RAD Lab
UC Berkeley

Beyond the Hype: A Berkeley View of Cloud Computing

Anthony D. Joseph*, EECS Dept, UC Berkeley
Reliable Adaptive Distributed Systems Lab

Toulouse – 14 January 2010

http://abovetheclouds.cs.berkeley.edu/

*Director, Intel Labs Berkeley

Image: John Curley http://www.flickr.com/photos/jay_que/1834540/

RAD Lab Cloud Computing Articles

http://abovetheclouds.cs.berkeley.edu/
Comm. of the ACM, Vol 53 Issue 4, April 2010

Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia

Cloud Computing: True Utility

Cloud Computing: App and Infrastructure over Internet

Software as a Service: Applications over the Internet

Utility Computing:
"Pay-as-You-Go" Datacenter Hardware and Software

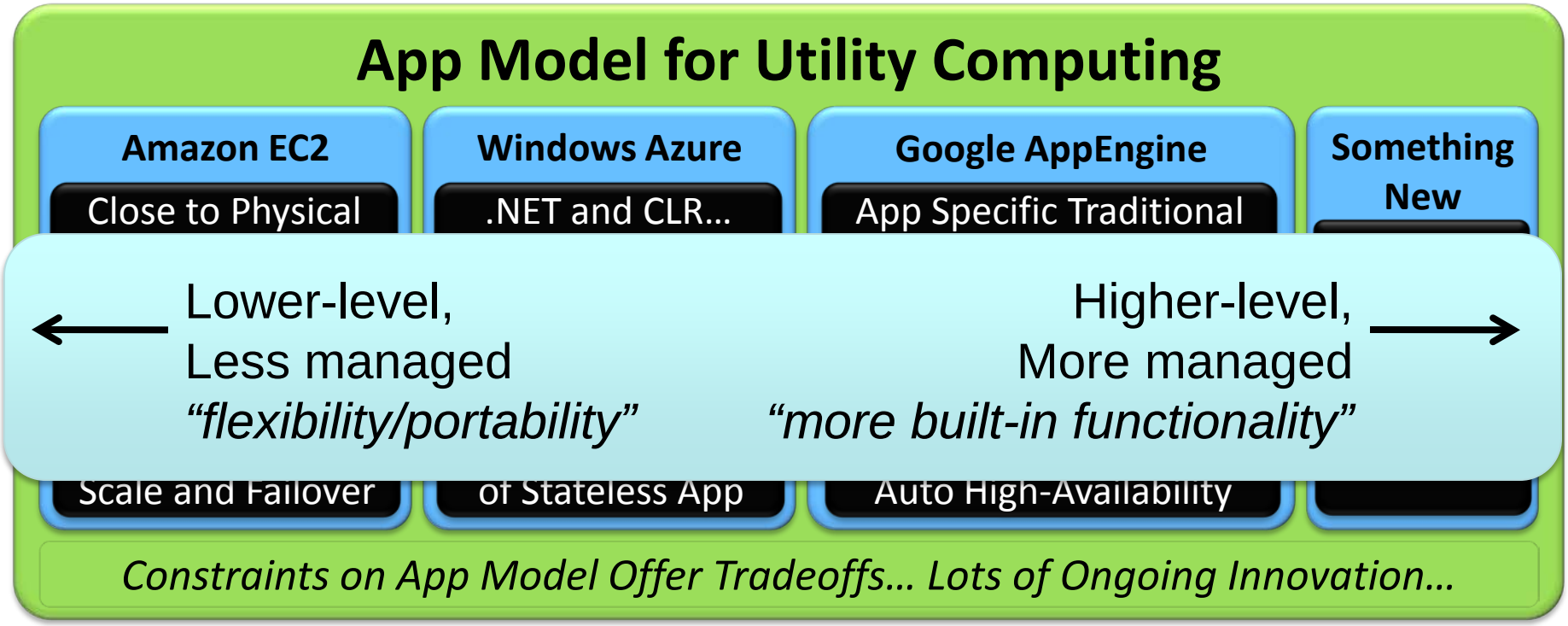
Three New Aspects to Cloud Computing

The Illusion of Infinite Computing Resources Available on Demand

The Elimination of an Upfront Commitment by Cloud Users

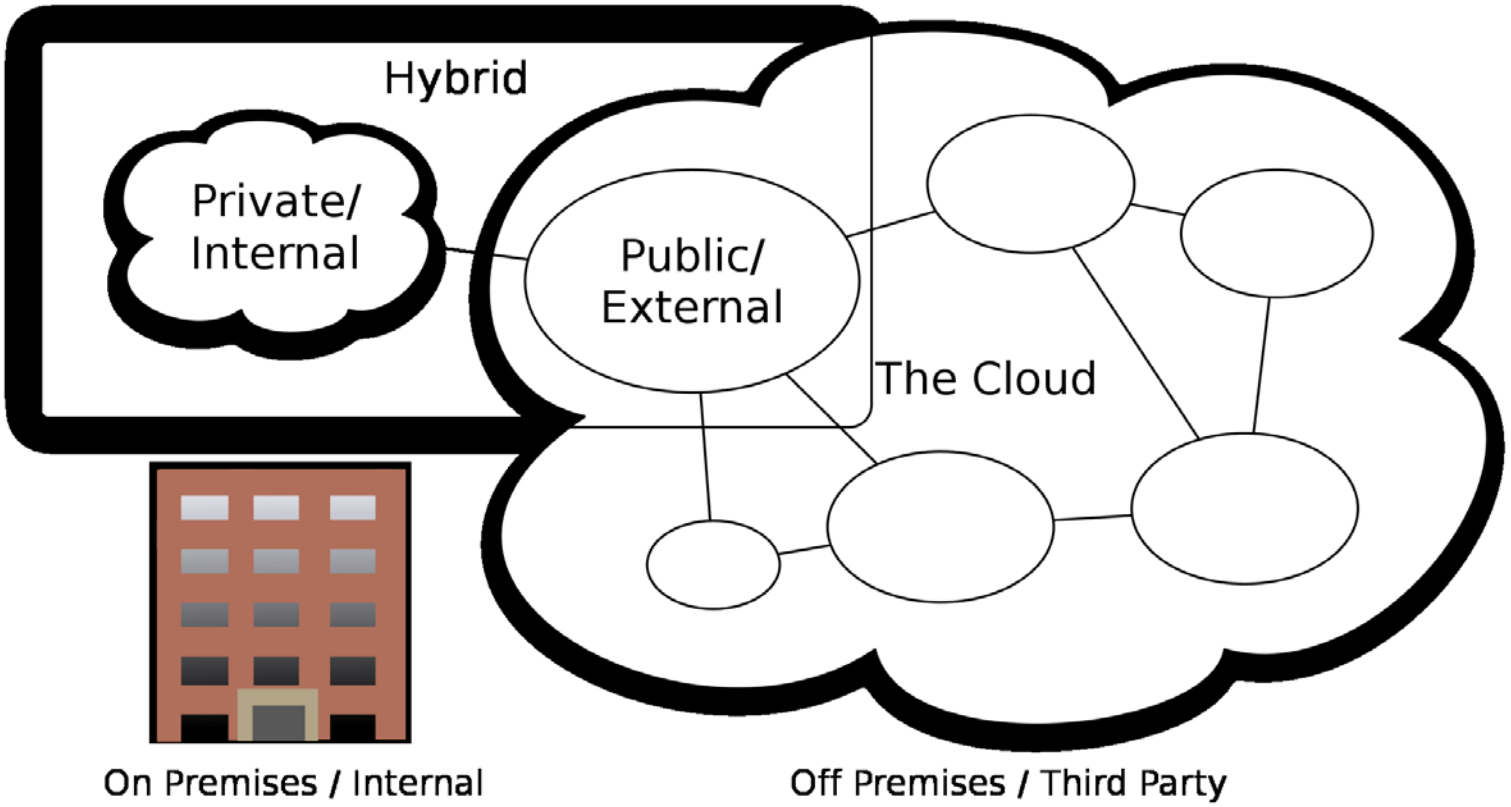
The Ability to Pay for Use of Computing Resources on a Short-Term Basis as Needed

