Cloud Computing

Research Challenges Overview

Carlos Eduardo Moreira dos Santos

IME-USP, Brazil

May 3, 2010

Table of Contents I

Table of Contents II

What Is Cloud Computing?

Some technologies are often compared to Cloud Computing and, in fact, they share some characteristics.

Grid Computing

- Both make distributed resources available to an application.
- A Cloud uses virtualization to
  - share resources,
  - provision resources dynamically.

Virtualization

- Abstracts details of hardware
- Provides virtualized resources
- Foundation of Cloud Computing
  - Virtualized resources are (re)assigned to applications on-demand

Utility Computing

Charges customers based on usage, like electricity.

Cloud providers can

- maximize resource utilization
- minimize operating costs

using

- utility-based pricing
- on-demand resource provisioning

Autonomic Computing

- IBM, 2001
  - Self-managing systems
  - Internal and external observations
  - No human intervention
  - Aims to reduce complexity
- Cloud Computing
  - Automatic resource provisioning
  - Server consolidation
  - Aims to lower cost

Is It New?

Cloud Computing is not a new technology, but a new operations model that combines existing technologies.

Definition

Cloud computing is a model for enabling convenient, on-demand network access to a shared *pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be* rapidly provisioned and released *with minimal management effort or service provider interaction.* [NIST, 2009]

Business

- Illusion of infinite resources on demand
- No up-front commitment
- Ability to pay on a short-term basis
- Lower
  - risks (hardware failure)
  - maintainance costs
  - hardware expenses (fabless companies)

