Queen's School of Computing

Queen's University

Cloud Computing
CISC 886

KHALID ELGAZZAR
GOODWIN 531
ELGAZZAR@CS.QUEENSU.CA

Learning Objectives

- ✓ Understand the motivation for, and the costs/benefits of, cloud computing.
- ✓ Become familiar with the issues surrounding big data.
- ✓ Become familiar with recent systems for the storage, processing and management of big data in the cloud.
- ✓ Achieve a depth of understanding of at least one data management research topic within cloud computing.
- ✓ Become familiar with key concepts and technologies in cloud computing.

Assumed Background

- ✓ Background in database management systems (CISC 432/832 or equivalent).
- ✓ Knowledge of distributed systems or service-oriented computing will be beneficial but is not required.

Housekeeping Stuff

Marking Scheme

- **Classroom participation (15%):**
  Students are expected to read all papers covered in a week, come to class prepared to discuss their thoughts and take part of the classroom discussions.
- **Paper presentation and discussion (20%):**
  Each paper will be assigned to two students; one will act as a presenter and the other as a discussant. The presentation will last 20 minutes and the discussion will last 15-20 minutes. Each student should upload their slides to the course wiki before the class.
- **Weekly critiques (20%):**
  Each student who is not assigned a role of presenter or discussant should pick one of the papers for that week and submit via email a one page critique of the paper before the start of class. The critique should offer a brief summary of the paper, points in favour, points against, and comments for improvement.
- **Project (45%):**
  One original project carried out individually or in a group of 2 students. The project will explore one or more of the topic areas covered in the course.

Paper Presentations

- **Role of presenter:** As a presenter you should not simply repeat the paper's content (remember you only have 15 minutes), instead you should point out the main important findings of the work. You should highlight any novel contributions, any surprises, and other possible applications of the proposed techniques. You should check the authors' other work related to the presented paper. Finally you should place the work relative other papers covered in the course (especially the papers covered in that particular week). You should prepare a presentation for the paper and submit the presentation, via email, at least one day prior to the presentation.
- **Role of discussant:** As a discussant, you should take an adversarial position by pointing out weak and controversial positions in the paper. You should present a short rebuttal of the paper and submit, via email, a summary of your rebuttal prior to the start of class. You should come prepared with problems and counterexamples for the presented work. *Make sure you list three things you liked and disliked about the paper.*

Project

The term project is an **IMPORTANT** part of this course. Projects can be conducted individually, or in teams of two students.

There are three types of projects that are acceptable in this course:

- **Research projects**
- **Survey projects**
- **Case study projects**

Project Deliverables

- **Project Proposal (15% of project mark) – Due Oct 10, 2014.**
- **Project Presentation (30% of project mark) – In last 2 weeks of the term.**
- **Project Report (55% of project mark) – Due Dec 5, 2014.**

You will be evaluated on the depth and novelty of your work, on the quality of your report, and on your research methodology (problem definition, choosing the correct level of abstraction, quality of implementation, experimentation methodology, etc.).

Academic Integrity

Academic integrity is constituted by the five core fundamental values of honesty, trust, fairness, respect and responsibility (see www.academicintegrity.org). These values are central to the building, nurturing and sustaining of an academic community in which all members of the community will thrive. Adherence to the values expressed through academic integrity forms a foundation for the "freedom of inquiry and exchange of ideas" essential to the intellectual life of the University (see the Senate Report on Principles and Priorities)

Students are responsible for familiarizing themselves with the regulations concerning academic integrity and for ensuring that their assignments conform to the principles of academic integrity. Information on academic integrity is available in the Arts and Science Calendar (see Academic Regulation 1), on the Arts and Science website (see http://www.queensu.ca/artsci/sites/default/files/Academic%20Regulations.pdf), and from the instructor of this course.

Departures from academic integrity include plagiarism, use of unauthorized materials, facilitation, forgery and falsification, and are antithetical to the development of an academic community at Queen's. Given the seriousness of these matters, actions which contravene the regulation on academic integrity carry sanctions that can range from a warning or the loss of grades on an assignment to the failure of a course to a requirement to withdraw from the university.

