

8-2014

A Process Framework for Managing Quality of Service in Private Cloud

Arvind Maskara
Walden University

Follow this and additional works at: <https://scholarworks.waldenu.edu/dissertations>

 Part of the [Business Administration, Management, and Operations Commons](#), [Computer Engineering Commons](#), and the [Technology and Innovation Commons](#)

This Dissertation is brought to you for free and open access by the Walden Dissertations and Doctoral Studies Collection at ScholarWorks. It has been accepted for inclusion in Walden Dissertations and Doctoral Studies by an authorized administrator of ScholarWorks. For more information, please contact ScholarWorks@waldenu.edu.

Walden University

College of Management and Technology

This is to certify that the doctoral study by

Arvind Maskara

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee

Dr. Ronald McFarland, Committee Chairperson, Doctor of Business Administration
Faculty

Dr. Maurice Dawson, Committee Member, Doctor of Business Administration Faculty

Dr. Ify Diala, University Reviewer, Doctor of Business Administration Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2014

Abstract

A Process Framework for Managing Quality of Service in Private Cloud

by

Arvind Maskara

MISM, Walden University, 2011

MBA, Magadh University, 1994

B.Eng., Birla Institute of Technology, 1988

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Business Administration

Walden University

August 2014

Abstract

As information systems leaders tap into the global market of cloud computing-based services, they struggle to maintain consistent application performance due to lack of a process framework for managing quality of service (QoS) in the cloud. Guided by the disruptive innovation theory, the purpose of this case study was to identify a process framework for meeting the QoS requirements of private cloud service users. Private cloud implementation was explored by selecting an organization in California through purposeful sampling. Information was gathered by interviewing 23 information technology (IT) professionals, a mix of frontline engineers, managers, and leaders involved in the implementation of private cloud. Another source of data was documents such as standard operating procedures, policies, and guidelines related to private cloud implementation. Interview transcripts and documents were coded and sequentially analyzed. Three prominent themes emerged from the analysis of data: (a) end user expectations, (b) application architecture, and (c) trending analysis. The findings of this study may help IT leaders in effectively managing QoS in cloud infrastructure and deliver reliable application performance that may help in increasing customer population and profitability of organizations. This study may contribute to positive social change as information systems managers and workers can learn and apply the process framework for delivering stable and reliable cloud-hosted computer applications.

A Process Framework for Managing Quality of Service in Private Cloud

by

Arvind Maskara

MISM, Walden University, 2011

MBA, Magadh University, 1994

B.Eng., Birla Institute of Technology, 1988

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Business Administration

Walden University

August 2014

Acknowledgements

I would like to acknowledge my sincere thanks to Dr. Ronald McFarland for his mentoring, guidance, and encouragement throughout this process. I would like to also thank my committee members Dr. Maurice Dawson, DBA methodologist Dr. Reginald Taylor, and URR Dr. Ify Diala for their valuable contribution to my study. My special thanks to Dr. Freda Turner for her guidance and encouragement.

Table of Contents

Section 1: Foundation of the Study	1
Background of the Problem.....	1
Problem Statement.....	2
Purpose Statement	3
Nature of the Study.....	4
Research Questions	5
Interview Questions.....	7
Conceptual Framework	8
Definition of Terms	9
Assumptions, Limitations, and Delimitations	12
Assumptions	12
Limitations.....	13
Delimitations	13
Significance of the Study.....	14
A Review of the Professional and Academic Literature	15
Cloud Computing	17
Cloud Computing Architecture	17
Deployment Models	24
Virtualization.....	27
Agility and Flexibility in IT Infrastructure.....	29
Effect of IT Agility and Flexibility on Business Performance.....	30

Cloud Computing Provides Agility and Flexibility.....	31
Benefits of Cloud Computing.....	33
Privacy and Security.....	35
Requirement Management.....	38
Information Technology Processes.....	39
Cloud Computing Decision Frameworks.....	39
Quality of Service (QoS).....	43
Cloud Computing Adoption Challenges.....	46
Cloud computing and environmental sustainability.....	48
Case Studies on Cloud Computing Based Solutions.....	49
Cloud Computing Economies.....	50
Cloud Computing Technologies.....	51
Transition and Summary.....	53
Section 2: The Project.....	54
Purpose Statement.....	54
Role of the Researcher.....	55
Participants.....	56
Research Method.....	59
Research Design.....	60
Population and Sampling.....	62
Ethical Research.....	65
Data Collection.....	66

Instruments	66
Data Collection Technique	69
Data Organization Techniques	69
Data Analysis.....	71
Reliability and Validity	73
Transition and Summary	74
Section 3: Application to Professional Practice and Implications for Change.....	75
Overview of Study.....	75
Presentation of the Findings	76
Theme 1: End User Expectations	77
Theme 2: Service Level Agreements	78
Theme 3: Architecture	79
Theme4: Sizing.....	82
Theme 5: Monitoring.....	84
Theme 6: Trending	85
Research Question 1	86
Research Question 2	89
Research Question 3	89
Research Question 4	89
Application to Professional Practice	91
Implication for Social Change.....	92
Recommendations for Action.....	93

Recommendations for Further Study.....	94
Reflections	94
Summary and Study Conclusions.....	95
References	98
Appendix A: Participant Selection	116
Appendix B: Interview Questions	118
Appendix C: Confidentiality/Consent to Participants Form	120
Appendix D: Permission from Participating Organization	122
Appendix E: NIH Training Certificate	123
Curriculum Vitae	124

Section 1: Foundation of the Study

Cloud computing emerged as a new paradigm with the convergence of popular computing trends such as utility computing, grid computing, *pay-as-you-go* pricing model for service delivery over the Internet, virtualization, elasticity, and web 2.0 (Madhavaiah, Bashir, & Shafi, 2012; Pallis, 2010). Cloud computing has many different meanings; for ordinary consumers, cloud is a ubiquitous service accessible through a network-connected computing device (Low, Chen, & Wu, 2011). This device could be a mobile device such as a smartphone, a tablet, or a personal computer. For example, Apple iCloud allows users to store their data (address book, pictures, notes, music, and documents) in a computing facility managed by Apple Inc. (Fernando, Loke, & Rahayu, 2013), which they can access from any Internet-connected device anywhere in the world. Other examples of cloud-based services include Google Docs, Google Gmail, Dropbox, Box.net, and many more (Fernando et al., 2013; Hofmann & Woods, 2010). The meaning of cloud computing may differ depending on the context in which it is used (Armbrust et al., 2010). For the purpose of this study, cloud computing is the underlying information technology (IT) infrastructure that includes hardware installed in datacenter and software used for managing hardware and virtualization (Armbrust et al., 2010).

Background of the Problem

IT professionals look at cloud computing from a different perspective. For IT professionals, cloud is a computer infrastructure available for use without any upfront cost or long-term commitment (Armbrust et al., 2010) that can meet their varying resource (computing, storage, and bandwidth) needs and has a cost of service

proportional to usage. Thus, cloud computing provides opportunities for both service providers and service users.

Cloud computing service providers expect to make a profit through economies of scale. Cloud service providers use virtualization to rent hardware resource units to multiple users. Another aspect of cloud computing is the self-service model. Service users can order cloud service by filling an online form and providing a method of payment such as a credit card (Armbrust et al., 2010; Skiba, 2011). However, adoption of cloud computing requires overcoming few obstacles such as information security and quality of service (QoS).

