Cloud Computing and Data Management in the Cloud

INTRODUCTION

1

Outline

- Motivation
  - what is cloud computing?
  - what is cloud data management?
- Challenges, opportunities and limitations
  - what makes data management in the cloud difficult?
- New solutions
  - key/value, document, column family, graph, array, and object databases
  - scalable SQL databases
- Application
  - graph data and algorithms
  - usage scenarios

2

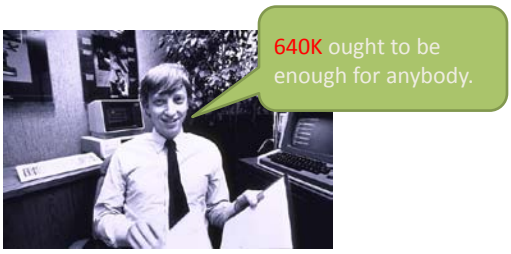
Cloud Computing

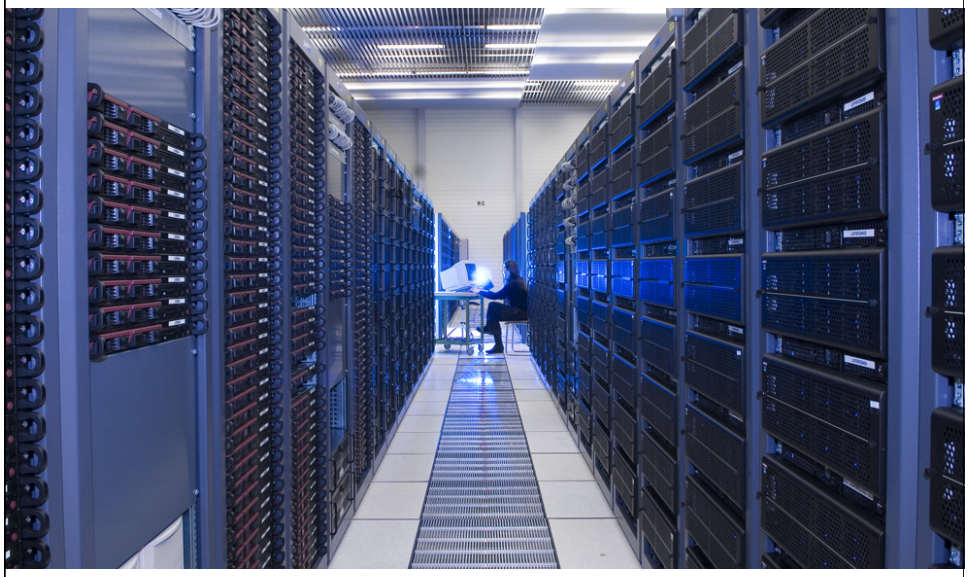
- Different definitions for "Cloud Computing" exist
  - http://tech.slashdot.org/article.pl?sid=08/07/17/2117221
- Common ground of many definitions
  - processing power, storage and software are **commodities** that are readily available from large infrastructure
  - **service-based view**: "everything as a service (*aaS)", where only "Software as a Service (SaaS)" has a precise and agreed-upon definition
  - utility computing: **pay-as-you-go** model

3

How much data? (from Jimmy Lin, U. Md)

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's LHC will generate 15 PB a year (??)





Maximilien Brice, © CERN

Cloud Computing Enterprises

rackspace HOSTING

amazon web services™

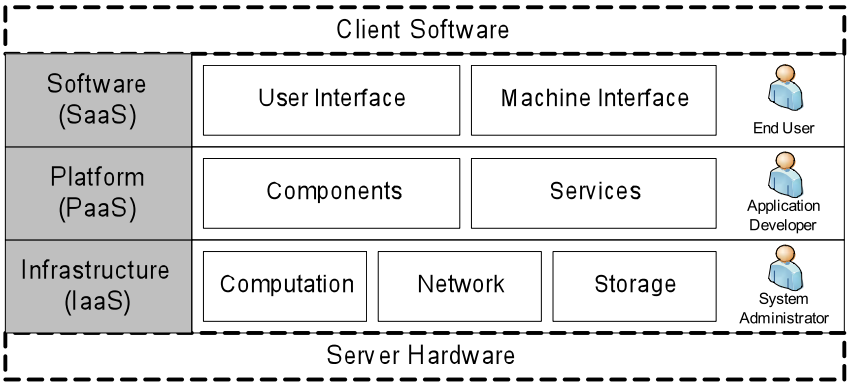
RIGHTSCALE™

Eucalyptus Systems™

elastra THE ENTERPRISE CLOUD COMPANY

hadoop

Service-Based View on Computing



Source: Wikipedia (http://www.wikipedia.org)

7

Terminology

- Term **cloud computing** usually refers to both
  - **SaaS**: applications delivered over the Internet as services
  - **The Cloud**: data center hardware and systems software
- Public clouds
  - available in a **pay-as-you-go** manner to the public
  - service being sold is **utility computing**
  - Amazon Web Service, Microsoft Azure, Google AppEngine
- Private clouds
  - internal data centers of businesses or organizations
  - normally not included under **cloud computing**

