RAD Lab
UC Berkeley

Cloud Computing: Obstacles & Opportunities

David Patterson, UC Berkeley
Reliable Adaptive Distributed Systems Lab

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

Outline

- What is Cloud Computing?
- What is new? Why has it happened now?
- Why good for users?
- Do cloud providers make money?
- Quick: Software as a Service / Cloud Computing in Education at UC Berkeley
- Quick: UC Berkeley RAD Lab Research Program in Cloud Computing
- Q&A

“Cloud computing is 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)

Above the Clouds: A Berkeley View of Cloud Computing

abovetheclouds.cs.berkeley.edu

- 2/09 White paper by RAD Lab PI's and students
  - Shorter version: "A View of Cloud Computing," *Communications of the ACM*, April 2010
  - Clarify terminology around Cloud Computing
  - Quantify comparison with conventional computing
  - Identify Cloud Computing challenges & opportunities
  - 60,000+ downloads of paper!
- Why can we offer new perspective?
  - Strong engagement with industry
  - Using cloud computing in research, teaching since 2008
- Goal: stimulate discussion on *what's really new*

Utility Computing Arrives

- Amazon Elastic Compute Cloud (EC2)
- "Compute unit" rental: $0.08-0.64/hr.
  - 1 CU ≈ 1.0-1.2 GHz 2007 AMD Opteron/Xeon core

"Instances"	Platform	Cores	Memory	Disk
Small - $0.08 / hr	32-bit	1	1.7 GB	160 GB
Large - $0.32 / hr	64-bit	4	7.5 GB	850 GB – 2 spindles
XLarge - $0.64 / hr	64-bit	8	15.0 GB	1690 GB – 3 spindles

- 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?

What is it? What's new?

- Old idea: Software as a Service (SaaS)
  - Basic idea predates MULTICS (timesharing in 1960s)
  - Software hosted in the infrastructure vs. installed on local servers or desktops; dumb (but brawny) terminals
- **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, no storage
  - *Public (utility)* vs. *private* clouds

Why Now (not then)?

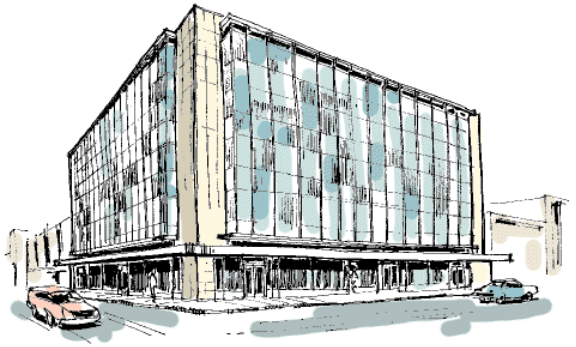
- "The Web Space Race": Build-out of extremely large datacenters (10,000's of ***commodity*** PCs)
  - Build-out driven by growth in demand (more users)

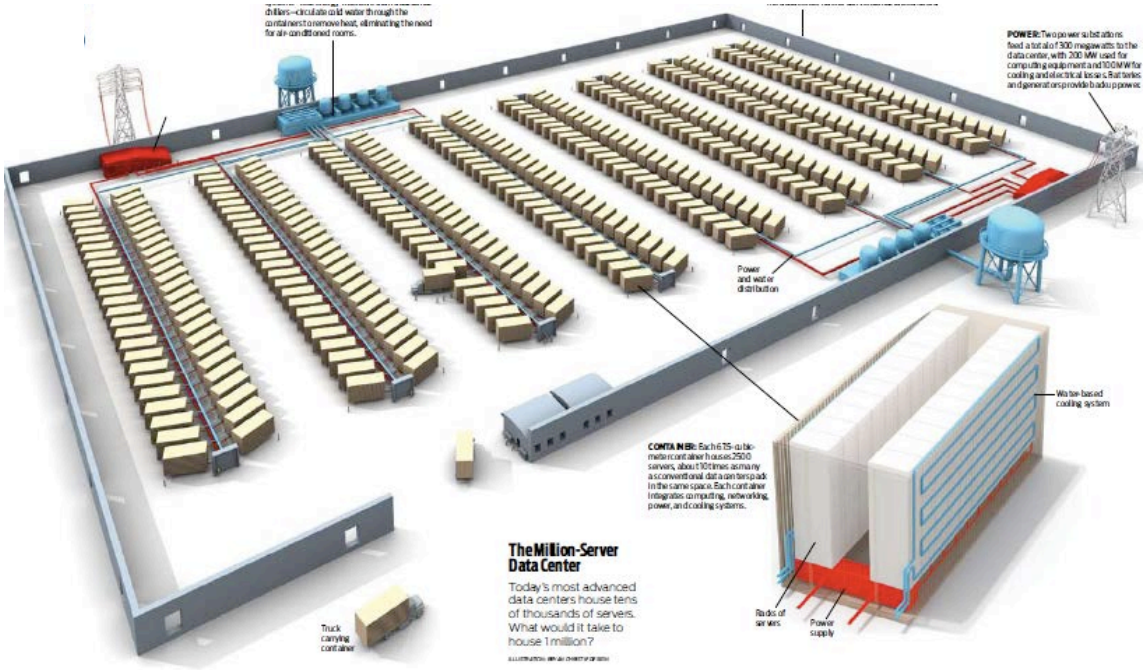
  => Infrastructure software: e.g., Google File System