Challenges associated with the adoption of cloud computing include concerns for information security and data privacy, vendor lock-ins, QoS, system availability, and support issues (Armbrust et al., 2010). There is also an increased risk of data breach in the public cloud due to multi-tenancy and Internet facing design (Takabi, Joshi, & Ahn, 2010). Therefore, large firms are building their own cloud infrastructure, private cloud (Marston, Li, Bandyopadhyay, Zhang, & Ghalsasi, 2011). These private clouds offer the same benefits as by external service providers, but allow firms to keep data and infrastructure under their control. The focus of this study is on private cloud.

Problem Statement

Cloud computing infrastructure offers on demand availability of computing resources, reduced upfront infrastructure cost, and a pay for use cost model without any long-term commitment (Armbrust et al., 2010). Total value of global cloud computing-based services may reach 150 billion by the year 2014 (Budrienė & Zalieckaitė, 2012).

About 50% of information system executives expect that more than half of their organizations transactions will take place in the cloud by the year 2016 (Budrienė & Zalieckaitė, 2012). Information security in the cloud is the biggest concern for 75% of the IT executives (Sultan, 2013). QoS, a factor behind performance and availability of applications hosted in the cloud, is the second biggest concern for 63% of the IT executives (Sultan, 2013). The general problem for business leaders is the lack of an agile and flexible in-house computing infrastructure (Bhatt, Emdad, Roberts, & Grover, 2010; Lu & Ramamurthy, 2011). The specific problem for businesses is the lack of well-defined processes for managing QoS in private cloud (Ferrer et al., 2012; Sunyaev & Schneider, 2013).

Purpose Statement

The purpose of this qualitative single-case study was to explore how information systems leaders manage QoS in successful private cloud implementations. The target population consisted of companies located in the state of California that successfully implemented a private cloud within last 2 years. This population was appropriate for the study because QoS is the second most important concern for 63% of the IT executives in adoption of cloud computing (Sultan, 2013). This study may contribute to positive social change by providing guidelines to IT leaders for maintaining QoS in cloud. By managing QoS in cloud, firms could ensure consistent and reliable application performance, which in turn would help businesses to meet their goals and achieve a competitive advantage.

Nature of the Study

The focus of this qualitative single-case study was on understanding processes related to management of QoS requirements in private cloud. Qualitative research is appropriate for exploring new research areas where there are unknown variables (Horsewood, 2011). Quantitative researchers use scientific methods for validating results of past studies (Horsewood, 2011). Quantitative methods were not suitable for this study due to a lack of past research on this topic. In qualitative research, the researcher interviews participants with a set of open-ended questions (Arghode, 2012). For complex research studies, the researcher may use mixed method, which is a combination of qualitative and quantitative methods (Yin, 2009). Mixed method research is time-consuming because it requires analysis of both qualitative and quantitative data. A mixed method was not required because sufficient information could be collected using qualitative methods.

The goal of this study was to explore ways to develop a process framework for managing QoS in the private cloud. I explored technology and processes related to IT infrastructure management at the selected organization. This study was different from traditional social science and natural science studies. I selected information rich participants and extracted information from them by asking open-ended questions; this methodology aligned to qualitative research design. The qualitative research method was the right fit for this study.

The single-case study design was most appropriate for this study. I evaluated phenomenological, grounded theory, and ethnography as possible designs for this study.

Phenomenological studies are about capturing and analyzing feelings and lived experiences of participants about a phenomenon (Norlyk & Harder, 2010).

Phenomenological design was not suitable for the goals of this study. Goals of studies based on grounded theory are to develop new theories on the basis of data collected during research (Dawes & Larson, 2011). This study was not about developing a theory from the data collected; therefore, grounded theory design was not be used. Ethnography is the study of specific cultural groups (Gagnon, Carnevale, Mehta, Rousseau, & Stewart, 2013). The purpose of this study was to explore implementation of private cloud in an organization; not the study of a specific cultural group. Therefore, ethnography did not align with goals of this study.

Research Questions

The primary objective of this study was to suggest a framework to information systems leaders and managers that may allow them to maintain QoS in private cloud. The overarching research question was as follows: How can organizations successfully manage QoS in a private cloud platform? During this study, to the best of my knowledge, there was no published process framework for ensuring QoS in the cloud. There was limited literature in the field of QoS in private cloud (Lawrance & Silas, 2013). The overarching research question was expanded to the following four specific research questions:

1. What processes should be implemented in private cloud environment for guaranteeing QoS (Liu et al., 2012)?

2. How can information systems managers identify QoS requirements of private cloud service users (Sunyaev & Schneider, 2013)?
3. How can information systems managers proactively identify QoS issues, which may cause application performance degradations (Faniyi, Bahsoon, & Theodoropoulos, 2012)?
4. What strategy should be followed for prioritizing allocation of computing resources in private cloud (Chung et al., 2013)?

In the first research question, I sought to identify necessary IT processes for ensuring QoS. These processes were used for building the process framework sought in the main research question. In the second research question, I focused on requirement gathering from users and an appropriate method for documenting such QoS requirements. These methods are necessary components of the solution of the main research problem. The answer of the third question would help information systems managers in developing processes for proactive detection of potential QoS issues before application performance is affected. The fourth question is about strategy for prioritizing cloud resources. The solutions of the third and the fourth questions contributed to the solution of the main research problem. Answers of all the above research questions contributed to the main research question.

Participants' responses to the interview questions helped in answering above research questions. After analysis of data collected from respondents, I was able to formulate answers of these research questions. This information was analyzed and

aggregated for presenting a process framework for managing QoS in private cloud, the answer to the main research question.

Interview Questions

The interview questions were formulated for collection of data for the study. The target of each interview question was to provide information leading to the answer of one or more of the research questions. The analyses of answers lead to the answer of the overarching research question. Some of the questions were followed by a probing question that helped in extracting additional information from participants. Interview questions are listed below:

1. What is your experience in application development, application support, infrastructure design, or infrastructure support? Please explain.
2. What processes you followed for ensuring consistent application performance in the private cloud?
3. How do you determine performance requirements for applications and ensure that application performance requirements are consistently met?
4. How do you determine hardware resource requirements for applications?
5. What challenges did you face in migrating applications to private cloud?
6. How do your determine performance, availability, and security SLAs for applications? How do you document these SLAs?
7. What proactive actions you take for ensuring that applications hosted in private cloud consistently meet established performance SLAs?

8. What information do you periodically receive and share with different IT groups in your organization?
9. How do you ensure that applications hosted in private cloud continue to get adequate system resources?
10. What additional information would you like to share that you may not have had the opportunity to address?

Conceptual Framework

This study was based on the theory of disruptive innovation. Christensen (2011) defined disruptive innovation, a radical innovation which changes the way firms carry out value creation activities. Disruptive innovation requires dismantling existing units and creating new units that perform different value creation tasks (Christensen, 2011). Customers, who could not afford certain goods and services in past, or were not possessing necessary skills to use such goods or services could access redesigned products or services. Introduction of such redesigned products and services is called disruptive innovation (Christensen, 2011).

