Journal of Internet Services and Applications
May 2010, Volume 1, Issue 1, pp 7-18

Cloud computing: state-of-the-art and research challenges

Abstract

Cloud computing has recently emerged as a new paradigm for hosting and delivering services over the Internet. Cloud computing is attractive to business owners as it eliminates the requirement for users to plan ahead for provisioning, and allows enterprises to start from the small and increase resources only when there is a rise in service demand. However, despite the fact that cloud computing offers huge opportunities to the IT industry, the development of cloud computing technology is currently at its infancy, with many issues still to be addressed. In this paper, we present a survey of cloud computing, highlighting its key concepts, architectural principles, state-of-the-art implementation as well as research challenges. The aim of this paper is to provide a better understanding of the design challenges of cloud computing and identify important research directions in this increasingly important area.

J Internet Serv Appl (2010) 1: 7–18
DOI 10.1007/s13174-010-0007-6

ORIGINAL PAPERS

Cloud computing: state-of-the-art and research challenges

Qi Zhang · Lu Cheng · Raouf Boutaba

Received: 8 January 2010 / Accepted: 25 February 2010 / Published online: 20 April 2010

Abstract Cloud computing has recently emerged as a new paradigm for hosting and delivering services over the Internet. Cloud computing is attractive to business owners as it eliminates the requirement for users to plan ahead for provisioning, and allows enterprises to start from the small and increase resources only when there is a rise in service demand. However, despite the fact that cloud computing offers huge opportunities to the IT industry, the development of cloud computing technology is currently at its infancy, with many issues still to be addressed. In this paper, we present a survey of cloud computing, highlighting its key concepts, architectural principles, state-of-the-art implementation as well as research challenges. The aim of this paper is to provide a better understanding of the design challenges of cloud computing and identify important research directions in this increasingly important area.

Keywords Cloud computing · Data centers · Virtualization

1 Introduction

With the rapid development of processing and storage technologies and the success of the Internet, computing resources have become cheaper, more powerful and more ubiquitously available than ever before. This technological trend has enabled the realization of a new computing model called cloud computing, in which resources (e.g., CPU and storage) are provided as general utilities that can be leased and released by users through the Internet in an on-demand fashion. In a cloud computing environment, the traditional role of service provider is divided into two: the *infrastructure providers* who manage cloud platforms and lease resources according to a usage-based pricing model, and *service providers*, who rent resources from one or many infrastructure providers to serve the end users. The emergence of cloud computing has made a tremendous impact on the Information Technology (IT) industry over the past few years, where large companies such as Google, Amazon and Microsoft strive to provide more powerful, reliable and cost-efficient cloud platforms, and business enterprises seek to reshape their business models to gain benefit from this new paradigm. Indeed, cloud computing provides several compelling features that make it attractive to business owners, as shown below.

No up-front investment: Cloud computing uses a pay-as-you-go pricing model. A service provider does not need to invest in the infrastructure to start gaining benefit from cloud computing. It simply rents resources from the cloud according to its own needs and pay for the usage.

Lowering operating cost: Resources in a cloud environment can be rapidly allocated and de-allocated on demand. Hence, a service provider no longer needs to provision capacities according to the peak load. This provides huge savings since resources can be released to save on operating costs when service demand is low.

Highly scalable: Infrastructure providers pool large amount of resources from data centers and make them easily accessible. A service provider can easily expand its service to large scales in order to handle rapid increase in service demands (e.g., flash-crowd effect). This model is sometimes called surge computing [5].

Q. Zhang · L. Cheng · R. Boutaba (✉)
University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1
e-mail: rboutaba@cs.uwaterloo.ca

Q. Zhang
e-mail: q8zhang@cs.uwaterloo.ca

L. Cheng
e-mail: l32cheng@cs.uwaterloo.ca

No Body Text -- translate me!
Page 2

Easy access: Services hosted in the cloud are generally web-based. Therefore, they are easily accessible through a variety of devices with Internet connections. These devices not only include desktop and laptop computers, but also cell phones and PDAs.

Reducing business risks and maintenance expenses: By outsourcing the service infrastructure to the clouds, a service provider shifts its business risks (such as hardware failures) to infrastructure providers, who often have better expertise and are better equipped for managing these risks. In addition, a service provider can cut down the hardware maintenance and the staff training costs.

However, although cloud computing has shown considerable opportunities to the IT industry, it also brings many unique challenges that need to be carefully addressed. In this paper, we present a survey of cloud computing, highlighting its key concepts, architectural principles, state-of-the-art implementations as well as research challenges. Our aim is to provide a better understanding of the design challenges of cloud computing and identify important research directions in this fascinating topic.

The remainder of this paper is organized as follows. In Sect. 2 we provide an overview of cloud computing and compare it with other related technologies. In Sect. 3, we describe the architecture of cloud computing and present its design principles. The key features and characteristics of cloud computing are detailed in Sect. 4. Section 5 surveys the commercial products as well as the current technologies used for cloud computing. In Sect. 6, we summarize the current research topics in cloud computing. Finally, the paper concludes in Sect. 7.