Introduction to Cloud Computing

Outline

1. What is cloud computing?
2. Why is cloud computing ***hot***?
3. Challenges and opportunities?
4. Key concepts and technologies
5. Challenges for data management

References

[1] L. Vaquero, L. Rodero-Merino, J. Caceres and M. Linder. **A Break in the Clouds: Towards a Cloud Definition**, *ACM SIGCOMM Computer Communication Review 39(1)*, 50 – 55, January 2009. http://portal.acm.org/citation.cfm?id=1496100

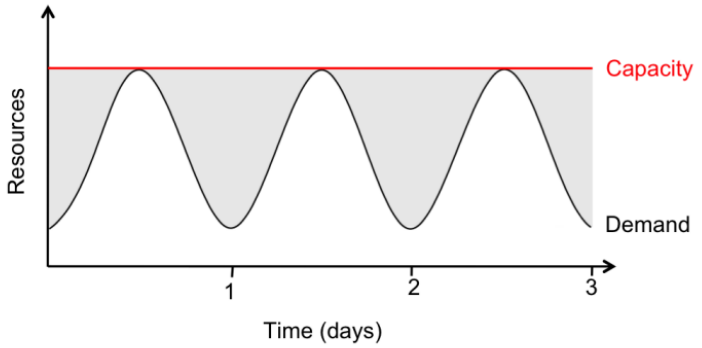
[2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia. **Above the Clouds: A Berkeley View of Cloud Computing**, Technical Report No. UCB/EECS-2009-28, Electrical Engineering and Computer Sciences, University of California at Berkeley, February 2009. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

1. What is Cloud Computing?

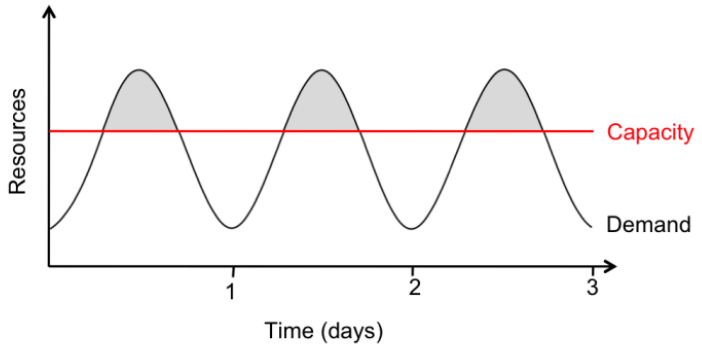
Key Properties

- Clouds are a large pool of ***virtualized resources***
- Resources can be ***dynamically reconfigured***
- Pay-per-use model (***utility computing***)