  => Operational expertise: failover, DDoS, firewalls...
  - Discovered economy of scale: 5-7x cheaper than provisioning a medium-sized (1000 servers) facility
- More pervasive broadband Internet
- Commoditization of HW & SW
  - Fast Virtualization
  - Standardized software stacks

Datacenter is the new Server

Utility computing: enabling innovation in new services without first building & capitalizing a large company.



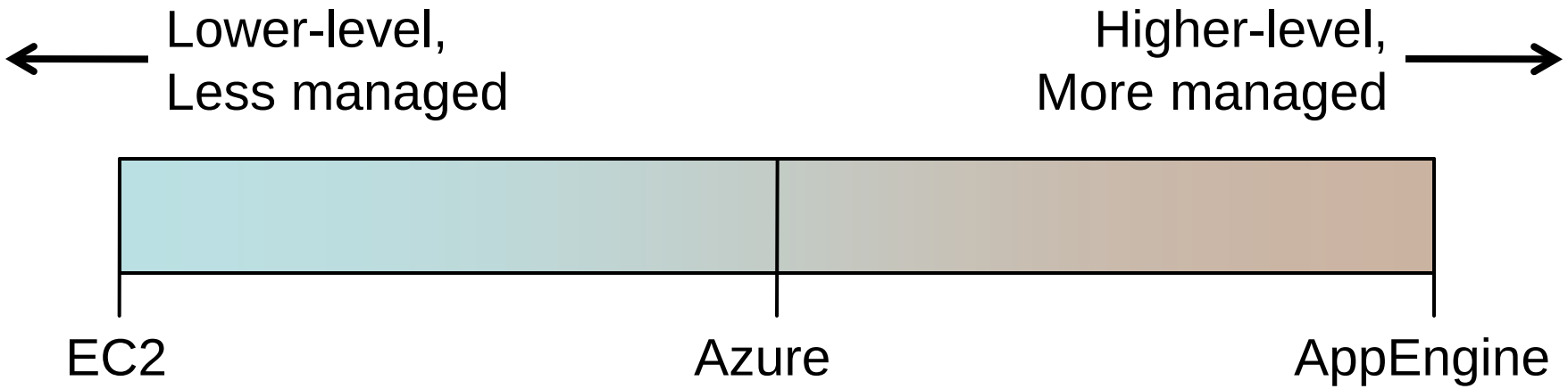


The Million Server Datacenter

- 24000 sq. m housing 400 containers
  - Each container contains 2500 servers
  - Integrated computing, networking, power, cooling systems
- 300 MW supplied from two power substations situated on opposite sides of the datacenter
- Dual water-based cooling systems circulate cold water to containers, eliminating need for air conditioned rooms