Based on: "Above the Clouds: A Berkeley View of Cloud Computing", RAD Lab, UC Berkeley

8

Utility Computing

- Illusion of infinite computing resources
  - available on demand
  - no need for users to plan ahead for provisioning
- No up-front cost or commitment by users
  - companies can start small
  - increase resources only when there is an increase in need
- Pay for use on short-term basis as needed
  - processors by the hour and storage by the day
  - release them as needed, reward conservation

Based on: "Above the Clouds: A Berkeley View of Cloud Computing", RAD Lab, UC Berkeley

9

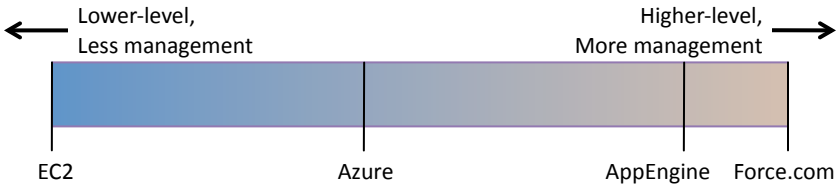
Virtualization

- Virtual resources abstract from physical resources
  - hardware platform, software, memory, storage, network
  - fine-granular, lightweight, flexible and dynamic
- Relevance to cloud computing
  - centralize and ease administrative tasks
  - improve scalability and work loads
  - increase stability and fault-tolerance
  - provide standardized, homogenous computing platform through hardware virtualization, i.e. **virtual machines**

10

Spectrum of Virtualization

- Computation virtualization
  - Instruction set VM (Amazon EC2, 3Tera)
  - Byte-code VM (Microsoft Azure)
  - Framework VM (Google AppEngine, Force.com)
- Storage virtualization
- Network virtualization

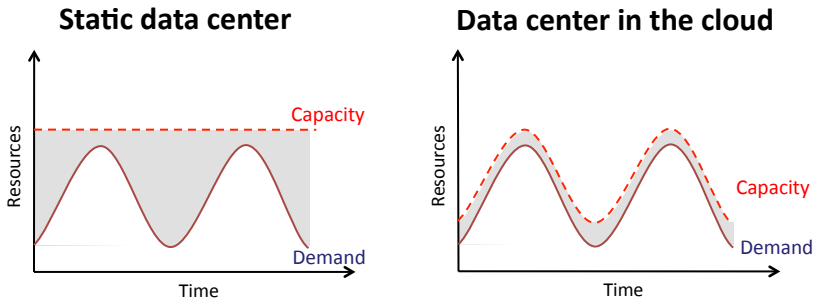


Slide Credit: RAD Lab, UC Berkeley

11

Economics of Cloud Users

- Pay by use instead of provisioning for peak



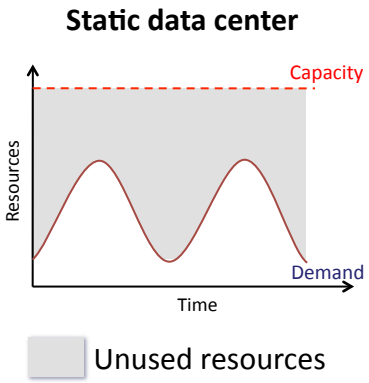
Unused resources

Slide Credit: RAD Lab, UC Berkeley

12

Economics of Cloud Users

- Risk of over-provisioning: underutilization

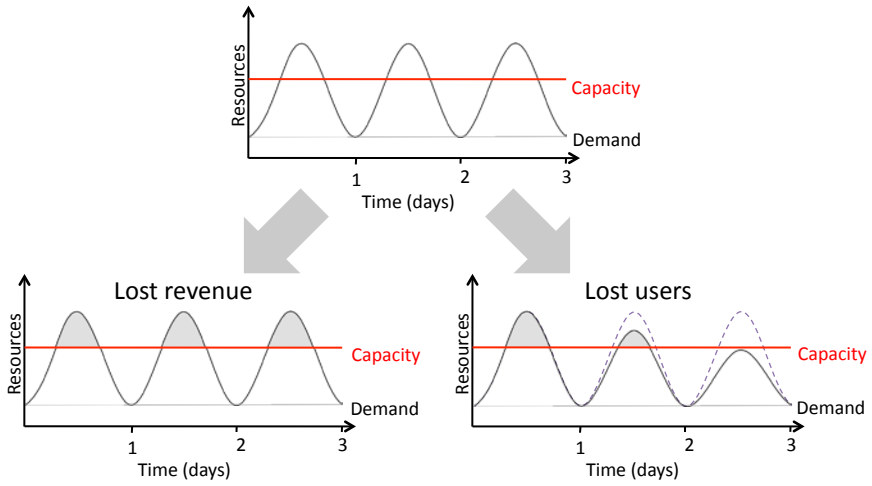


Slide Credit: RAD Lab, UC Berkeley

13

Economics of Cloud Users

- Heavy penalty for under-provisioning



Slide Credit: RAD Lab, UC Berkeley

14

Economics of Cloud Providers

Resource	Cost in Medium Data Center	Cost in Very Large Data Center	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