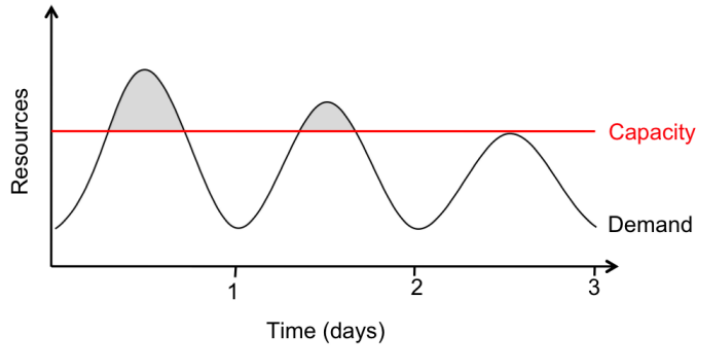
Resource Utilization



(a) Provisioning for peak load



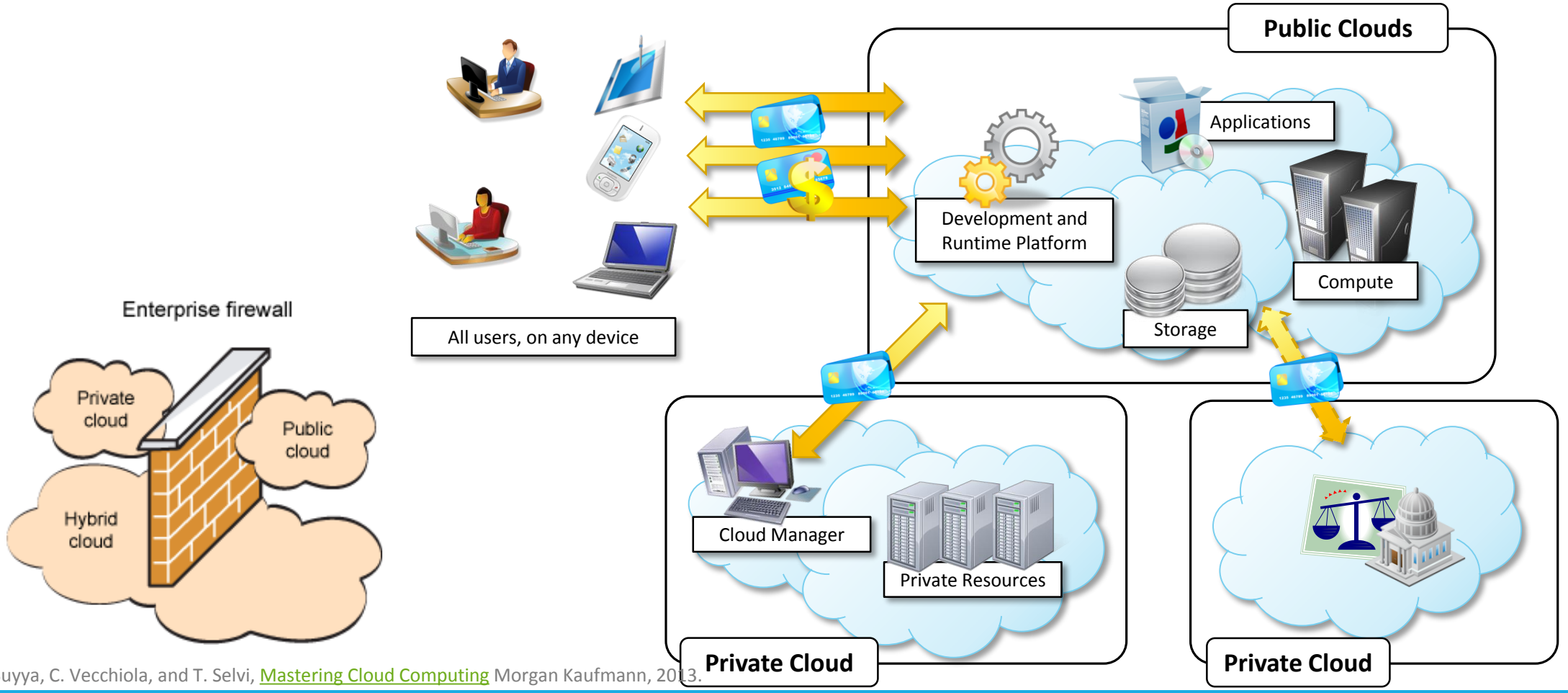
(b) Underprovisioning 1



(c) Underprovisioning 2

R. Buyya, C. Vecchiola, and T. Selvi, Mastering Cloud Computing Morgan Kaufmann, 2013.

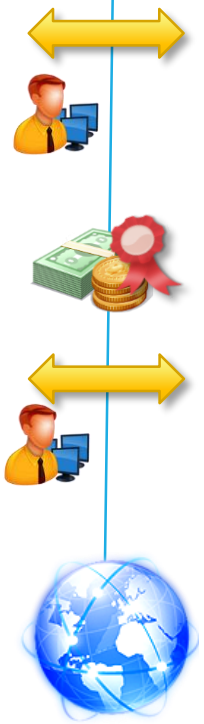
Types of Cloud



R. Buyya, C. Vecchiola, and T. Selvi, Mastering Cloud Computing Morgan Kaufmann, 2013.

Cloud Service Provision Models [?aaS]

Web 2.0 Interfaces



Software as a Service

Applications run by cloud and offered to users with on-demand licensing

Platform as a Service

Cloud provides software platform to run systems

Infrastructure as a Service

Infrastructure Providers use virtualization to dynamically assign resources

SaaS • Software

Office Communications Online · NETSUITE · salesforce.com · ZOHO Work. Online · workday.