Business Model

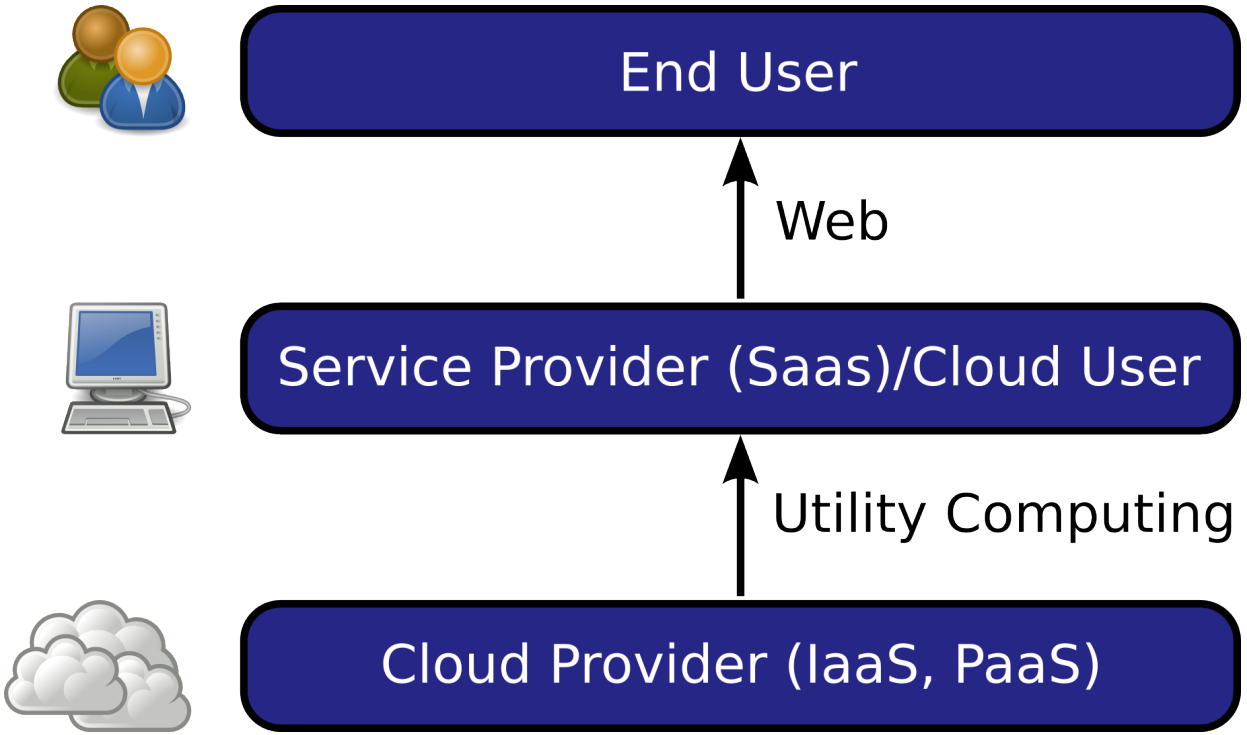

Figure: Cloud Computing business model

Elasticity

- Datacenters utilization
  - average: 5% to 20%
  - peak workload: 10% to 200%
- Provisioning errors
- Surges (Animoto - 50 to 3500 servers in 3 days)
- DDoS example: 1 GB/s attack, 500 bots/EC2 instance
  - Attacker (week): $500,000$ bots $\times \$0.03 = \$15,000$
  - Cloud (hour): $\$360 + \$100$ (32 hours)

Amazon EC2

- Virtual Machines on top of Xen
- Full control of software stack
- Images can be used to launch other VMs
- Conditions can trigger VM addition or removal
- Multiple locations: US, Europe, Asia

Google AppEngine

- Platform for traditional web applications
- Supports python and java
- Non-relational database
- Scaling is automatic and transparent

Microsoft Windows Azure

- Three components
  1. Windows based environment (applications, data)
  2. SQL Azure based on SQL Server
  3. .NET Services for distributed infraestructure
- Platform runs applications in the cloud or locally
- Supports .NET, C#, VB, C++, etc
- User must specify application needs to scale

Service Categories

- Infraestructure as a Service (IaaS)
  - On-demand resources, usually VMs
  - Amazon EC2
- Platform as a Service (PaaS)
  - Operating system support
  - Software development frameworks
  - Google App Engine
- Software as a Service (SaaS)
  - On-demand applications over the Internet
  - Salesforce (CRM)