Source: James Hamilton (http://perspectives.mvdirona.com)

- Cloud computing is 5-7x cheaper than traditional in-house computing
- Added benefits
  - utilize off-peak capacity (Amazon)
  - sell .NET tools (Microsoft)
  - reuse existing infrastructure (Google)

Slide Credit: RAD Lab, UC Berkeley

15

Data Management in the Cloud

- Data management applications are potential candidates for deployment in the cloud
  - **industry:** enterprise database systems have significant up-front cost that includes both hardware and software costs
  - **academia:** manage, process and share mass-produced data in the cloud
- Many "Cloud Killer Apps" are in fact data-intensive
  - Batch Processing as with map/reduce
  - Online Transaction Processing (OLTP) as in automated business applications
  - Offline Analytical Processing (OLAP) as in data mining or machine learning

16

Scientific Data Management Applications

- Old model
  - "Query the world"
  - data acquisition coupled to a specific hypothesis
- New model
  - "Download the world"
  - data acquired en masse, in support of many hypotheses
- E-science examples
  - astronomy: high-resolution, high-frequency sky surveys, …
  - oceanography: high-resolution models, cheap sensors, satellites, …
  - biology: lab automation, high-throughput sequencing, ...

Slide Credit: Bill Howe, U Washington

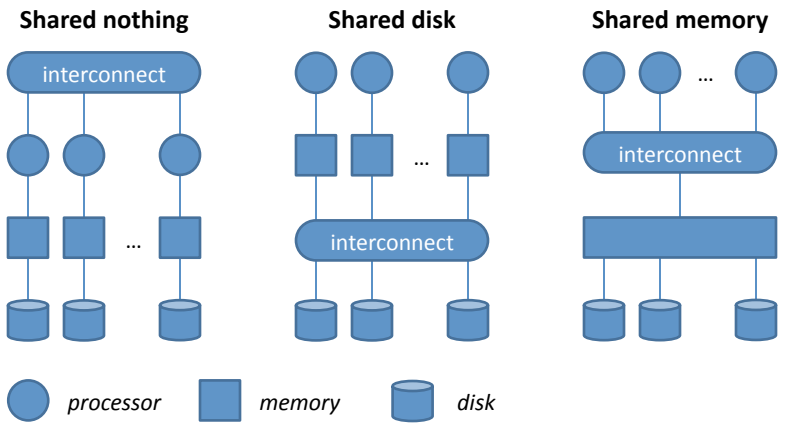
17

Scaling Data Management Systems

- Flavors of scalability
  - lots of (small) transactions
  - lots of copies of the data
  - lots of processor running on a single query (compute intensive tasks)
  - extremely large data set for one query (data intensive tasks)
- Data replication
  - move data to where it is needed
  - managed replication for availability and reliability

18

Parallel Database Architectures



Source: D. DeWitt and J. Gray: "Parallel Database Systems: The Future of High Performance Database Processing", CACM 36(6), pp. 85-98, 1992.

19

Revisit Cloud Characteristics

- Compute power is elastic, but only if workload is parallelizable
  - transactional data management systems do not typically use a shared-nothing architecture
  - shared-nothing is a good match for analytical data management
- Scalability
  - **in the past**: out-of-core, works even if data does not fit in main memory
  - **in the present**: exploits thousands of (cheap) nodes in parallel

Based on: "Data Management in the Cloud: Limitations and Opportunities", IEEE, 2009.

20

Revisit Cloud Characteristics

- Data is stored at an untrusted host
  - there are risks with respect to privacy and security in storing transactional data on an untrusted host
  - particularly sensitive data can be left out of analysis or anonymized
  - sharing and enabling access is often precisely the goal of using the cloud for scientific data sets

Based on: "Data Management in the Cloud: Limitations and Opportunities", IEEE, 2009.

21

Revisit Cloud Characteristics

- Data is replicated, often across large geographic distances
  - it is hard to maintain ACID guarantees in the presence of large-scale replication
  - full ACID guarantees are typically not required in analytical applications
- Virtualizing large data collections is challenging
  - data loading takes more time than starting a VM
  - storage cost vs. bandwidth cost
  - online vs. offline replication

Based on: "Data Management in the Cloud: Limitations and Opportunities", IEEE, 2009.

22

Challenges

- Scalability
  - today's SQL databases cannot scale to the thousands of nodes deployed in the cloud context
  - hard to support multiple, distributed updaters to the same data set
  - hard to replicate huge data sets for availability, due to capacity (storage, network bandwidth, …)
  - **storage:** different transactional implementation techniques, different storage semantics, or both
  - **query processing and optimization:** limitations on either the plan space or the search will be required
  - **programmability:** express programs in the cloud

Based on: "The Claremont Report on Database Research", 2008

23

Challenges

- Data privacy and security
  - protect from other users and cloud providers
  - specifically target usage scenarios in the cloud with practical incentives for providers and customers
- New applications: "*mash up*" interesting data sets
  - expect services pre-loaded with large data sets, stock prices, web crawls, scientific data
  - data sets from private or public domain
  - might give rise to federated cloud architectures

Based on: "The Claremont Report on Database Research", 2008

24