The concept of disruptive innovation is about developing and offering goods or services which may have a lower performance compared to established goods and service (Ahlstrom, 2010). Consumers adopt these goods and service due to other feature such as convenience, portability, lower price, or ease of use. Disruptive innovation requires changing a firm's core logic. The shifting of core logic may lead to resistance from stakeholders due to structural inertia, competitive inertia, organizational moment, and the current management (Lengnick-Hall & Wolff, 1999).

Managers find it difficult to replace old trusted technologies and business models with unproven ones (Crockett, McGee, & Payne, 2013). Managers see a risk of hurting their existing business and profitability while adopting disruptive innovations.

Christensen (2011) argued that new entrants do better than established firms in adopting disruptive innovation. Unlike established firms, new entrants do not have an existing infrastructure and business model. They build new infrastructure specifically for the disruptive innovation and do not have to go through changes similar to the established players. Therefore, new entrants have higher success rate in adopting disruptive innovation.

Cloud computing is a disruptive innovation for enterprise (Aljabre, 2012; Dikaiakos, Katsaros, Mehra, Pallis, & Vakali, 2009). Cloud computing makes IT infrastructure available without the upfront cost and delays, which was not available in traditional IT infrastructure (Rogers & Cliff, 2012). Cloud computing also presents challenges because enterprise processes may not remain effective and new processes are to be implemented. IT organizations may also require reorganization with adoption of cloud computing.

Definition of Terms

Cloud computing is another paradigm in the area of IT. It uses most of the terminologies that were used in IT. There are some new terms such as cloud computing itself. Some of the previously defined terms may have a slightly different meaning now.

Application: Application software or computer application is a set of computer programs that provide users the ability to perform certain tasks on a computing device (Ciznicki, Kierzyńska, Kopta, Kurowski, & Gepner, 2012).

Cloud computing: Cloud computing is a new computing paradigm in which users can access computing resources using a network or Internet-connected device, such as a personal computer, a tablet, or a smart phone (Fernando et al., 2013; Garrison, Kim, & Wakefield, 2012). Cloud computing represents the underlying infrastructure such as data-centers, computing hardware, and software used for managing virtualization and hardware resources (Armbrust et al., 2010).

Cloud service provider: Firms that own and manage computer hardware, software, network, and storage and provide cloud computing services to consumers (Garrison et al., 2012).

Cloud service user: Cloud service users are the direct consumers of cloud services (Shin, 2013).

Compute: Computer hardware resources used for execution of computer programs (Ciznicki et al., 2012).

Data-center: A secured facility where computer servers, storage, network equipment, and other related equipment are installed. These resources are accessed through a local area network, wide area network, or Internet.

Infrastructure: All equipment and software used for hosting computer applications (Henfridsson & Bygstad, 2013).

Infrastructure as a service (IaaS): A cloud computing service model, in which cloud service users are given access and full control of virtualized computers (Garrison et al., 2012).

Performance: The performance of computer applications is the time it takes to respond to user requests (Ciznicki et al., 2012). The performance requirements of applications are documented in the service level agreement (SLA). Application performance requirements vary, depending on application usage. For the purpose of this study, the performance of an application was considered slow when its response time exceeded the time documented in SLA.

Platform as a service (PaaS): A cloud computing service model, in which cloud service users are given access to set of web and application servers, software libraries, and storage (Garrison et al., 2012). Cloud service users can deploy their computer programs on the platform for running their application.

Quality of service (QoS): QoS is the method of prioritizing compute and network resource and guaranteeing adequate resources to real-time and critical application for the purpose of maintaining desired performance level (Tolosana-Calasanz, Bañares, Pham, & Rana, 2012).

Software as a service: A cloud computing service model, in which a computer application is made available to users through Internet (Garrison et al., 2012). All processing and data storage are done in cloud.

Service level agreement (SLA): A formal agreement between service provider and service users on all aspects of service, such as availability, performance, and service request response time (C. Huang et al., 2013).

Storage: Computer equipment used for storing and retrieval of large volume of data in digital format (Liu & Dong, 2012).

Utility computing: An IT infrastructure service model in which computing services are offered like other utilities such as electricity, telephony, and water (Low et al., 2011). Consumers just need a network connection and utility computing providers are responsible for managing hardware, software, storage, and network.

Virtualization: A technique in which a single computer (host) can behave like multiple virtual computers. Each virtual computer (also called guest) could run its own operating system, has central processing unit (CPU), memory, storage, and network interface that are drawn from the resource pool on the host computer (Pal & Pattnaik, 2013). Each guest computer is logically isolated from other guests running on the same host.

Assumptions, Limitations, and Delimitations

Assumptions

There were two assumptions associated with this study. The first assumption was that the participants of the study would answer all interview questions with sincerity and without any bias or prejudice. The second assumption was that the themes and patterns, identified in data gathered from the interview response of participants were true

representations of participants' experience. The participants of study were IT professionals responsible for managing performance of computer applications.

Limitations

The results of this study were limited by three factors. First, the accuracy of results was dependent on the credibility and reliability of the information provided by participant professionals. Second, because cloud computing is still evolving, the participants of study did not have long-term experience with cloud computing; therefore, some of the information gathered from the participants was based on their experience with traditional IT infrastructure instead of direct experience with cloud computing. Third, because this was a single-case study, private cloud implementation was studied on only one organization. The results of this study may have been influenced by the organization culture of the organization under study.

Delimitations

This study was delimited to (a) the problem to be studied, (b) the organization selected for the single-case study (c) selection of participants, (d) the population, and (e) the size of population. The two key delimiters of this study were the problem being studied and organization selected for case study. This study was delimited to analysis and interpretation of data collected from the selected organization. I collected data by reviewing documents such as SOPs, policies, and guidelines related to cloud computing implementation and by interviewing 23 IT professional selected using a purposeful sampling method.

Significance of the Study

The results of this study will provide guidelines for setting up processes that may help in managing of QoS in private cloud. The process framework identified in this study could be used by IT organizations for determining the level of private cloud resources to be assigned to various applications that would ensure desired application performance under varying load. In this study, I focused on the business problem of adopting cloud computing. I worked on finding a solution to the problem of QoS management in the private cloud. There are two aspects of QoS management in private cloud the technical aspect and the process aspect. On the technical side, researchers have addressed issues such as dynamic resource allocation under varying load on computational, network, and storage resources (Lawrance & Silas, 2013). On the process side, researchers have focused on SLAs guaranteeing cloud resources such as CPU speed, I/O bandwidth, and memory size (Lawrance & Silas, 2013). In this study, I focused on the process aspect. At the time of writing this report, I could not find any literature about a process framework for managing QoS in private cloud.

The management of QoS in private cloud is important due to a smaller resource pool compared to public clouds (Quarati, Clematis, Galizia, & D'Agostino, 2013). Applications running in cloud infrastructure share infrastructure components, such as compute, storage, operating system, and network. The resource consumption by applications may increase due to change in application dynamics, such as more users on a web application or processing of a data intensive report that requires complex calculations (Yang, Yu, Jian, Qiu, & Li, 2011). Cloud infrastructure is elastic, but no

cloud infrastructure has infinite resources; resources may exhaust at some point when the load keeps on increasing. In such a situation, high usage of one application may affect performance of other applications. Such performance degradation may be acceptable for some applications; for example, a backup job, but may not be acceptable for other applications, such as an online shopping application or a video streaming service. QoS management could ensure that applications or services receive adequate resource for meeting their quality requirements.