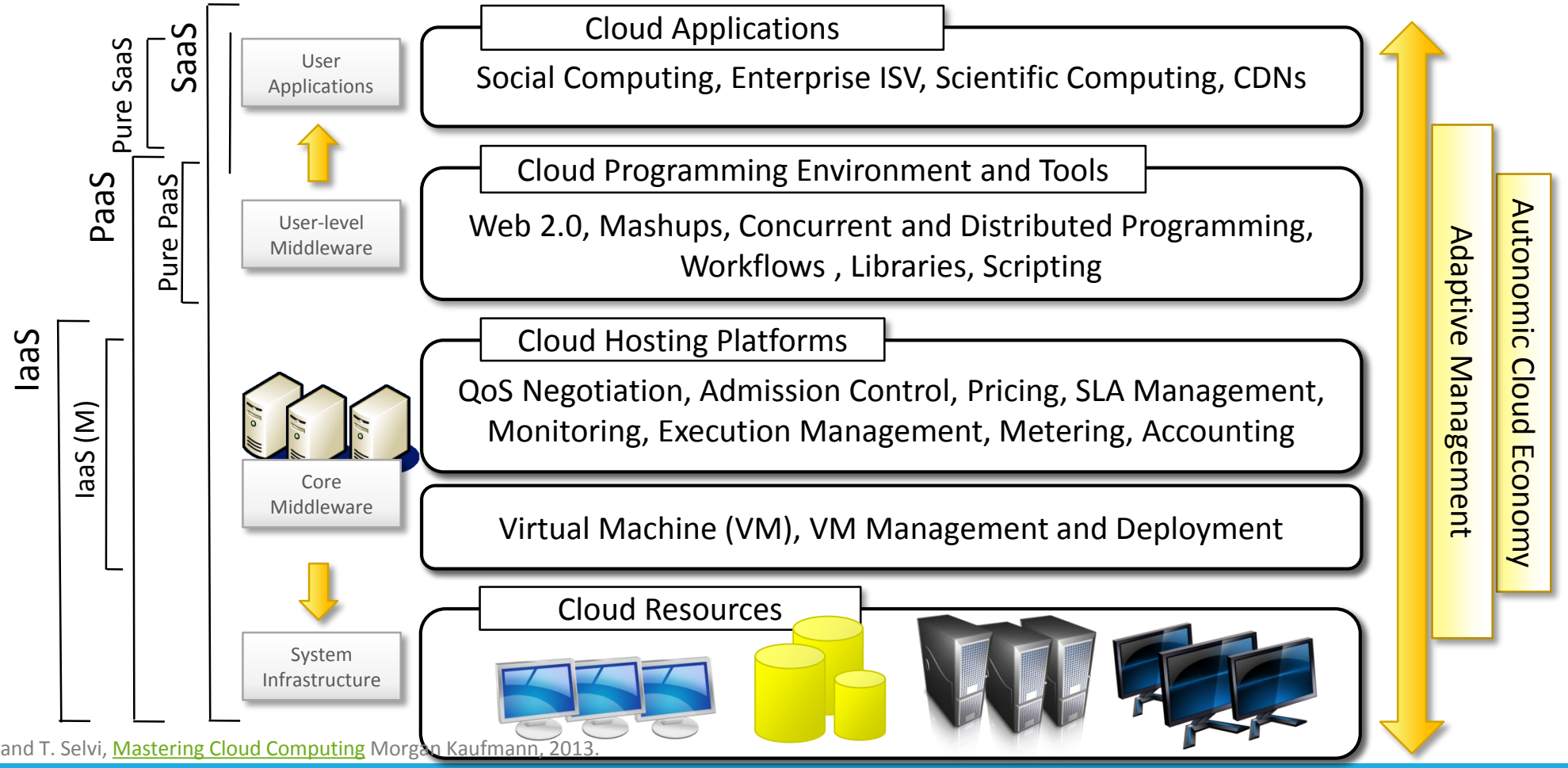
PaaS • Platform

Windows Azure Platform · Live Services · Google App Engine · force.com platform as a service

