Services in the Cloud

Louise Moser, *Member, IEEE*, Bhavani Thuraisingham, *Fellow, IEEE*, and Jia Zhang, *Member, IEEE*

1 Introduction

Cloud computing promotes the sharing of resources, to achieve economies of scale and to enable the end-users to focus on their core competencies, instead of the underlying computing technologies. Cloud resources not only are shared by multiple end-users, but also are dynamically allocated on demand.

Clouds can provide services in many different application domains, such as financial and retail services, engineering design and simulation, scientific investigations in molecular biology, particle physics, etc., online education and training, healthcare and personalized medicine, digital media and multi-player games, government services, etc.

The economic benefit of cloud computing is that it allows an organization to move from the CAPEX model to the OPEX model. In the CAPEX model, the organization purchases (or licenses) the hardware and software, and depreciates it over time. In the OPEX model, the organization essentially rents the hardware and/or software from a cloud provider when it is needed. Doing so allows the organization to avoid upfront hardware and software costs, as well as ongoing management and maintenance costs.

2 Underlying Cloud Technologies

Cloud computing depends on virtualization, which creates multiple logical (virtual) computers on top of a single physical computer. The virtual computers are used to perform multiple tasks concurrently, and enable the physical computer to be used more efficiently. Virtualization speeds up computing operations, and increases the utilization of the physical computers.

Cloud computing employs resource pooling and autonomic processing, to provide scalability, agility and elasticity. It involves dynamic provisioning (allocation and reallocation) of resources to multiple users on-demand. Tasks are run on multiple virtual machines to meet changing workloads, and load balancers are used to distribute the work across the virtual machines.

Cloud computing provides its resources as services. It adopts the idea of the service oriented architecture (SOA) which enables the end-users to integrate or compose their services to meet their computing needs.

3 Different Kinds of Cloud Services

The services that clouds provide are characterized as software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS), as illustrated in Fig. 1.

SaaS cloud providers offer applications and install, operate and maintain that software. End-users access the software in the cloud, which eliminates the need for the end-users to install and run the software on their own computers, and simplifies maintenance and support.

PaaS cloud providers offer computing platforms that include operating systems, program development and execution environments, Web servers and databases. End-users can develop and run their software on the cloud platform, without the need to purchase and manage their own platforms.

IaaS cloud providers offer physical computers, storage devices, and networks, as well as facilities that manage these resources. To deploy their applications, end-users install their operating system images and application software on the cloud infrastructure and maintain their own software.

An end-user accesses the cloud services using a Web browser, or software specific to the particular application, on the client platform. Client platforms include desktop computers, laptops, tablets and mobile phones.

4 Different Kinds of Clouds

Clouds are characterized as public, private and hybrid clouds, as illustrated in Fig. 2.

A public cloud involves services that are accessed by the general public over a public network, typically the Internet, in the cloud provider's data center. It typically involves multiple end-users, who are not related and who share the cloud services (multi-tenancy).

A private cloud involves services that are provided for a single organization, either internally or by a third party, and that are hosted internally or externally. Typically but not necessarily, the end-users communicate with a private cloud using a private network.

A hybrid cloud comprises both public and private clouds. Hybrid clouds present the challenges of both public and private clouds, as discussed below.

5 Challenges Presented by Clouds

The high degree of complexity of the cloud infrastructure, platform and software presents considerable challenges. Multi-tenancy in public clouds, and multiple and hybrid clouds, present even more challenges.

- *L. Moser is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106. E-mail: moser@ece.ucsb.edu.*
- *B. Thuraisingham is with the Department of Computer Science, University of Texas, Dallas, TX 75080. E-mail: bhavani.thuraisingham@utdallas.edu.*
- *J. Zhang is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Silicon Valley, CA 94035. E-mail: jia.zhang@sv.cmu.edu.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TSC.2015.2410351