There was limited information available on this topic in literature. Most process frameworks in studies were focused on migration of an application from the noncloud environment to cloud environment and did not address the issue of QoS. There were several studies on managing QoS in the cloud, but none of these scholars addressed processes that could be used for determining the QoS requirements of application or service when deploying in a private cloud. In this study, I addressed this gap in the literature.

A Review of the Professional and Academic Literature

The literature for this study was gathered by searching various research databases through Google Scholar and research databases at the Walden University library. Research databases at the Walden University library included Business Source Complete/Premier, Computer and Applied Science Complete, IEEE Xplore Digital Library, Science Direct, ACM digital library, and ProQuest Central. The Gartner research database was searched for professional literature. The keywords used in the search were *cloud computing*, *cloud security*, *cloud research*, *cloud frameworks*,

disruptive innovation, cloud computing business model, cloud computing infrastructure, cloud database, quality of service, private clouds, cloud process framework, and a combination of these keywords. I searched peer-reviewed articles published in the year 2010 or newer. An open source software Qiqqa was used for cataloging and managing articles in electronic format.

The literature review is divided into a number of sections for researching the purpose of study and investigating the research question. The purpose of this study was to identify a process framework for managing QoS in private cloud by exploring processes followed in the successful private cloud implementation in an organization. The overarching research question was the following: How organizations can successfully manage QoS in private cloud?

The literature review on cloud computing QoS covers broad categories such as (a) cloud computing, (b) cloud computing architecture, (c) deployment models, (d) virtualization, (e) agility and flexibility in IT infrastructure, (f) effect of IT agility and flexibility on business performance, (g) cloud computing provides agility and flexibility, (h) benefits of cloud computing, (i) privacy and security, (j) requirement management, (k) IT processes, (l) cloud computing decision frameworks, (m) QoS, (n) cloud computing adoption challenges, (o) cloud computing and environmental sustainability, (p) case studies on cloud computing-based solutions, (q) cloud computing economics, and (r) cloud computing technologies.

Cloud Computing

Traditionally, computer applications required building infrastructure by buying hardware, software, and network connectivity. This requires an upfront cost and delays when building and deploying new computer applications. In case the computer application cannot be fully used by users, the investment in the infrastructure would go to waste. Similarly, if there has been growth in an application usage, which requires more compute power, adding new infrastructure is time consuming and expensive. Also, the infrastructure has to be sized for the peak usage of applications; therefore, most of the times server hardware at data-centers remained underused. Most servers in corporate data-centers run at 10-30% capacity use (Marston et al., 2011). Cloud computing is the solution to these problems.

There are three aspects of cloud computing that change the completed landscape of IT infrastructure (Armbrust et al., 2010). These three aspects are (a) a perception of unlimited computing power available on demand to meet load fluctuations, (b) availability of computing infrastructure without any upfront cost, and (c) pay-per-use services. Companies such as software startups can now start their business without much investment in server infrastructure and will be able to scale up with growth in usage without any delay.

Cloud Computing Architecture

The word cloud in cloud computing came from network diagrams. The cloud symbolizes a part of the network for which interfaces are available, but internal architecture such as routing and interconnectivity between various components is not

known. Internet, in network diagrams, is also represented by a cloud. Cloud provides an abstraction to the complexities of the underlying infrastructure components. As defined by the National Institute of Standards and Technology (NIST), cloud computing is a pool of computing resources such as servers, storage, networks, applications, and services (NIST, 2012). These resources are available on demand with little or no interaction with cloud service provider. The definition of cloud computing by NIST includes five essential characteristics (a) on-demand self-service, (b) broad network access, (c) resource pooling, (d) rapid elasticity, and (e) measured service. The cloud computing architecture should cover all of the above five characteristics.

The cloud computing environment is made of three layers as shown in Figure 1. The lower-most layer contains the physical components, such as servers, storage, and network switches. The next layer is the virtualization layer, made of virtual servers, storage, and network; this hides complexities of the underlying infrastructure. Virtualization has been a common practice in IT infrastructure. The top layer, the cloud management layer, distinguishes cloud computing from traditional IT infrastructure (Grobauer, Walloschek, & Stocker, 2011; NIST, 2012). The cloud management layer provides a self-service portal for auto provisioning, a management interface for monitoring and capacity planning, and the resource metering and billing capabilities. Public cloud service providers may charge their customers based on their resource usage; private cloud service provider may internally charge-back their clients (Low et al., 2011) or just use other features of cloud computing depending on organizational policies.

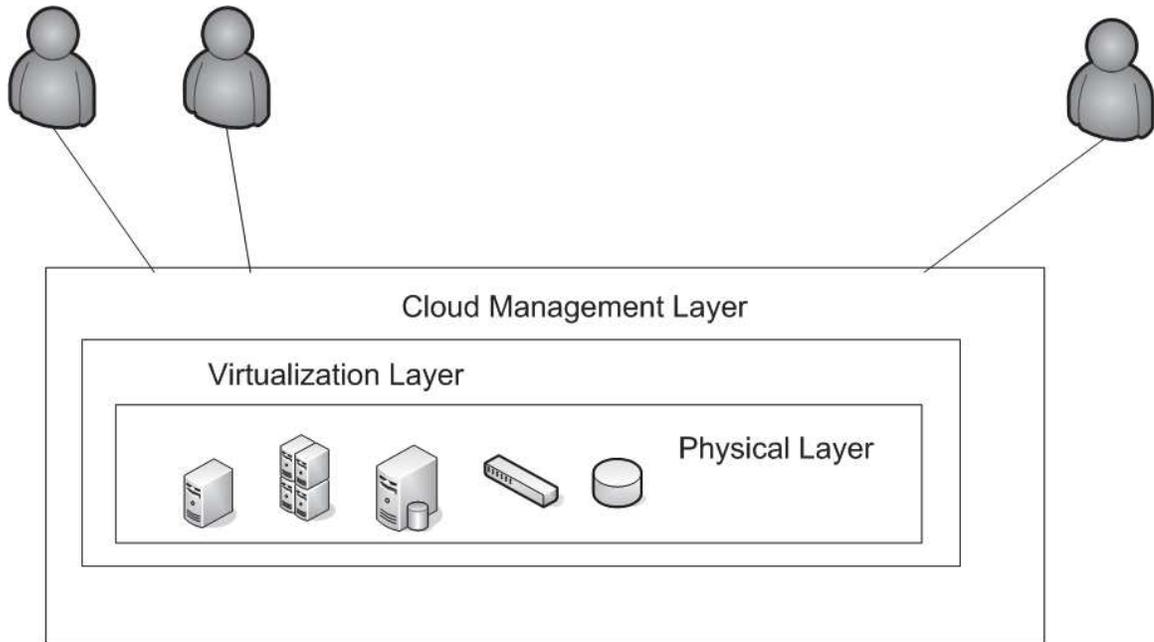


Figure 1. The general architecture of cloud computing.