2 Overview of cloud computing

This section presents a general overview of cloud computing, including its definition and a comparison with related concepts.

2.1 Definitions

The main idea behind cloud computing is not a new one. John McCarthy in the 1960s already envisioned that computing facilities will be provided to the general public like a utility [39]. The term "cloud" has also been used in various contexts such as describing large ATM networks in the 1990s. However, it was after Google's CEO Eric Schmidt used the word to describe the business model of providing services across the Internet in 2006, that the term really started to gain popularity. Since then, the term cloud computing has been used mainly as a marketing term in a variety of contexts to represent many different ideas. Certainly, the lack of a standard definition of cloud computing has generated not only market hypes, but also a fair amount of skepticism and confusion. For this reason, recently there has been work on standardizing the definition of cloud computing. As an example, the work in [49] compared over 20 different definitions from a variety of sources to confirm a standard definition. In this paper, we adopt the definition of cloud computing provided by The National Institute of Standards and Technology (NIST) [36], as it covers, in our opinion, all the essential aspects of cloud computing:

NIST definition of cloud computing *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.*

The main reason for the existence of different perceptions of cloud computing is that cloud computing, unlike other technical terms, is not a new technology, but rather a new operations model that brings together a set of existing technologies to run business in a different way. Indeed, most of the technologies used by cloud computing, such as virtualization and utility-based pricing, are not new. Instead, cloud computing leverages these existing technologies to meet the technological and economic requirements of today's demand for information technology.

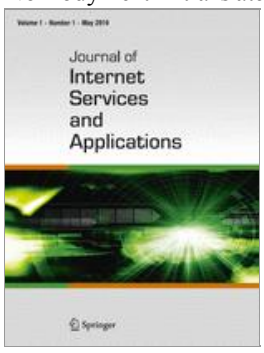
2.2 Related technologies

Cloud computing is often compared to the following technologies, each of which shares certain aspects with cloud computing:

Grid Computing: Grid computing is a distributed computing paradigm that coordinates networked resources to achieve a common computational objective. The development of Grid computing was originally driven by scientific applications which are usually computation-intensive. Cloud computing is similar to Grid computing in that it also employs distributed resources to achieve application-level objectives. However, cloud computing takes one step further by leveraging virtualization technologies at multiple levels (hardware and application platform) to realize resource sharing and dynamic resource provisioning.

Utility Computing: Utility computing represents the model of providing resources on-demand and charging customers based on usage rather than a flat rate. Cloud computing can be perceived as a realization of utility computing. It adopts a utility-based pricing scheme entirely for economic reasons. With on-demand resource provisioning and utility-based pricing, service providers can truly maximize resource utilization and minimize their operating costs.

Virtualization: Virtualization is a technology that abstracts away the details of physical hardware and provides



References (57)

1. Al-Fares M et al (2008) A scalable, commodity data center network architecture. In: Proc SIGCOMM
2. Amazon Elastic Computing Cloud, aws.amazon.com/ec2
3. Amazon Web Services, aws.amazon.com
4. Ananthanarayanan R, Gupta K et al (2009) Cloud analytics: do we really need to reinvent the storage stack? In: Proc of HotCloud
5. Armbrust M et al (2009) Above the clouds: a Berkeley view of cloud computing. UC Berkeley Technical Report
6. Berners-Lee T, Fielding R, Masinter L (2005) RFC 3986: uniform resource identifier (URI): generic syntax, January 2005
7. Bodik P et al (2009) Statistical machine learning makes automatic control practical for Internet datacenters. In: Proc HotCloud
8. Brooks D et al (2000) Power-aware microarchitecture: design and modeling challenges for the next-generation microprocessors, IEEE Micro
9. Chandra A et al (2009) Nebulas: using distributed voluntary resources to build clouds. In: Proc of HotCloud
10. Chang F, Dean J et al (2006) Bigtable: a distributed storage system for structured data. In: Proc of OSDI
11. Chekuri C, Khanna S (2004) On multi-dimensional packing problems. SIAM J Comput 33(4):837–851 CrossRef
12. Church K et al (2008) On delivering embarrassingly distributed cloud services. In: Proc of HotNets
13. Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A (2005) Live migration of virtual machines. In: Proc of NSDI
14. Cloud Computing on Wikipedia, en.wikipedia.org/wiki/Cloudcomputing, 20 Dec 2009
15. Cloud Hosting, CLoud Computing and Hybrid Infrastructure from GoGrid, http://www.gogrid.com
16. Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters. In: Proc of OSDI
17. Dedicated Server, Managed Hosting, Web Hosting by Rackspace Hosting, http://www.rackspace.com
18. FlexiScale Cloud Comp and Hosting, www.flexiscale.com
19. Ghemawat S, Gobioff H, Leung S-T (2003) The Google file system. In: Proc of SOSP, October 2003
20. Google App Engine, URL http://code.google.com/appengine
21. Greenberg A, Jain N et al (2009) VL2: a scalable and flexible data center network. In: Proc SIGCOMM
22. Guo C et al (2008) DCell: a scalable and fault-tolerant network structure for data centers. In: Proc SIGCOMM
23. Guo C, Lu G, Li D et al (2009) BCube: a high performance, server-centric network architecture for modular data centers. In: Proc SIGCOMM
24. Hadoop Distributed File System, hadoop.apache.org/hdfs
25. Hadoop MapReduce, hadoop.apache.org/mapreduce
26. Hamilton J (2009) Cooperative expendable micro-slice servers (CEMS): low cost, low power servers for Internet-scale services In: Proc of CIDR
27. IEEE P802.3az Energy Efficient Ethernet Task Force, www.ieee802.org/3/az
28. Kalyvianaki E et al (2009) Self-adaptive and self-configured CPU resource provisioning for virtualized servers using Kalman filters. In: Proc of international conference on autonomic computing
29. Kambatla K et al (2009) Towards optimizing Hadoop provisioning in the cloud. In: Proc of HotCloud
30. Kernal Based Virtual Machine, www.linux-kvm.org/page/MainPage
31. Krautheim FJ (2009) Private virtual infrastructure for cloud computing. In: Proc of HotCloud
32. Kumar S et al (2009) vManage: loosely coupled platform and virtualization management in data centers. In: Proc of international conference on