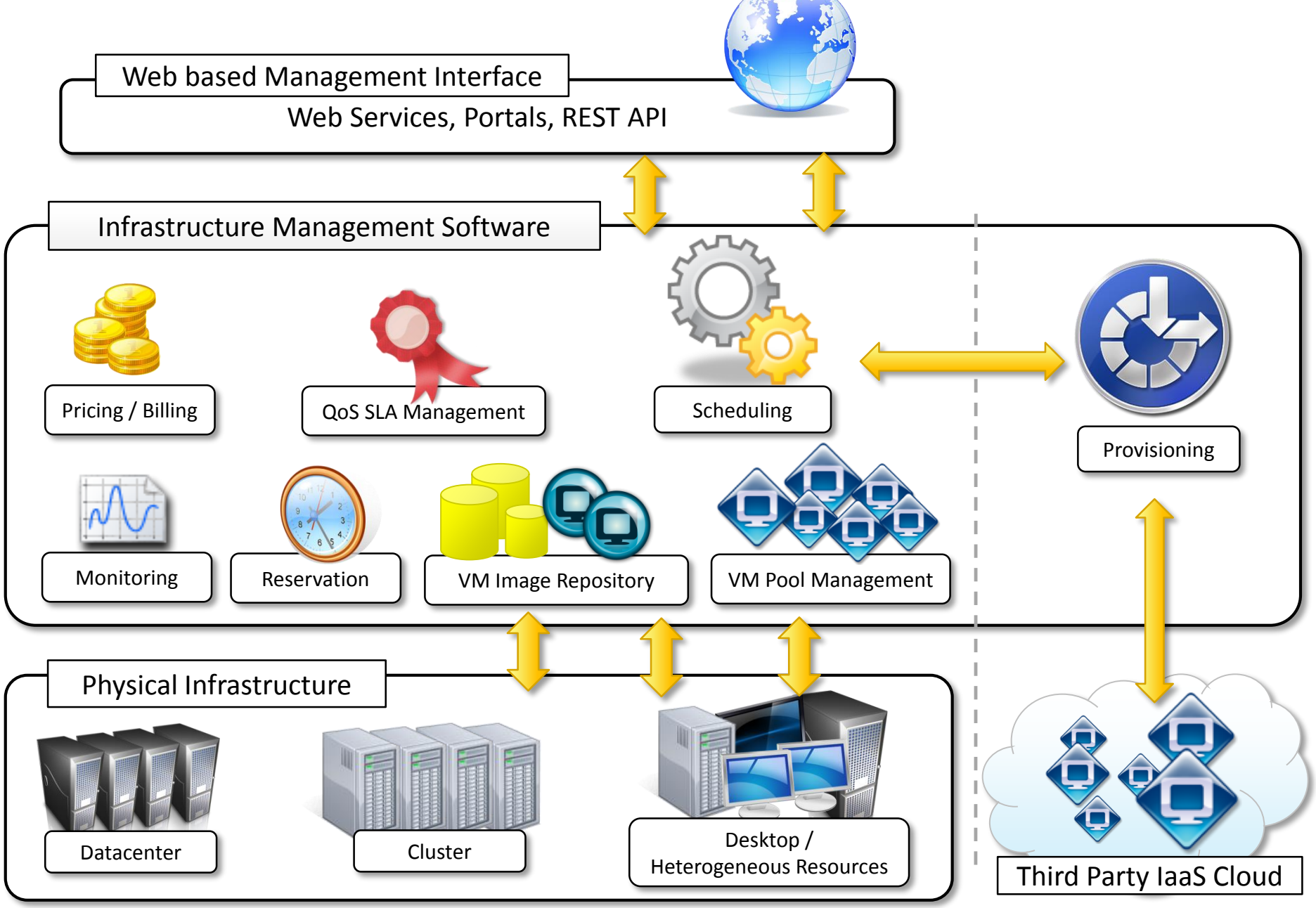
IaaS • Infrastructure

rackspacecloud · amazon web services · GOGRID · GIGASPACES

Cloud Layers