Cloud computing implementations vary among different service providers. Three distinct types of architecture are the three service models IaaS, PaaS, and SaaS and four deployment models are public cloud, community cloud, private cloud, and hybrid cloud (NIST, 2012). There are scholarly articles about cloud computing architecture on security, QoS, power management, computational resource optimization, high performance computing, and mobile computing. A review of such a broad range of architecture is beyond the scope of this study. For the purpose of this study, I will be discussing the common cloud architectures.

Infrastructure as a service architecture: IaaS architecture provides maximum control to the cloud users. Cloud users could request providers for specific operating systems and install software of their choice on the server assigned to them. IaaS servers in most cases are virtual servers. Some vendors also provide servers with dedicated

hardware at a higher cost. For example, Amazon EC2 gives an option of dedicated instance to their customer, which is hosted on a single unshared server (Amazon.com Inc., 2013). Consumers of IaaS are typically system administrators; IaaS consumers get access to network attached storage and network resources such as firewalls and virtual private networks (VPN; NIST, 2012). The total cost of IaaS service may include the cost of (a) server and other fixed infrastructure, (b) per GB of storage, (c) network bandwidth used, and (d) value added services such as monitoring. These services are typically charged on a per hour basis.

Most commercial cloud infrastructures are proprietary, and their architecture is publically not available (NIST, 2012). There is several open source cloud management software available, and the description of their architecture is available in the literature. Eucalyptus is an open source cloud management software developed by Wolski (Milojicic & Wolski, 2011). The architecture diagram of an IaaS infrastructure using Eucalyptus is presented in Figure 2. The three layers in the Eucalyptus architecture are cloud manager, cluster managers, and computer managers (NIST, 2012; Q. Huang et al., 2013). These layers manage different aspects of cloud computing infrastructure.

The cloud manager is at the top level of hierarchy and provides the public interface to the customers. The cloud manager provides secured authentication and authorization, accepts resource requests, and provisions resources. The cloud manager accesses data object repository (DOS), which contains cloud administration data and facilitates billing. The cluster manager is the middle layer; cluster managers queries computer manager and connects customers to the computer manager which has sufficient

resources for meeting customer requests. The cluster manager does the load balancing among multiple computer managers.

Computer managers are at the lowest level of hierarchy; each computer manager works with a hypervisor running on a physical server. The computer manager communicates with cluster manager about the status of virtual machines and available capacity on the server. The computer manager does the start, stop, suspend, and reconfiguration of virtual machine by interfacing with hypervisor. All of the computer managers and servers share a persistent storage and are connected with a high speed network. Virtual machine images are stored on the persistence storage and could be started on any of the hypervisors.

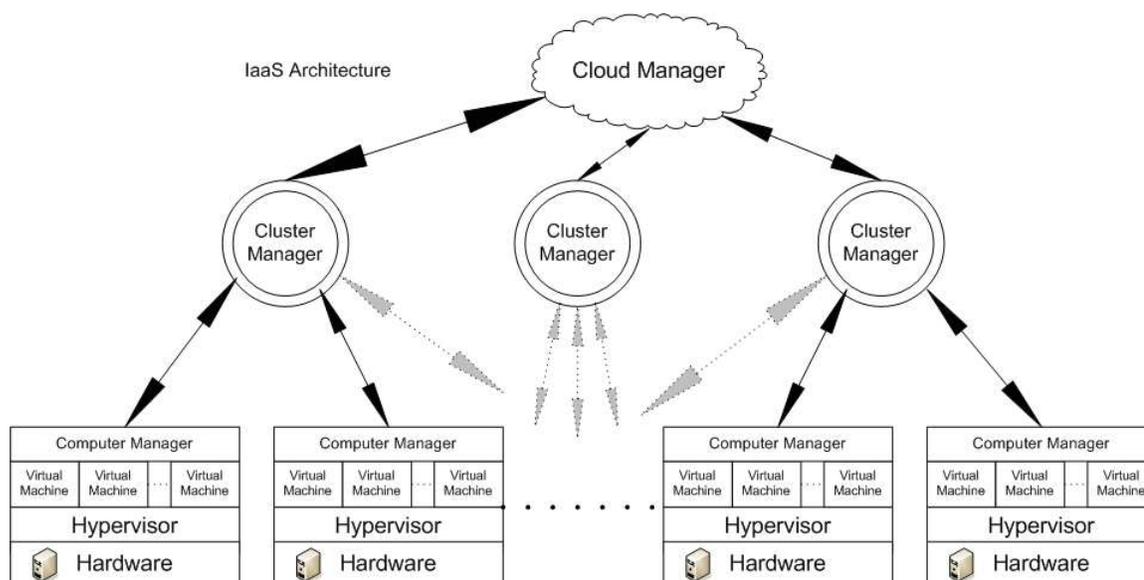


Figure 2. The IaaS architecture.

Platform as a service architecture: In the PaaS model, service providers maintain underlying IT infrastructure including processing, storage, operating systems, and underlying application hosting components such as database servers, web servers,

and application server (Goscinski & Brock, 2010). Consumers write their own code and deploy on the platform provided by the service providers. It gives developers ability to develop and deploy web applications to cloud infrastructure using browser based tools (Liu et al., 2012). Consumers have full control over their code but do not control underlying infrastructure including hardware, network, operating systems and platform components. Examples of PaaS are Google App Engine, and Microsoft Azure. This model provides less flexibility to cloud service users compared to IaaS.

The infrastructure of PaaS is generally built on the IaaS infrastructure. An application development and execution environment allows developers to develop and deploy code without knowing the complexities of the underlying infrastructure. Cloud computing is a generalized concept, and there is no single architecture for PaaS (Boniface et al., 2010). Boniface et al. presented a design for PaaS which allows managing QoS in PaaS for the benefit of SaaS service.

One of the challenges of cloud computing is the management of QoS commitments to the customers (Boniface et al., 2010). This challenge could be handled at the PaaS level by using some special tools and services. A PaaS architecture supporting applications with guaranteed QoS requires five key features (Boniface et al., 2010) (a) real time QoS specifications, (b) event prediction, (c) dynamic SLA negotiation, (d) on demand resource provisioning, and (e) QoS event monitoring. By managing QoS at the PaaS level, applications hosted at the SaaS layer could also take advantage of QoS management; because, this design makes QoS management features available to SaaS applications.

Software as a service architecture: In the SaaS model, consumer can access software applications using the client software on a network-connected computer, typically the web browser. In this service model, service providers are responsible for maintaining complete IT infrastructure and software. Common example of SaaS is web based email services, such as Gmail by Google and Hotmail by Microsoft. Another example of SaaS is Google Docs that provides online word processing, electronic spreadsheet, and presentation software (Skiba, 2011). Another example is Customer Relationship Management (CRM) application by Salesforce.com.