Classifying Clouds



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

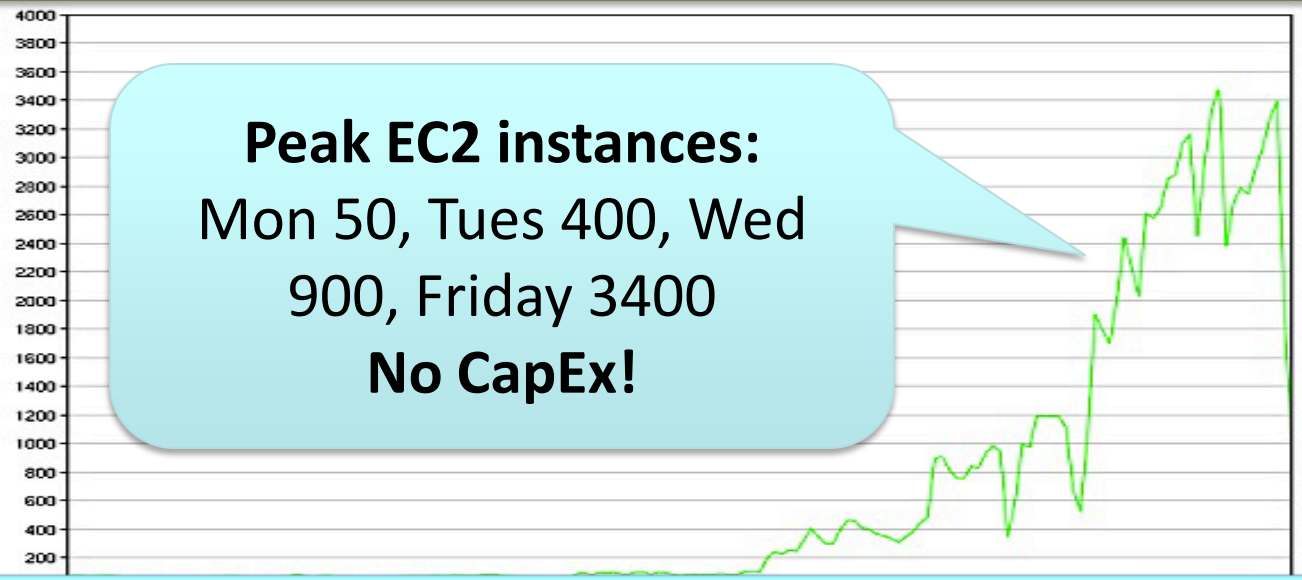
Private vs. Public



Cloud as Major Enabler

***Major enabler* for SW as a Service (SaaS) startups**

Animoto traffic doubled every 12 hours for 3 days
when released as Facebook plug-in in April 2008



Example #2: Target.Com (Large Retailer site on Amazon AWS)
28 Nov 2008 (Black Friday) – Many ECommerce Sites Failed
Target and Amazon Slower by Only About 50%

Why Now (not then)?

Economies of Scale for Humongous Datacenters

(1,000's to 10,000's of ***commodity*** computers)

Electricity	Network	Operations	Hardware
Put Datacenters at Cheap Power	Put Datacenters on Main Trunks	Standardize and Automate Ops	Containerized Low-Cost Servers

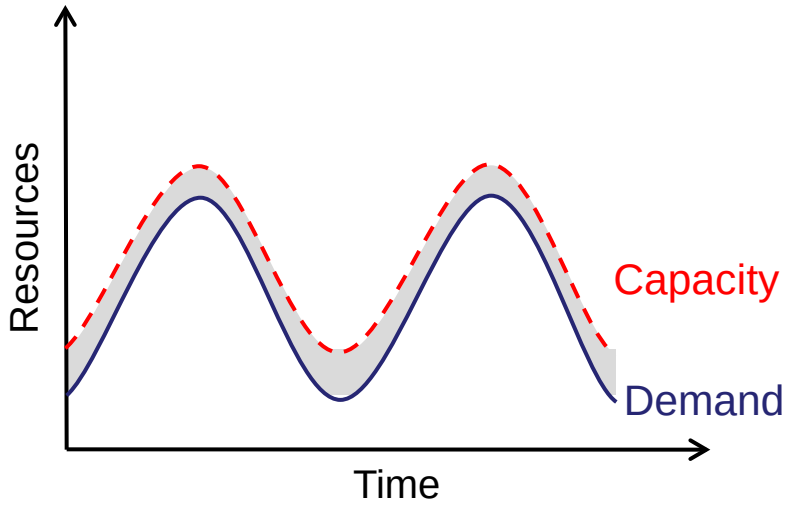
5 to 7 Times Reduction in the Cost of Computing vs. medium-sized (100's of machines) private facility