Layers

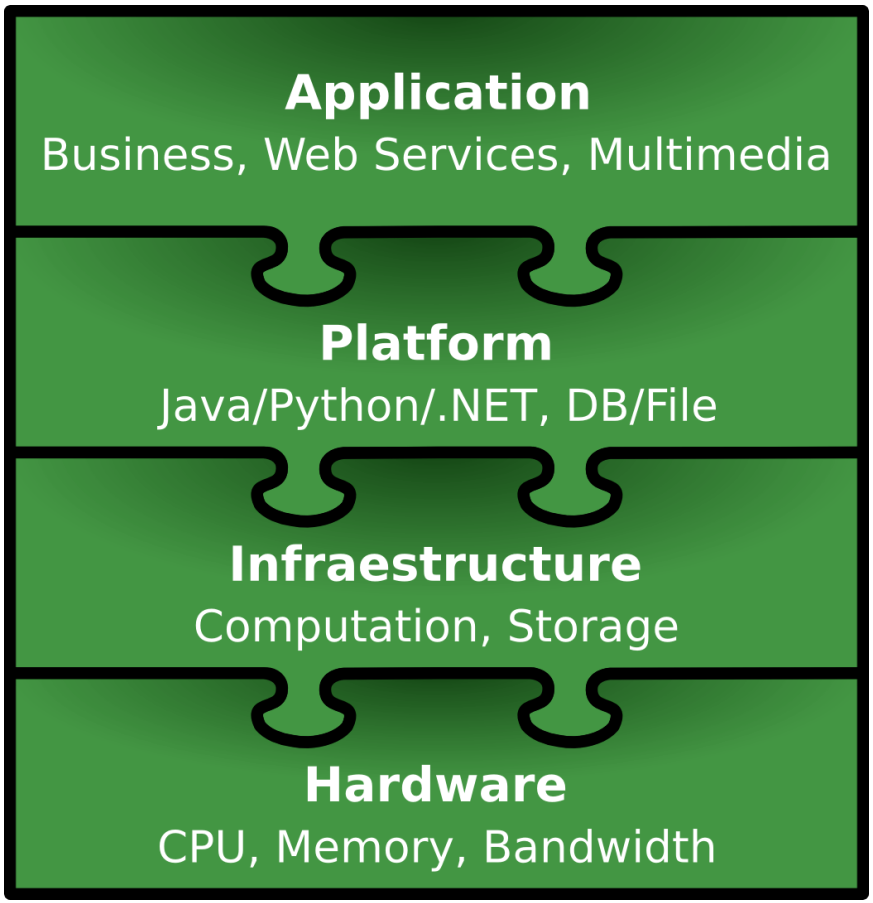

Figure: Architecture of Cloud Computing

Automated Service Provisioning

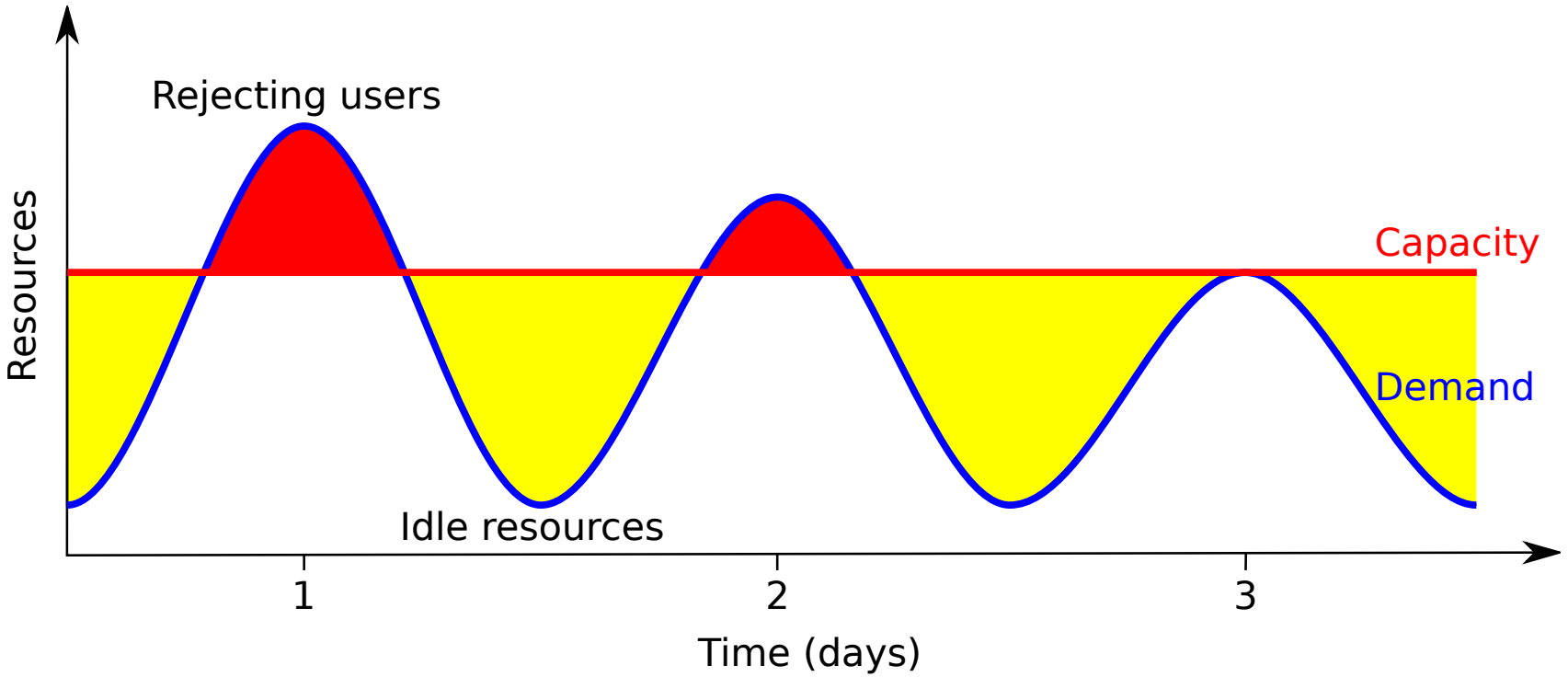

Figure: Fixed capacity problems. Most of the time opportunities are lost (red) or resources are wasted (yellow).

Automated Service Provisioning

Resource (de-)allocation to rapid demand flunctuations.

1. Predict number of instances to handle demand
   - Queuing theory
   - Control theory
   - Statistical Machine Learning
2. Predict future demand
3. Automatic resource allocation

Energy Management

Problems:

- Government regulations, environmental standards
- 53% of the cost is powering and cooling

Researches:

- Energy-efficient hardware
- Energy-aware job-scheduling
- Server consolidation
- Energy-efficient network protocols and infraestructures

Figure: HP Performance-Optimized Dataceter (POD): 40 feet and a capacity of 3,520 nodes or 12,000 LFF hard drives.

Figure: HP POD inside.

Figure: On the left, Microsoft's new $500 million Chicago data center. On the right, NASA cloud computing application hosted in a container and below it, IBM Portable Modular Data Center (PMDC).

Server Consolidation

- Maximize servers in energy-saving state
- VM dependencies (communication requirements)
- Resource congestions
- Variant of vector bin-packing problem (NP-hard)

Virtual Machine Migration

- Xen and VMWare "live" VM migration (<1 sec)
- Detecting hotspots versus sudden workload changes

Traffic analysis

- Important for management and planning decisions
- Much higher density of links
- Most existing methods have problems here:
  - Compute only a few hundreds end hosts
  - Assume flow patterns (MapReduce jobs)

Data Security