Software as a Service model is built over IaaS and PaaS infrastructure. One of the requirements of SaaS infrastructure is multi-tenancy (Mauch, Kunze, & Hillenbrand, 2013). Multi-tenancy facilitates efficient use of system resources (Espadas et al., 2013). A tenant in SaaS service is an application user; although there are multiple application users at a time, a multi-tenant aware application responds to user in such a way, as if it was the only user of the application. Multi-tenancy could be achieved at one of the four cloud infrastructure layers: the application layer, middleware layer, virtual machine layer, or the operating system layer. Also, there are two common multi-tenancy patterns: native multi-tenancy, which is achieved within a single instance of software and multiple instances, which is achieved by creating a separate instance for each user (Espadas et al., 2013). Native multi-tenancy is more scalable, it may support thousands of users. Multiple instances may support hundreds to few thousand users; multiple instances provide better isolation to users in comparison to native multi-tenancy.

A group of researchers at Massachusetts Institute of Technology (MIT) presented the design of a database cloud. They proposed a relational database management system (RDBMS) that could make the database available as a service (Curino et al., 2011). They have named this concept Database as a Service (DaaS). This is an extension of the concepts of IaaS, PaaS, and SaaS. This design overcomes three major challenges of database service in cloud (a) efficient multi-tenancy, (b) elastic scalability, and (c) database privacy.

Cloud service providers deliver services to clients through Internet or private network (private cloud). Timesharing and application hosting are not new concepts; timesharing was common in 1960s and 1970s, application hosting in 1980s and 1990s, and web hosting in 2000s (Cusumano, 2010). Cusumano argued that cloud is different because it has emerged as a platform. Cusumano recommended that the cloud vendors must open their platform to others who may complement their services, or in some cases compete with them. Cloud vendor can establish their SaaS, IaaS, or other services as an industry standard platform only after opening to other collaborators.

Deployment Models

There are four cloud deployment models defined by NIST. These models use similar technologies and architecture. The major difference among these deployment models is the user community. The user community of cloud could be members of an organization, general public, or special groups.

Private cloud. Large organizations build their own cloud infrastructure for meeting application hosting needs of their internal consumers in various business units

(Quarati et al., 2013). Private clouds are similar to public clouds in architecture, but the consumers of private cloud are from within the organization. Setup and operations of private cloud requires both the traditional IT skills and the cloud IT skills (NIST, 2012). Private clouds offer better security and data privacy compared to the public cloud, due to an additional layer of security at network.

Organizations have an obligation to meet the regulatory requirements that may include the Sarbanes-Oxley (SOX) act 2002, Health and Human Services Health Insurance Portability Act (HIPAA), and Payment Card Industry (PCI) data security standards. Also, organizations must safeguard their confidential information, such as employee personal information, customer information, and the financial information (Hofmann & Woods, 2010). Private clouds provide a more secured environment compared to public cloud, which helps organizations in meeting their security and compliance requirements (Hofmann & Woods, 2010).

Private cloud may be built in a single data-center or a group of geographically distributed data-centers. Secured high-speed network links interconnect private cloud infrastructure located in different locations (NIST, 2012). Some organizations outsource private cloud to external service providers that are hosted off-site at service providers premises. The outsourced private cloud requires a secured high-speed link between the customer location and the service provider.

Public cloud. Public cloud service providers offer cloud services to individuals and other organizations (Garrison et al., 2012). Public cloud service providers manage hardware in their own data-centers and offer services using various pricing models.

Public cloud services could be compared with the traditional utility services like electricity, water, and gas. In the public cloud, computing resources are made available to consumer on demand for a price.

One of the biggest advantages of public cloud is no upfront cost for cloud users. Therefore, this is an ideal solution for the startup companies, small and medium organizations, and individuals. Public cloud may offer better safety and security to a small business owners compared to the in-house infrastructure, because small business owners may not be able to hire expert IT staff (NIST, 2012). Public cloud service provider Microsoft offers Microsoft office live, an online office application for businesses at a small subscription fee (Gupta, Seetharaman, & Raj, 2013). Google offers similar subscription based service, Google docs for business. Another popular service is Cisco WebEx web conferencing that facilitates online meetings with computer screen sharing. Businesses have achieved significant cost reductions by utilizing such services (Gupta et al., 2013).

Community cloud. This type of cloud infrastructure is built for meeting the computing needs of a community of consumers with shared interests (Pallis, 2010). For example, a shared cloud infrastructure of community members of a group of educational institutions is a community cloud. Community cloud shares most of the characteristics of private cloud. The only difference between a community cloud and the private cloud is in the cloud users. Private cloud access is restricted to a single organization, its business partners, and guests; whereas community cloud is open to more than one organizations and individuals associated with those organizations.

Hybrid cloud. A hybrid cloud is a combination of two or more of the above deployment models (Garrison et al., 2012). For example, some organizations may have a hybrid of public and private cloud. Hybrid cloud infrastructure is built with composition of two or more distinct cloud infrastructures that remain separate entities, but work together by interfacing through standard or proprietary technologies.

Virtualization

Virtualization is the key underlying technology in the cloud computing implementations. It is present in the most cloud implementation and many people think it as a feature of cloud computing, not the implementation detail (Montero, Moreno-Vozmediano, & Llorente, 2011). Virtualization allows sharing of hardware resources by multiple virtual computers. Each virtual computer could run its own operating system and assigned a slice of hardware resources such as CPU, memory, storage, network, and input-output devices (Lin, Qi, Wu, Dong, & Guan, 2012). The virtual computer is also called virtual machine (VM). A VM is an abstraction layer between the hardware and user (Pal & Pattnaik, 2013); often VMs are also called virtual servers when they run server software such as a web server, application server, or a database server.

The real hardware on which multiple VMs run, uses a special software called hypervisor or virtual machine monitors (VMMs). Hypervisors allow the creation and execution of multiple VMs on the host hardware (Pal & Pattnaik, 2013). Hypervisors have special features that allow multiple virtual machines to exist in their own secured environment, as if all the machines are running on separate hardware. Hypervisors allow VMs to run different operating systems on the same host hardware. Modern CPUs and

server hardware are designed to support virtualization; hypervisors are capable of taking advantage of hardware virtualization features.

There are three types of virtual machine implementations: software virtual machines, bare metal virtual machines, and virtual OS/containers (Pal & Pattnaik, 2013). Software virtual machines work as a layer between the host operating system and the guest operating system. An example of software virtual machine is Microsoft Virtual Server 2005. Hypervisors, in a bare metal virtual machine, run directly on the hardware without any underlying operating system. Example of bare metal hypervisor is VMWare vSphere. Some operating systems have capabilities of partitioning in different containers of zones. Examples of OS/container virtualization are Solaris Zones and BSD Jail.

Virtual Machines are the building blocks of most cloud computing infrastructure. Management of QoS in the cloud computing requires proper resource allocation to virtual machines. A given hardware may support only a limited number of virtual machine. With an increase in computational load, network load, or disk input-output rate; there is a risk of missing agreed QoS requirements. For example, a web application with an agreed SLA of less than 5 second per transaction could be at the risk of taking more than 5 seconds, when computational load on the server is too high. One of the methods to manage QoS is migrating virtual machines to hardware with adequate available resources (Beloglazov, Abawajy, & Buyya, 2012). A cloud resource manager could effectively manage resources based on agreed QoS requirement (Beloglazov et al., 2012).

Agility and Flexibility in IT Infrastructure

The words agile and flexible are sometimes used interchangeably. While flexibility means the ability to effectively change, in response to the changing business needs; agility means the ability to quickly change, for effectively meeting business needs (Bernardes & Hanna, 2009). The difference between the two is their relative speeds (Gong & Janssen, 2012). The terms agile and flexible are used for both business services and IT services. For the purpose of this study, agility and flexibility relate to only IT services. For IT services, the meanings of flexibility and agility are different, with some overlap.