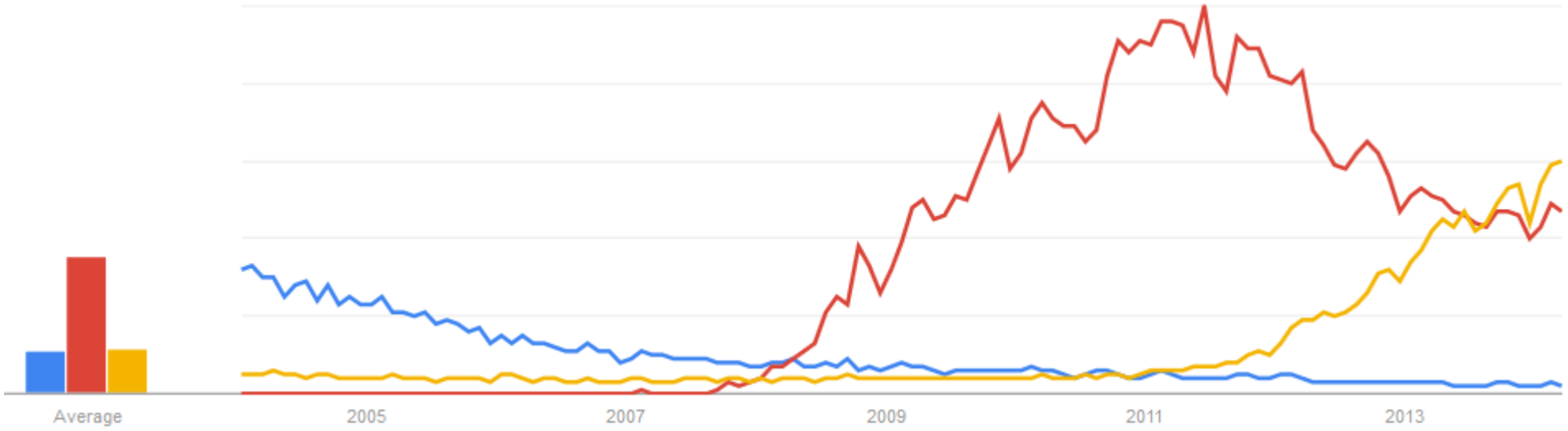
R. Buyya, C. Vecchiola, and T. Selvi, Mastering Cloud Computing Morgan Kaufmann, 2013.