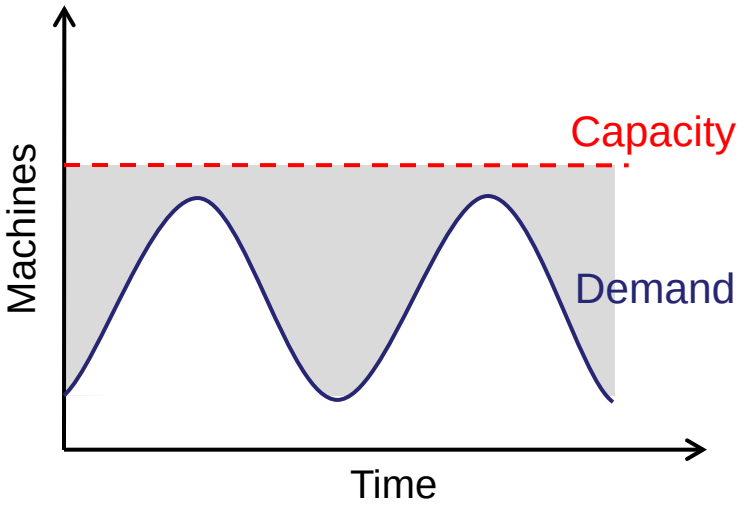
Classifying Clouds

- Instruction Set VM (Amazon EC2)
- Managed runtime VM (Microsoft Azure)
- Framework VM (Google AppEngine)
- *Tradeoff: flexibility/portability vs. "built in" functionality*

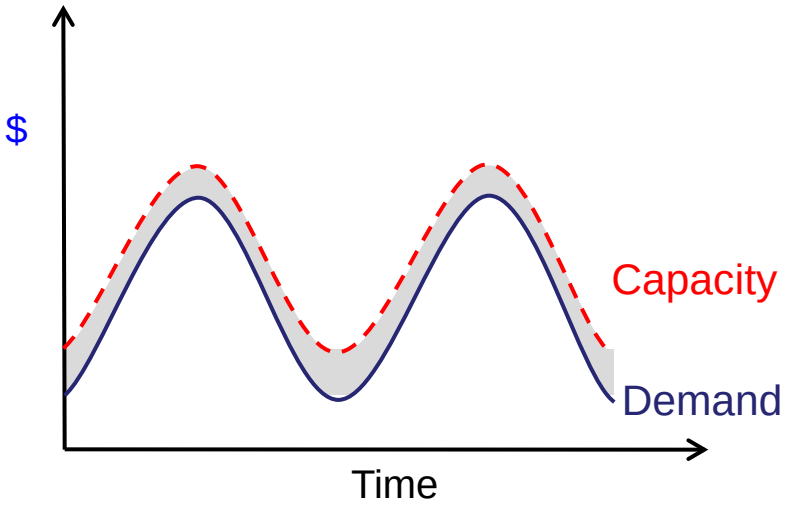


Cloud Economics 101

- Cloud Computing User: Static provisioning for peak - wasteful, but necessary for SLA



