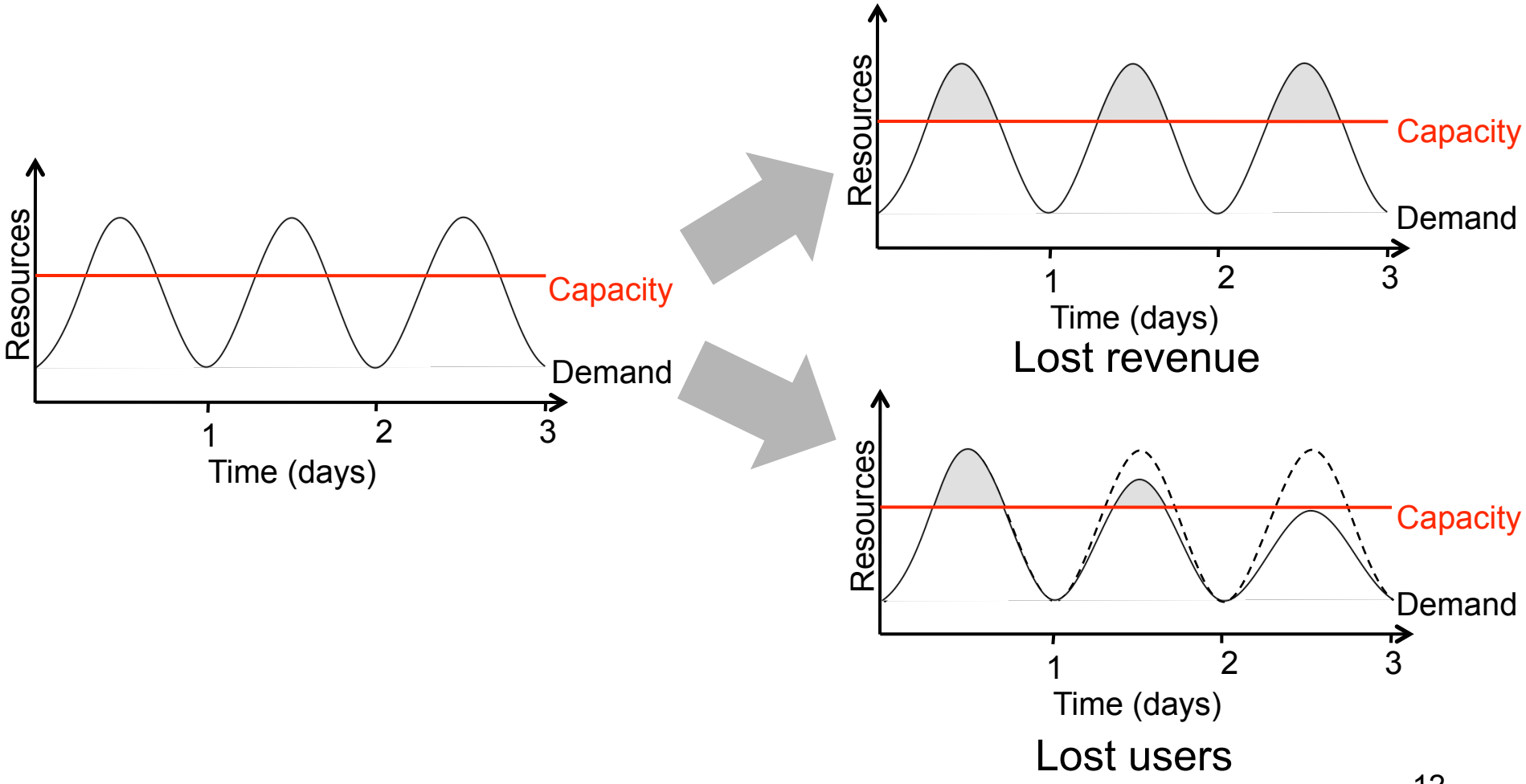


Unused resources

Static data center

Economics of Cloud Users

- Heavy penalty for under-provisioning



To cloud or not to cloud?

$$\text{UserHours}_{\text{cloud}} \times (\text{revenue} - \text{Cost}_{\text{cloud}}) \geq \text{UserHours}_{\text{datacenter}} \times (\text{revenue} - \frac{\text{Cost}_{\text{datacenter}}}{\text{Utilization}})$$

Revenue using public cloud vs revenue using private cloud

Hybrid clouds combine the benefits of both!

September 2011 Amazon EC2 Instance Costs

Region: US East (Virginia)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.12 per hour
Large	$0.34 per hour	$0.48 per hour
Extra Large	$0.68 per hour	$0.96 per hour
Micro On-Demand Instances		
Micro	$0.02 per hour	$0.03 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.62 per hour
Double Extra Large	$1.00 per hour	$1.24 per hour
Quadruple Extra Large	$2.00 per hour	$2.48 per hour
Hi-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.29 per hour
Extra Large	$0.68 per hour	$1.16 per hour
Cluster Compute Instances		
Quadruple Extra Large	$1.60 per hour	N/A*
Cluster GPU Instances		
Quadruple Extra Large	$2.10 per hour	N/A*
* Windows® is not currently available for Cluster Compute or Cluster GPU Instances		

Source: http://aws.amazon.com/ec2/pricing/

September 2011 Amazon Data Transfer Costs

Region: US East (Virginia)

	Pricing
Data Transfer IN	
All data transfer in	$0.000 per GB
Data Transfer OUT	
First 1 GB / month	$0.000 per GB
Up to 10 TB / month	$0.120 per GB
Next 40 TB / month	$0.090 per GB
Next 100 TB / month	$0.070 per GB
Next 350 TB / month	$0.050 per GB
Next 524 TB / month	Contact Us
Next 4 PB / month	Contact Us
Greater than 5 PB / month	Contact Us

Source: http://aws.amazon.com/ec2/pricing/

September 2011 Amazon Free Trials Available!