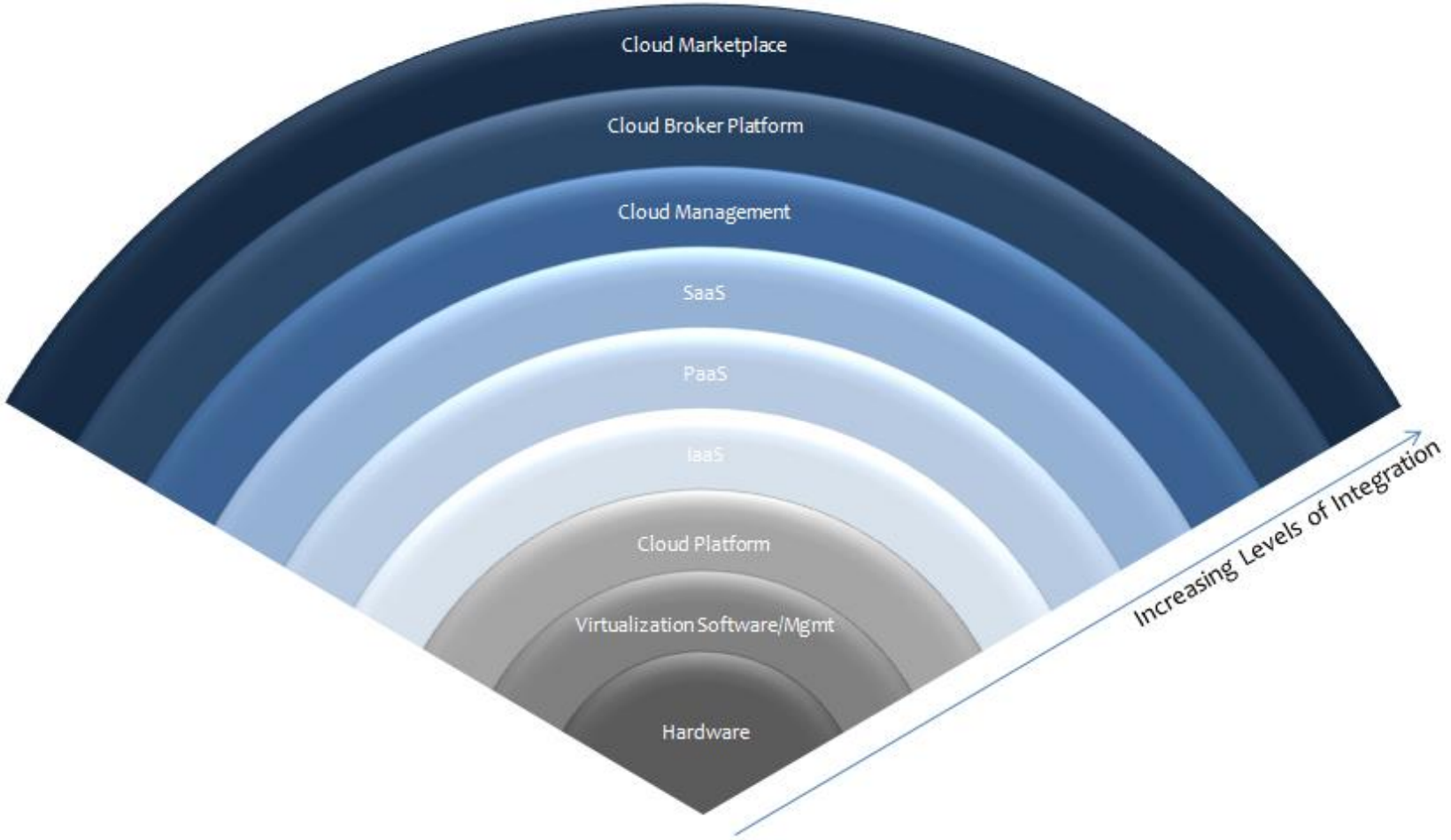
R. Buyya, C. Vecchiola, and T. Selvi, Mastering Cloud Computing Morgan Kaufmann, 2013.

Cloud Technology Spectrum

Interest over time
{grid, cloud, big data} computing



Cloud Technology Spectrum



Cloud Characteristics

Computing resources are elastic

- Service providers can quickly adapt to demand by reconfiguring their resources
- Workload must be parallelizable so additional resources can be exploited
  - Best for apps that run in shared-nothing environment

Computing resources shared among multiple users/providers

- Centralization of infrastructure and economies of scale
- Guarantees of privacy and security are needed

Cloud Characteristics (cont.)

Data is stored at an untrustworthy host

- Moving data offsite means potential security risks
- Data may be stored in a different country and under its regulations
- Service users required to encrypt their data.

Applications are service-oriented

- Compose applications out of loosely coupled services.

Data and services replicated across widely distributed sites

- Availability and durability achieved through replication
- Data needs to close to processing for performance reasons
- Users can access services / data from anywhere.

What's New?

Illusion of infinite computing resources available on demand

- Eliminates need to plan far-ahead for provisioning

Elimination of up-front commitment by cloud users

- Can start small and increase resources with demand

Ability to pay for computing resources only as needed

- Reward conservation by releasing resources

2. Why is cloud computing *HOT*?

New Technology Trends and Business Models

Increasing pressure on companies

- Heavier and more complex applications
- Higher availability demands
- Need to cut costs – people, energy, IT resources

Emergence of Web 2.0

- Shift from *contracted supported provisioning of services* to *do-it-yourself pay-as-you-go services*.