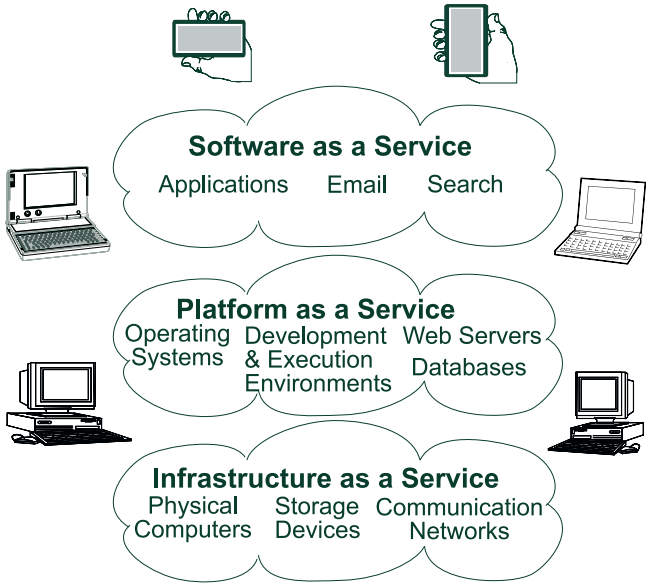

Fig. 1. Different kinds of cloud services.

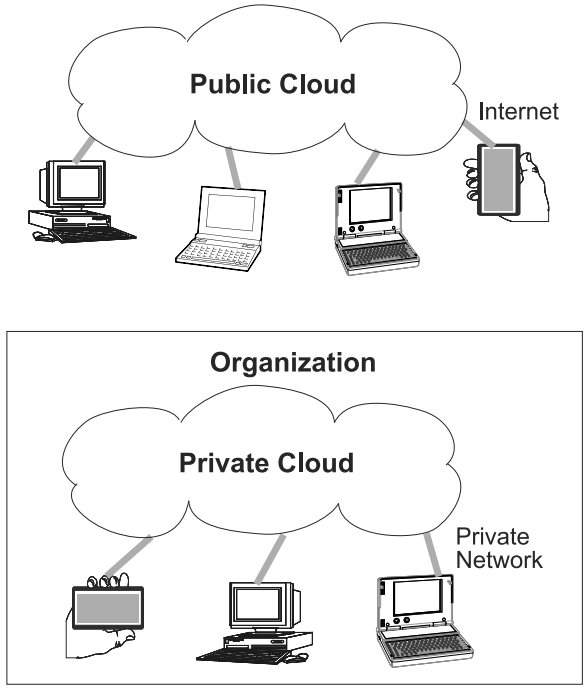

Fig. 2. Different kinds of clouds.

Maintaining availability of the data and processing is a challenge presented by clouds, particularly because the resources are located remotely and are shared by multiple end-users. Multiple clouds, with redundant hardware, software and data, can achieve higher availability, business continuity and diaster recovery. However, employing redundancy of the data and/or processing leads to the issues of out-of-date data and inconsistency.

Clouds present security/privacy challenges not only because there are multiple tenants sharing the same resources but also because the cloud provider has access to the end-users' data, and can alter or even delete that data. Moreover, the cloud provider can share the end-users' data with a third party, and can even profit from it. Security/privacy can be addressed by policy and legislation or by encryption to prevent unauthorized access. However, if the data are processed in the cloud, they need to be decrptyed first or, otherwise, processed while they are encrypted.

Other challenges presented by clouds relate to the heterogeneity of infrastructures, platforms and software within and across clouds, particularly public and hybrid clouds. To reduce lock-in to proprietary infrastructures, platforms and software, greater portability and interoperability are needed. That is, applications need to be portable from one kind of infrastructure or platform to another, and applications running on different kinds of infrastructures or platforms must be able to interoperate.

6 Research Topics for Clouds

Clouds are based on a number of technological advances that have occurred over many years. Although clouds are now being used commercially, many research topics still remain to be investigated.

As the end-users make increasing demands for richer and more capable services, new approaches to creating and providing such services are needed. Recently, there has been much research on the composition of services to produce more complex services from simpler services. However, further innovations are needed to yield services that go beyond current cloud offerings.

In a mobile cloud, for example, mobile phones, with limited storage and processing capabilities, exploit the more extensive storage and processing capabilities of the cloud. Data, such as photos from the mobile phone's camera, are stored in the cloud, and complex image processing is performed in the cloud to yield composite images, which are returned to the user's mobile phone.

In the future, sensors and actuators might have access to the cloud. Data from sensors are communicated to and processed in the cloud, and results are returned to actuators that affect the real-world environment. Such sensors and actuators might be used in electrical power grids, healthcare applications, etc.

For all three kinds of cloud services (SaaS, PaaS, IaaS), an important research topic is how to inform the platform of the needs of the applications, and how to inform the application of the capabilities of the platform. Providing such information might be done transparently through middleware, or it might involve management components that allow such information to be specified by the end-user.