Free Tier*

As part of AWS's Free Usage Tier, new AWS customers can get started with Amazon EC2 for free. Upon sign-up, new AWS customers receive the following EC2 services each month for one year:

- 750 hours of EC2 running Linux/Unix Micro instance usage
- 750 hours of Elastic Load Balancing plus 15 GB data processing
- 10 GB of Amazon Elastic Block Storage (EBS) plus 1 million IOs, 1 GB snapshot storage, 10,000 snapshot Get Requests and 1,000 snapshot Put Requests
- 15 GB of bandwidth out aggregated across all AWS services
- 1 GB of Regional Data Transfer

Source: http://aws.amazon.com/ec2/pricing/

Economics of Cloud Providers

- 5-7x economies of scale [Hamilton 2008]

Resource	Cost in Medium DC	Cost in Very Large DC	Ratio
Network	$95 / Mbps / month	$13 / Mbps / month	7.1x
Storage	$2.20 / GB / month	$0.40 / GB / month	5.7x
Administration	≈140 servers/admin	>1000 servers/admin	7.1x

- Extra benefits
  - Amazon: utilize off-peak capacity
  - Microsoft: sell .NET tools
  - Google: reuse existing infrastructure

Economics of Cloud Providers

- Regional prices vary, e.g.:

Price per KWH	Where	Why
3.6 cents	Idaho	Hydroelectric power, not sent long distance
10.0 cents	California	Electricity transmitted long distance over the grid; no coal fired electricity
18.0 cents	Hawaii	Must ship fuel to generate electricity

Opportunities for geographical, seasonal, re-distribution of resources, e.g., cooling unneeded in northern/southern hemisphere: cloud on a boat!

Adoption Challenges

Challenge	Opportunity
Availability	Multiple providers & DCs
Data lock-in	Standardization
Data Confidentiality and Auditability	Encryption, VLANs, Firewalls; Geographical Data Storage

Lock-in/Business Continuity

Challenge	Opportunity
Availability / business continuity	Multiple providers & datacenters Open API's

- Few enterprise datacenters' availability is as good
- "Higher level" (AppEngine, Force.com) vs. "lower level" (EC2) clouds include proprietary software
  - \+ richer functionality, better built-in ops support
  - – structural restrictions
- FOSS reimplementations on way? (eg AppScale)

Data Lock-in

Challenge	Opportunity
Data lock-in	Standardization

- FOSS implementations of storage (eg HyperTable)
- 10/19/09: Google Data Liberation Front

Growth Challenges

Challenge	Opportunity
Data transfer bottlenecks	FedEx-ing disks, Data Backup/Archival
Performance unpredictability	Improved VM support, flash memory, scheduling VMs
Scalable storage	Invent scalable store
Bugs in large distributed systems	Invent Debugger that relies on Distributed VMs
Scaling quickly	Invent Auto-Scaler that relies on Machine Learning; Snapshots

Data is a Gravity Well

Challenge	Opportunity
Data transfer bottlenecks	FedEx-ing disks, Data Backup/Archiving

- Amazon now provides "FedEx a disk" service
- and hosts free public datasets to "attract" cycles

See: http://aws.amazon.com/publicdatasets/
→ Possible interesting course projects here…

Data is a Gravity Well

Challenge	Opportunity
Scale-up/scale-down structured storage	Major research opportunity

•Profileration of *non-relational* scalable storage:
SQL Services (MS Azure), Hypertable, Cassandra, HBase, Amazon SimpleDB & S3, Voldemort, CouchDB, NoSQL movement

Policy and Business Challenges

Challenge	Opportunity
Reputation Fate Sharing	Offer reputation-guarding services like those for email
Software Licensing	Pay-for-use licenses; Bulk use sales

Policy and Business Challenges

Challenge	Opportunity
Reputation Fate Sharing	Offer reputation-guarding services like those for email

4/2/09: FBI raid on Dallas datacenter shuts down legitimate businesses along with criminal suspects

10/28/09: Amazon will whitelist elastic-IP addresses and selectively raise limit on outgoing SMTP

Policy and Business Challenges

Challenge	Opportunity
Software Licensing	Pay-as-you-go licenses; Bulk licenses

2/11/09: IBM pay-as-you-go Websphere, DB2, etc. on EC2

Windows on EC2

FOSS makes this less of a problem for some potential cloud users

Short Term Implications

- Startups and prototyping
- One-off tasks
  - Washington post, NY Times
- Cost associativity for scientific applications
- Research at scale

Long Term Implications

- Application software:
  - Cloud & client parts, disconnection tolerance
- Infrastructure software:
  - Resource accounting, VM awareness
- Hardware systems:
  - Containers, energy proportionality