"Statically provisioned" data center

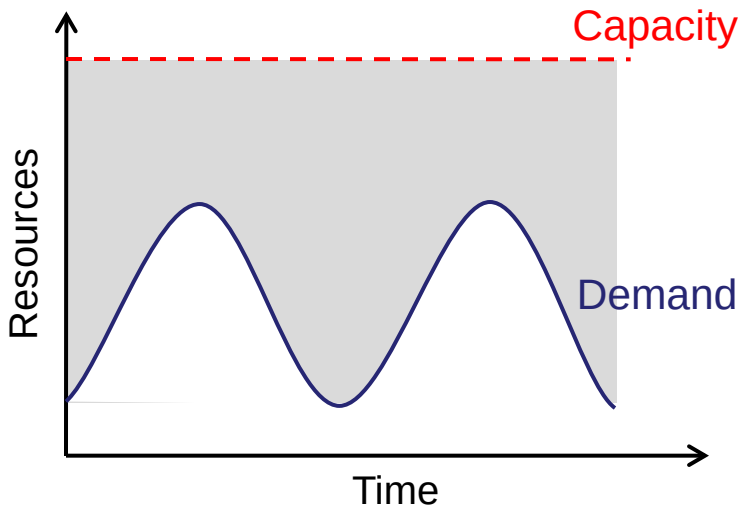


"Virtual" data center in the cloud

Unused resources

Risk of Under Utilization

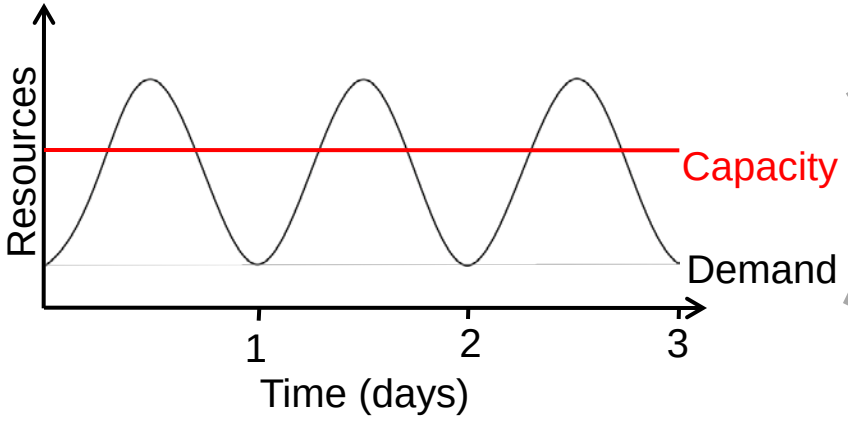
- Underutilization results if “peak” predictions are too optimistic

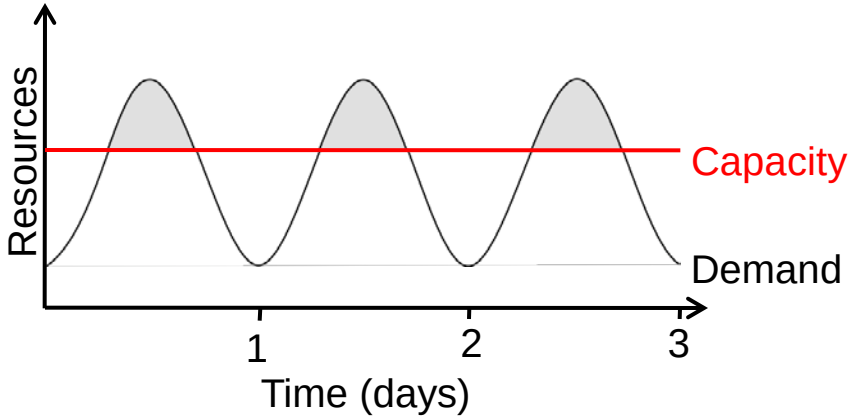


Unused resources

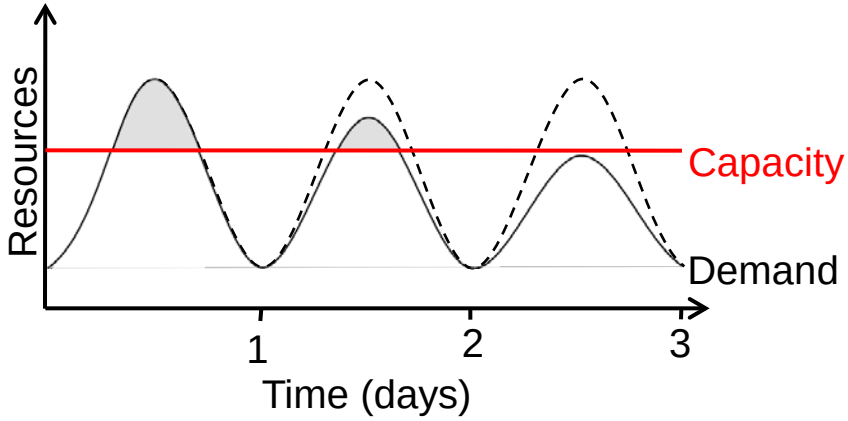
Static data center

Risks of Under Provisioning





Lost revenue



Lost users

New Scenarios Enabled by "Risk Transfer" to Cloud

- Not (just) Capital Expense vs. Operation Expense!
- "Cost associativity": 1,000 CPUs for 1 hour same price as 1 CPUs for 1,000 hours (@$0.08/hour)
  - Grad students demonstrate new idea on 1,000 servers
- *Major enabler* for SaaS startups
  - *Animoto* traffic 2X every 12 hours for 3 days when released as Facebook plug-in (50 to >3500 servers)
  - *FarmVille*: 1M users @ 4 days; 10M @ 60 days; 75M @ 270 days (28M daily users)
- Cloud gets IT gatekeepers out of the way
  - not unlike the PC revolution