Managing the data, including big data, different types of data, streaming data, sensor data, etc., requires improvements in structuring the data and metadata, as well as mechanisms for handling the data. In particular, non-proprietary metadata formats are needed to describe the services and resources. Such metadata formats can enable the discovery and composition of services by the end-users and can also enable the federation of clouds.

The key performance considerations of the cloud providers are high throughput, resource utilization and energy efficiency, whereas the key concerns of the end-users are low response time, high availability and security/privacy. With the growing requirements and expectations of the end-users, research on resource management that involves

the solution of complex optimization problems is needed. Moreover, in the presence of multiple tenants and mobile clients, providing low reponse times without long tails to the response time distribution is an important topic.

7 Articles in This Special Issue

This *IEEE TSC* Special Issue on Cloud Computing includes five articles that address some of these research topics.

The first article, "Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment," by X. Lin et al. addresses the task scheduling problem, which involves determining the tasks to be offloaded from the mobile device to the cloud, mapping the remaining tasks onto local cores of the mobile device, and scheduling the tasks on the local cores or communication channels. The objective is to satisfy the task precedence and application completion time requirements while minimizing the energy expended in the mobile device. The authors present an algorithm that starts with minimal delay scheduling and then performs energy reduction by migrating tasks among the local cores or between the local cores and the cloud. They also present a linear-time rescheduling algorithm for task migration.

In the second article, "Using ant colony system to consolidate VMs for green cloud computing," the authors F. Farahnakian et al. investigate the dynamic consolidation of virtual machines in a cloud data center to improve the resource utilization of physical machines and to reduce energy consumption by turning off unused physical machines. They also present a near-optimal virtual machine placement algorithm, along with experimental results using real workload traces from PlanetLab virtual machines.

In the third article, "Selecting optimum cloud availability zones by learning user satisfaction levels," the authors M. Unuvar et al. consider cloud providers that employ availability zones across multiple locations worldwide, to achieve higher availability and lower failure rates. More specifically, they present a predictive approach to identify the availability zone that maximizes user satisfaction against a set of user requirements.

In the fourth article, "PriDyn: Enabling differentiated I/O services in cloud using dynamic priorities," the authors N. Jain and J. Lakshmi consider the performance unpredictability and degradation that result from virtualization, due to the sharing of disk storage in the cloud. The authors present a scheduling framework, named PriDyn, that considers I/O performance metrics, such as latency, and converts them to priority values for disk access based on the current system state.

The last article, "Design support for performance aware dynamic application (re-)distribution in the cloud," by S. Gòmez Sàez et al. derives a set of functional and non-functional requirements and then presents a process-based approach to support the optimal distribution of applications in the cloud, in order to handle fluctuating workloads. The authors use the TCP-H benchmark to evaluate the performance of the applications under different deployment scenarios in the cloud.

Louise Moser received the PhD degree in mathematics from the University of Wisconsin, Madison. She is a professor in the Department of Electrical and Computer Engineering, University of California, Santa Barbara. Her research interests span the fields of services computing, distributed systems, computer networks, and software engineering. She has authored or coauthored more than 290 conference and journal publications. She has served as an associate editor for the *IEEE Transactions on Services Computing* and the *IEEE Transactions on Computers*, and as an area editor for *IEEE Computer*. She has also served as a technical chair and a general chair of the IEEE International Conference on Web Services, and the IEEE International Conference on Services Computing, and on many technical program committees. She is a member of the IEEE.

Bhavani Thuraisingham is the Louis A. Beecherl, Jr. distinguished professor in the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD) and the executive director in UTD's Cyber Security Research and Education Institute. Her current research focuses on integrating cyber security, social/cloud computing and big data analytics. She received several awards including the IEEE Computer Society's 1997 Technical Achievement Award, the ACM SIGSAC 2010 Outstanding Contributions Award, and a 2013 IBM Faculty Award. She is a fellow of the IEEE, the AAAS, and the British Computer Society.

Jia Zhang received the PhD degree in computer science from the University of Illinois at Chicago. She is an associate research professor at Carnegie Mellon University's Silicon Valley campus. Her research interests center on service oriented computing, with a focus on collaborative scientific workflows, Internet of Things, service-oriented architecture, and semantic service discovery. She has published more than 120 refereed journal papers, book chapters, and conference papers, and has co-authored a textbook titled *Services Computing*. Currently, she is an associate editor of the *IEEE Transactions on Services Computing*, an associate editor of the *International Journal of Web Services Research*, and editor-in-chief of the *International Journal of Services Computing*. She is a member of the IEEE.