The IT infrastructure of an organization is considered flexible when it is scalable, compatible, modular, and capable of handling multiple business applications (Bhatt et al., 2010). Scalability allows an organization to effectively meet its growing business needs, such as increase in application users, data, or complexity in processing. An organization's IT infrastructure, which is compatible with standard technologies, could be used for effectively communicating with its business partners. Compatible infrastructure would help in adoption of new business solutions when becomes available, without building a separate infrastructure. Modularity in IT infrastructure gives flexibility to modify, upgrade, and reconfigure specific IT components without affecting other components. Organizations could make more efficient use of IT infrastructure when they could deploy multiple business applications together. Most computer applications do not fully utilize hardware resources; by sharing infrastructure, organizations could save in hardware, software licensing, power and space, and maintenance and upgrades.

An agile IT infrastructure could quickly scale up or scale down with variation in computational, storage, or network bandwidth requirements (Faniyi et al., 2012). The measurement of agility could be done in terms of time taken in allocating additional computational power, storage, or network bandwidth. For example, in an agile IT infrastructure, computational power could be added by instantiating more virtual machines with predefined CPU and memory, or by allocating more virtual CPUs and/or memory to the virtual machine (Espadas et al., 2013). At the time of writing this paper, I could not find any literature that defines a numerical measurement of agility such as the time taken to instantiate N number of virtual machines.

Another measurement of infrastructure agility is how quickly a new instance of a virtual server or an application can be invoked (Aljabre, 2012). Invocation of virtual server is applicable to the IaaS, and invocation of application is applicable to PaaS and SaaS. An agile infrastructure allows a quick deployment of applications software (Narasimhan & Nichols, 2011). This increases application availability during application upgrade cycles.

Effect of IT Agility and Flexibility on Business Performance

Information technology is an integral part of all the modern businesses; businesses depend on their IT for managing their day to day operations. Business environments are very dynamic; businesses are required to quickly respond to external changes and competition. The changes in business may also require enhancement of existing IT capabilities or developing new capabilities (Wang, Liang, Zhong, Xue, & Xiao, 2012). This requires an agile and flexible IT infrastructure.

Cloud Computing Provides Agility and Flexibility

Agility and flexibility are the main features of cloud computing. Cloud computing infrastructure is highly scalable due to its inherent elasticity. A cloud computing based infrastructure gives a notion of the availability of unlimited computational resources, storage, and network bandwidth that could be utilized on demand (Armbrust et al., 2010). Although, in reality, no computing infrastructure has unlimited resources; cloud computing can provide a sense of unlimited resources up to a certain level of resource requirements. Another quality of a flexible infrastructure is compatibility; users of IaaS service can deploy most operating systems and software in cloud infrastructure, this provides compatibility to most software applications.

Another requirement of flexible infrastructure is modularity; cloud users can add or remove IT components on need basis. New virtual machines or extra storage could be added for a short term task and removed after the task is completed. Cloud service providers also provide the flexibility of adding extra computational nodes to corporate IT infrastructure through VPN (Marston et al., 2011). This gives flexibility of quickly expanding capacity of existing data-centers without buying and installing new hardware.

Organizations could extend their existing data-centers by connecting securely to a private cloud hosted off-site by an external service provider (Bossche, Vanmechelen, & Broeckhove, 2013). External service providers build a virtual network at their sites that becomes a logical extension of client's data-center network. One of the challenges in extension of data-center is maintaining a balance between cost and QoS. Bossche et al. presented a model, hybrid cloud construction and management (HICCAM), for

optimizing cost. The HICCAM model may help data-center managers in managing QoS requirements of applications while optimizing cost by using the service of external provider with the least cost. This is a viable solution when increased capacity is needed for a short term, needed quickly, or corporate data-centers are limited by space or power. Cloud infrastructure is designed for being shared by multiple applications; therefore, it also meets the flexibility requirement of sharing infrastructure by a diverse application mix.

Agility is one of the prominent features of cloud computing. Cloud infrastructure could be designed for dynamically adjusting resource allocation under varying computational load, data storage requirements, and network bandwidth. Cloud computing infrastructure may utilize techniques such as dynamic load-balancing of virtual servers (Espadas et al., 2013). Agility in cloud computing infrastructure is the ability to dynamically allocate resources with consideration to application QoS requirements (Laili et al., 2013). Two critical steps towards implementation of an agile cloud computing infrastructure are (a) service composition optimal selection (SCOS) and (b) optimal allocation of computing resources (OACR) (Laili et al., 2013). Service composition algorithms select resources on the basis of availability and communication-route. Optimal allocation requires analysis of application QoS requirements; OACR algorithms select application execution environments that meet QoS SLAs. Monitoring systems detect potential QoS violations. In response to monitoring, cloud management systems take preventive action by assigning adequate resources through dynamic configuration or by utilizing load-balancing.

Dispatching and load-balancing are used for managing agility by distributing the load amongst different infrastructure components. Load-balancing systems could dispatch user requests to infrastructure with adequate available resources for meeting application QoS requirements (Mondal, Dasgupta, & Dutta, 2012). Load-balancing could be session-based or request-based depending on the application design. Geographical load-balancing may be used by dispatching user requests to servers, in the proximity of users. This would minimize latency and provide quicker response leading to a superior user experience (UX) (Mondal et al., 2012). The management of agility is of a greater importance in private cloud due to smaller resource pool size compared to public clouds (Laili et al., 2013).

Agility of IT infrastructure is also measured by how quickly a new instances or server or application is created. The self-service feature of cloud computing makes it extremely quick to create a server or application instance in both public and private cloud. In the public cloud, filling an online form, agreeing with the terms and condition of contract, and providing a method of payment could prepare new instances in minutes (Amazon.com Inc., 2013). In private cloud new instance could be created based on organizations' internal policy and technical infrastructure. In many cases, these instances could be made available in few minutes or hours.

Benefits of Cloud Computing

Narasimhan and Nichols (2011) conducted a surveyed of 150 IT decision makers in various organizations who adopted at least one cloud based application. The results of the survey have shown that these early cloud adopters were highly satisfied with cloud

hosted applications (Narasimhan & Nichols, 2011). They were no longer skeptical about cloud and planned to host more applications in cloud in the future. Participants acknowledged the role of IT has significantly changed with adoption of cloud computing. Cloud computing is beneficial for business due to many reasons such as lower total cost of ownership (TCO), time to value, availability, ease of deployment, ease of integration, customizability, user adoption, reliability, and security (Narasimhan & Nichols, 2011).

One of the advantages of cloud computing is it provides flexibility in application hosting (Narasimhan & Nichols, 2011). Bhatt, Emdad, Roberts, and Grover (2010) surveyed senior executives of 105 manufacturing and service firm. From the results of the survey, they established that flexibility of IT infrastructure was positively related to the performance of business and gives firms competitive advantage.