Hybrid / Surge Computing

- Keep a local “private cloud” running same protocols as public cloud
- When need more, “surge” onto public cloud, and scale back when need fulfilled
- Saves capital expenditures by not buying and deploying power distribution, cooling, machines that are mostly idle

Do Cloud *Providers* Make $?

- James Hamilton Blog (now at Amazon)
  - http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx
- CAPEX for IT Equipment (46k servers), Warehouse, Power Distribution, Cooling
  - $80M for IT equip + $88M for building, infrastructure
- OPEX via Amortization (3 years for IT equip., 10 years for warehouse) + Electricity Costs
  - $3.6M/month => $0.11/hour/server
- Closest AWS server is High CPU XL @ $0.68/hr
- If sell 50% hours, gross margin of ~ 66%
  - Good margin for a service business

Energy & Cloud Computing?

- Cloud Computing saves Energy?
- Don’t buy machines for local use that are often idle
- Better to ship bits as photons over fiber vs. ship electrons over transmission lines to convert via local power supplies to spin disks and power processors and memories
  - Clouds use nearby (hydroelectric) power
  - Leverage economies of scale of cooling, power distribution

Energy & Cloud Computing?

- Techniques developed to stop using idle servers to save money in Cloud Computing can also be used to save power
  - Up to Cloud Computing Provider to decide what to do with idle resources
- New Requirement: Scale DOWN and up

Challenges & Opportunities

- “Top 10” Challenges to adoption, growth, & business/policy models for Cloud Computing
- Both technical and nontechnical
- Most translate to 1 or more *opportunities*
- Complete list in paper
- Paper also provides worked examples to quantify tradeoffs (“Should I move my service to the cloud?”)

5 Growth Challenges

Challenge	Opportunity
Programming for large distributed systems	MapReduce for batch processing, Major research opportunity
Scalable structured storage	Major research opportunity
Scaling quickly	Invent Auto-Scaler that relies on Statistical Machine Learning
Performance unpredictability (I/O)	Improved Virtual Machine support,scheduling, flash memory
Data transfer bottlenecks	FedEx-ing disks, Data Backup/Archival

3 Adoption Challenges

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

2 Policy and Business Challenges

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

Outline

- What is Cloud Computing?
- What is new? Why has it happened now?
- Why good for users?
- Do cloud providers make money?
- Quick: Software as a Service / Cloud Computing in Education at UC Berkeley
- Quick: UC Berkeley RAD Lab Research Program in Cloud Computing
- Q&A

Quick Overview Education

- Web 2.0 SaaS is a great motivator for teaching software skills
  - students get to build artifacts they themselves use
  - some projects continue after course is over
  - opportunity to (re-)introduce “big ideas” in software development/architecture
- Cloud computing is great fit for courses
  - elasticity around project deadlines
  - easier administration of courseware
  - students can take work product with them after course

RAD Lab 5-year Mission

*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: Fox/Katz/Patterson (systems/networks), Jordan (machine learning), Stoica (networks & P2P), Joseph (systems/security), Franklin (databases)
  - 2 postdocs, ~30 PhD students, ~10 undergrads

Machine Learning & Systems

- **Recurring theme**: cutting-edge Statistical Machine Learning (SML) works where simpler methods have failed
  - Predict performance of complex software system when demand is scaled up
  - Automatically add/drop servers to fit demand, without violating Service Level Objective (SLO)
  - Distill millions of lines of log messages into an operator-friendly “decision tree” that pinpoints “unusual” incidents/conditions

Conclusion

- Cloud Computing will transform IT industry
  - Pay-as-you-go utility computing leveraging economies of scale of Cloud provider
  - Anyone can create/scale next eBay, Twitter…
- Transform academic research, education too
- Cloud Computing offers $ for systems to scale down as well as up: save energy too
- RAD Lab addressing New Cloud Computing challenges