Public Clouds track cost trends better than private: Amazon EC2 price drop from $0.10 to $0.085

Benefits for Cloud Users

Static provisioning for peak: wasteful, but necessary for SLO

Risk of underutilization if peak predictions are too optimistic – Wasted CapEx

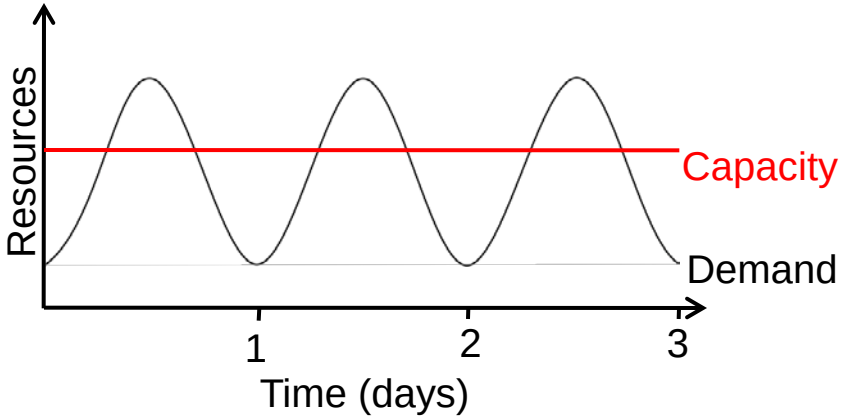


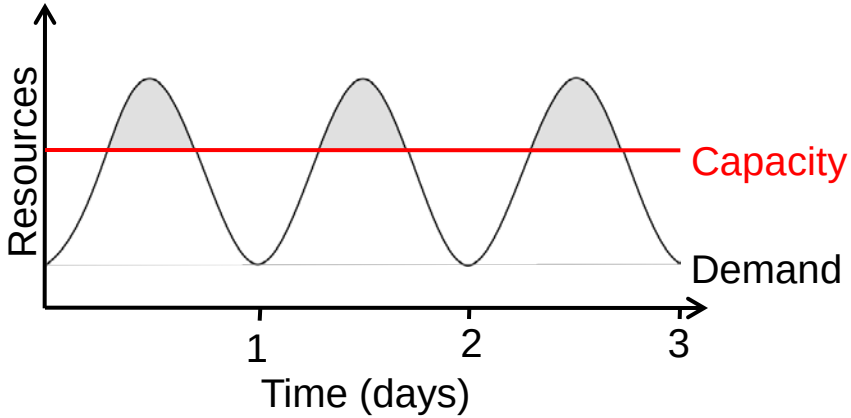
"Statically provisioned" data center

"Virtual" data center in the cloud

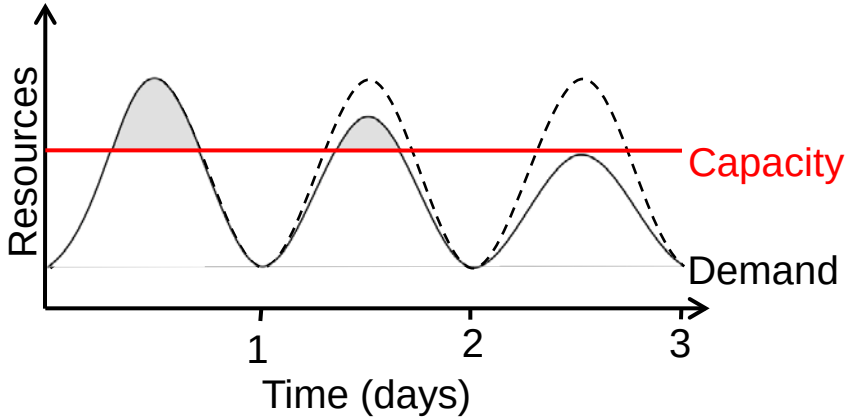
Unused resources

Risks of Underprovisioning





Lost revenue



Lost users

"Risk Transfer" Enables New Scenarios

More than (just) CapEx vs. OpEx!