- Confidentiality for secure data access and transfer
  - Cryptographic protocols
- Auditability
  - Remote attestation (system state encrypted with TPM)
  - With VM migration, it is not sufficient
  - Virtualization platform must be trusted (SVMM)
  - Hardware must be trusted using hardware TPM
  - Efficient protocols are being designed

Software Frameworks

Large-scale data-intensive applications usually leverage MapReduce frameworks.

- Scalable
- Fault-tolerant

Challenges:

- Performance and resource usage depends on application
- Better performance and cost can be achieved by:
  - Selecting configuration parameter values
  - Mitigating the bottleneck resources
  - Adaptive scheduling in dynamic conditions
  - Performance modeling of Hadoop jobs
- Energy-aware:
  - A node should sleep while waiting new jobs
  - HDFS must be energy-aware, also

Novel Cloud Architectures

Small data centers can be more advantageous

- Less power (cooling)
- Cheaper
- Better geographically distributed (interactive gaming)

Using voluntary resources for cloud applications

- Much cheaper
- More suitable for non-profit applications
- Must deal with heterogeneous resources
- Machines can be turned on or off at any time
- How to incentivate resource donations?

Bibliography

Zhang, Q et al (2010)
Cloud computing: state-of-the-art and research challenges
Journal of Internet Services and Applications 1(1):7-18

Armbrust M et al (2009)
Above the Clouds: A Berkeley View of Cloud Computing
UC Berkeley Technical Report

Mell, Peter and Grance, Tim (2009)
The NIST Definition of Cloud Computing (v15)
http://csrc.nist.gov/groups/SNS/cloud-computing/

Data Center Knowledge
http://www.datacenterknowledge.com/archives/category/technology/containers/