cloud computing

33. Li B et al (2009) EnaCloud: an energy-saving application live placement approach for cloud computing environments. In: Proc of international conf on cloud computing
34. Meng X et al (2010) Improving the scalability of data center networks with traffic-aware virtual machine placement. In: Proc INFOCOM
35. Mysore R et al (2009) PortLand: a scalable fault-tolerant layer 2 data center network fabric. In: Proc SIGCOMM
36. NIST Definition of Cloud Computing v15, csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc
37. Osman S, Subhraveti D et al (2002) The design and implementation of zap: a system for migrating computing environments. In: Proc of OSDI
38. Padala P, Hou K-Y et al (2009) Automated control of multiple virtualized resources. In: Proc of EuroSys
39. Parkhill D (1966) The challenge of the computer utility. Addison-Wesley, Reading
40. Patil S et al (2009) In search of an API for scalable file systems: under the table or above it? HotCloud
41. Salesforce CRM, http://www.salesforce.com/platform
42. Sandholm T, Lai K (2009) MapReduce optimization using regulated dynamic prioritization. In: Proc of SIGMETRICS/Performance
43. Santos N, Gummadi K, Rodrigues R (2009) Towards trusted cloud computing. In: Proc of HotCloud
44. SAP Business ByDesign, www.sap.com/sme/solutions/businessmanagement/businessbydesign/index.epx
45. Sonnek J et al (2009) Virtual putty: reshaping the physical footprint of virtual machines. In: Proc of HotCloud
46. Srikantaiah S et al (2008) Energy aware consolidation for cloud computing. In: Proc of HotPower
47. Urgaonkar B et al (2005) Dynamic provisioning of multi-tier Internet applications. In: Proc of ICAC
48. Valancius V, Laoutaris N et al (2009) Greening the Internet with nano data centers. In: Proc of CoNext
49. Vaquero L, Rodero-Merino L, Caceres J, Lindner M (2009) A break in the clouds: towards a cloud definition. ACM SIGCOMM computer communications review
50. Vasic N et al (2009) Making cluster applications energy-aware. In: Proc of automated ctrl for datacenters and clouds
51. Virtualization Resource Chargeback, www.vkernel.com/products/EnterpriseChargebackVirtualAppliance
52. VMWare ESX Server, www.vmware.com/products/esx
53. Windows Azure, www.microsoft.com/azure
54. Wood T et al (2007) Black-box and gray-box strategies for virtual machine migration. In: Proc of NSDI
55. XenSource Inc, Xen, www.xensource.com
56. Zaharia M et al (2009) Improving MapReduce performance in heterogeneous environments. In: Proc of HotCloud
57. Zhang Q et al (2007) A regression-based analytic model for dynamic resource provisioning of multi-tier applications. In: Proc ICAC

About this Article

Title
Cloud computing: state-of-the-art and research challenges
Journal
Journal of Internet Services and Applications
Volume 1, Issue 1 , pp 7-18

Cover Date
2010-05-01
DOI
10.1007/s13174-010-0007-6
Print ISSN
1867-4828

Online ISSN
1869-0238
Publisher
Springer-Verlag
Additional Links

- Register for Journal Updates
- Editorial Board
- About This Journal
- Manuscript Submission

Topics

- Processor Architectures
- Computer Applications
- Business Information Systems
- Information Systems and Communication Service
- Computer Communication Networks
- Computer Systems Organization and Communication Networks

Keywords

- Cloud computing
- Data centers
- Virtualization

Industry Sectors

- Electronics
- IT & Software
- Telecommunications

Authors

- Qi Zhang [1]
- Lu Cheng [1]
- Raouf Boutaba [1]

Author Affiliations

- 1. University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

HERON

Health Inequalites Research Network Conference 2014

14-15 May 2014
Southwark
London

Keynote Speaker
Dr David Williams

Click here