Companies like Google and Amazon developed very large data centers

- Offer *pay-as-you-go computing* with no contract
- Offer a customer their own *virtual machine*

New Application Opportunities

Mobile interactive applications

Parallel batch processing

Business analytics

Extension of compute-intensive desktop applications

3. Challenges and opportunities

1. Availability of a Service

Obstacle

- Organizations worry if utility computing will have adequate availability. Expect similar availability as services like Google search, Gmail
- Threat of distributed denial of service attacks

Challenge

- High-availability provided by high scalability and multiple cloud providers
- Cloud needs to provide DDoS protection

2. Data Lock-In

Obstacle

- Cloud computing APIs are essentially proprietary so customers cannot easily extract their data and programs to move to another provider

Challenge

- Standardize APIs
- Surge computing model

3. Data Confidentiality and Auditability

Obstacle

- Customers concerned about putting sensitive data out in the cloud for both security and audit reasons
- Data may be moved outside national boundaries so other countries laws apply

Challenge

- Add security and audit layers

4. Data Transfer Bottlenecks

Obstacle

- Parallelism makes data placement and transfer complicated
- Data transfer among clouds costly if across WAN

Challenges

- Develop new services to exploit data in cloud, eg backup to cloud and create searchable indices of archived data
- Improve WAN routers, LAN switches

5. Performance Unpredictability

Obstacle

- I/O sharing is problematic
- Standard scheduling of virtual machines not appropriate for HPC applications

Challenges

- Efficiently virtualize storage systems
- Use flash memory to reduce I/O interference
- Gang scheduling for HPC apps

6. Scalable Storage

Obstacle

- How are cloud properties of short-term usage, no up-front cost and infinite capacity on-demand supported with respect to persistent storage?

Challenge

- Open research problem to provide appropriate storage system

7. Bugs in Large Scale Distributed Systems

Obstacle

- Common that bugs cannot be reproduced in smaller configuration so debugging must occur at large scale

Challenge

- Need to develop new approaches to debugging. Eg, exploit use of virtual machines to capture more information that can help with debugging

8. Scaling Quickly

Obstacle

- Pay-as-you-go applies to storage and to network bandwidth, both of which count bytes used
- Computation is slightly different (by cycles or hours)

Challenge

- Releasing and acquiring resources only when necessary

9. Reputation Fate Sharing

Obstacle

- One customer's bad behavior can affect the reputation of the cloud as a whole
- Legal liability – cloud providers want liability to remain with the customer

Challenge

- Create reputation-guarding services similar to "trusted email" services

10. Software Licensing

Obstacle

- Current software licensing not a good match for clouds
- Pay-as-you-go not good for salespersons

Challenges

- Improve on use of open-source software
- Convince software companies to change licenses, eg pay-as-you-go software or prepaid plans for bulk use

4. Key Concepts and Technologies

Distributed Computing

Clouds support distributed computing models

- A problem is partitioned into subtasks and they are solved on independent (virtual) machines
- Coordination of subtasks via message passing

Two popular models

- MapReduce
- Service-Oriented Computing

MapReduce

Programming model developed by Google for processing large data sets.

Based on 2 functions:

- Data partitioned and given to instances of Map() that return result lists
- Reduce() instances process results from Map() instances to produce final result

Apache Hadoop is an open-source implementation of *MapReduce*

Service-Oriented Computing

Service-oriented computing uses *services* as fundamental components and develops apps by composing services

Service-Oriented Architecture (SOA) supports loosely-coupled, standards-based and protocol-independent distributed computing

Web services are best known SOA implementation

Utility Computing

Clouds are a way to provide utility computing

Utility computing is the packaging of computing resources as a metered service

- Similar to public utilities like electricity, water
- Customers pay only for what they use
- Resources used can vary with demand
- No initial costs for hardware, software, expertise

Virtualization

- Server presents illusion of many smaller *virtual machines (VMs)*
- Each VM has its own partition of server resources, its own instance of OS and possibly other software
- VMs isolated from each other
- Service providers acquire cloud resources through VMs
- Software like VMWare and Xen control VMs

Software-as-a-Service (SaaS)

- Clouds are a way to provide SaaS
- Model of software deployment where provider licenses software to a client as a service on demand
- Software hosted by provider and accessed via Web
- Frequently integrated into a larger collection of communicating software through a mashup or as a component of a platform-as-a-service

5. Challenges for Data Management

Challenges

- Large-scale data analysis and warehousing
- Elastic resource allocations
  - Adaptable query plans
  - Fault tolerant query execution
- Distributed data
  - Effective placement of replicas
  - Indexing strategies
- Shared-nothing servers
  - Query scheduling and execution
- Operating with outsourced data