"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 day after released
- RAD Lab graduate students demonstrate improved Hadoop (batch job) scheduler—on 1,000 servers

Adoption Challenges

Challenge	Opportunity
Availability of Service	Multiple providers and datacenters
Data lock-in	Standardization
Data Confidentialit	Encryption, VLANs, lls; Geographical torage

Open source reimplementations of Google AppEngine (AppScale), EC2 API (Eucalyptus), BigTable (HyperTable)

Growth Challenges

Challenge	Opportunity
Data transfer bottlenecks	FedEx-ing disks, Data ackup/Archival
Performance unpredictability	
Scalable storage	
Bugs in large distribu systems	
Scaling quickly	Invent Auto-Scaler that relies on ML; Snapshots

Freedom OSS partnership with Amazon to allow FedEx-ing disks into their datacenters, Amazon hosting free public datasets to "attract" cycles

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

2/11/09: IBM WebSphere™ and other service-delivery software available on AWS with *pay-as-you-go* pricing

Cloud or Earthbound: "Should I Move to the Cloud?"

Compelling Apps

- Surge computing: overflow into the cloud
- Extend desktop apps into cloud: Matlab, Mathematica
- Batch processing to exploit cost associativity, e.g. for business analytics

Challenged Apps

- Bulk data movement expensive, slow
- Jitter-sensitive apps (long-haul latency & transient performance distortion due to virtualization)

Summary

Economics of Cloud Computing Are Very Attractive to Some Users

Predicting Application Growth Hard

Cost-Associativity: Time is Money

Avoid Investment Risks from Peak Provisioning (CapEx -> OpEx)

Many Challenges: Availability, Data Gravity Well, …

http://abovetheclouds.cs.berkeley.edu/
Comm. of the ACM, Vol 53 Issue 4, April 2010

BACKUP SLIDES

Utility Computing Arrives

- Amazon Elastic Compute Cloud (EC2)
- "Compute unit" rental: ~~$0.10-0.80~~ 0.085-0.68/hour
  - 1 CU ≈ 1.0-1.2 GHz 2007 AMD Opteron/Intel Xeon core

	Platform	Units	Memory	Disk
Small - ~~$0.10~~ $.085/hour	32-bit	1	1.7GB	160GB
Large - ~~$0.40~~ $0.35/hour	64-bit	4	7.5GB	850GB – 2 spindles
X Large - ~~$0.80~~ $0.68/hour	64-bit	8	15GB	1690GB – 4 spindles
High CPU Med - ~~$0.20~~ $0.17	64-bit	5	1.7GB	350GB
High CPU Large - ~~$0.80~~ $0.68	64-bit	20	7GB	1690GB
High Mem X Large - $0.50	64-bit	6.5	17.1GB	1690GB
High Mem XXL - $1.20	64-bit	13	34.2GB	1690GB
High Mem XXXL - $2.40	64-bit	26	68.4GB	1690GB

- No up-front cost, no contract, no minimum
- Billing rounded to nearest hour (also regional,spot pricing)
- New paradigm(!) for deploying services?, HPC?

Northern VA cluster

Public vs. Private Clouds

- Building a Very Large-Scale Datacenter Very Is Expensive
  - $100+ Million (Minimum)
- Large Internet Companies Already Building Huge DCs
  - Google, Amazon, Microsoft…
- Large Internet Companies Already Building Software
  - MapReduce, GoogleFS, BigTable, Dynamo

Technology	Cost in Medium-Sized DC	Cost in Very Large DC	Ratio
Network	$95 per Mbit/sec/month	$13 per Mbit/sec/Month	7.1
Storage	$2.20 per GByte/month	$0.40 per Gbyte/month	5.7
Administration	≈ 140 Servers / Administrator	> 1000 Servers / Administrator	7.1

James Hamilton, *Internet Scale Service Efficiency*, Large-Scale Distributed Systems and Middleware (LADIS) Workshop Sept'08

Huge DCs 5-7X as Cost Effective as Medium-Scale DCs