Business enterprises in the United States and Europe spent about 44% of their IT budget on IT infrastructure during 2011 (Luftman et al., 2012). Cost reduction is a major concern for IT leaders (Luftman et al., 2012). IT leaders can reduce costs by optimizing IT assets. Transforming IT infrastructure to private and public clouds is an effective strategy for the optimization of IT assets (Aljabre, 2012).

Dwivedi and Mustafee (2010) described six aspects of cloud computing, which are not discussed in many cloud related publications. These six aspects are (a) low-cost access and computing devices (LCADs), (b) parallel programming, (c) communication networks, (d) open source software, (e) cloud access to high performance computing, and (f) green computing. These technologies can be implemented in cloud at a cost, lower than the traditional infrastructure.

Grid computing infrastructure provides a platform for hosting compute resources intensive applications. Using cloud resources on demand can expand the capacity of the grid. With the option of using cloud resources, organizations may not build in-house grid infrastructure capacity matching their peak usage, which reduces infrastructure and operational cost. This also gives organizations flexibility of utilizing infinite computational resources offered by cloud computing on the pay-per-use basis (Vazquez, Huedo, Montero, & Llorente, 2011). Vázquez et al. presented a framework for monitoring and expanding capacity of grid infrastructure. When grid capacity approaches saturation, grid expands by utilizing the public cloud or a partner grid. The resource provisioning framework ensures adherence to the QoS SLAs of enterprise grid.

The computer science department of Hochschule Furtwangen University, in Furtwangen, Germany uses a private cloud for hosting e-learning system being used by students and staff (Doelitzscher, Sulistio, Reich, Kuijs, & Wolf, 2011). University IT and computer science department have collaboratively developed a private cloud. Other than the e-learning system, private cloud at the University is also being used for meeting the computing requirements of research and development. The private cloud at University has significantly reduced the hours spent by IT administrators. Students are able to spend more time on learning because they are not required to install and configure servers; the server and platform services are available to them as IaaS and PaaS.

Privacy and Security

Grobauer, Walloschek, and Stocker (2011) argued that cloud computing has more security vulnerabilities compared to traditional information systems. Cloud environment

is an attractive target for hackers because if they break into cloud, they would have access to much more information than traditional infrastructure. The damage done by a hacker in cloud based infrastructure would have wider implications due to multi-tenancy model used in cloud. Due to this, all the previously known security vulnerabilities of traditional information systems are amplified in cloud computing. Cloud infrastructure also has several cloud specific vulnerabilities which presents more security challenges compared to traditional infrastructure.

Cloud architecture is an improvised technology; the underlying infrastructure and technology remains the same (Subashini & Kavitha, 2011). All the security vulnerabilities of the underlying infrastructure components remain present in the cloud infrastructure. Subashini and Kavitha utilized secondary data sources for identifying types of threats as applicable to various service delivery models of cloud computing. They concluded firms should utilize private clouds for applications containing private and confidential information.

Identity access management is another challenge with cloud computing. Daniels (2011) presented a model *Assured Identity Management System (AIMS)*, for managing identity and access in a cloud-computing system. The goal of AIMS framework is to promote reusable components for furthering the adoption of cloud computing, as standards are created and implemented (Daniels, 2011). The AIMS framework provides a homogeneous security context to cloud based systems for managing identities.

Five common attack techniques that could be used against cloud are denial of service attack, cross virtual machine side-channel attack, malicious insiders attack,

attacks targeting shared memory, and Phishing attack (Khorshed, Ali, & Wasimi, 2012). Khorshed et al. collected data on different types of data security threats in cloud computing environment. They used different techniques to analyze data and developed a model for proactive attack detection. They used confusion matrix analysis for studying the performance of various techniques used in attacking a networked computer system. They developed support vector machine (SVM), a statistical learning theory based data mining technique for proactively detecting an attack on cloud computing infrastructure.

Ryan (2011) presented an analysis of conference management system, a web based tool which allows program committee members and conference chairs browse papers and contribute reviews via web. This conference management system is available both as a PC installable software and a cloud based applications. Ryan identified various privacy related issues in the cloud based systems, which were not present in the standalone system. Ryan argued that the issues present in the cloud based conference management are universal and present in all the cloud-based applications.

Information and data security is the biggest obstacle to the adoption of cloud computing by business organizations (Che, Duan, Zhang, & Fan, 2011). Che et al. surveyed existing security models such as (a) multiple-tenancy model, (b) risk accumulation model, and (c) cube model of computing. They recommended several security strategies for hosting applications in cloud computing environment. They recommended further research for mapping these security strategies to technology and management methods.

Requirement Management

The design of IT infrastructure is preceded by requirement determination, structuring of requirements, and the selection of the best design from available options. Requirement determination and structuring of requirement are parts of requirement management (Ullah & Lai, 2011). The goal of requirement management is to understand the needs of customer, the problem to be solved, and constraints such as resource and time constraints (Ullah & Lai, 2013).

Information systems managers and analysts identify the requirements by having interviews or discussions with the stakeholders. They identify user needs during the process of stakeholder requirement definition and requirement analysis (Carrillo de Gea et al., 2012). Carrillo de Gea et al. surveyed requirement engineering tools and identified eight features of these tools. These features are – (a) elicitation, (b) analysis, (c) specification, (d) modeling, (e) verification and validation, (f) management, and (g) traceability. Ruiz-López, Noguera, Rodríguez, Garrido, and Chung (2013) introduced a requirement engineering method for analysis of ubiquitous system, called REUBI. The REUBI method utilizes a goal-centered and scenario-assisted approach for designing ubiquitous systems. This method is applicable to the different domains of ubiquitous systems, such as context-aware, dynamic, and interactive systems.

Information systems managers should select right cloud platform after analysis of four attributes of requirement – (a) who are the stakeholders, (b) desired location of the system, (c) in what timeframe system is needed, and (d) what needs of stakeholder the system will fulfill (Ullah & Lai, 2011). By following proper requirement management

process, IT managers and architects can collect detailed user requirements; such as functionality, availability, data storage, security, and QoS requirements. Information system managers may select right cloud platform after analysis of the requirements.

Information Technology Processes

Processes are a series of activities that produce a specific outcome. Organizations have multiple processes. A process may have dependencies on other processes. Process definition includes a description of actions, dependencies, and sequence (Škrinjar & Trkman, 2013). Processes are essential part of various departments of organizations, including IT departments. Every department may have their own set of processes.

There are four main characteristics of processes (Škrinjar & Trkman, 2013) (a) process are measurable and performance driven, (b) processes have individually identifiable and countable results, (c) processes deliver results to customers or stakeholders, and (d) processes are triggered by specific events. Processes are measured in different contexts; managers may measure cost and quality, while engineers may be measuring efficiency and duration of the process. Processes must deliver some output or result, which may become input for other processes. A group of processes may provide a specific service in organization which provides value to the customers.

Cloud Computing Decision Frameworks

Ross (2010) completed a survey of 38 IT managers located in upstate New York, involved in the cloud computing adoption decisions. Ross found a strong relationship between the decision to adopt cloud computing and four independent variables: cost-effectiveness, need, reliability, and security effectiveness of cloud computing.

Li, Zhang, O'Brien, Cai, and Flint, (2013) reviewed literature on cloud computing and identified three main aspects of cloud computing (a) performance, (b) economics, and (c) security. Main criteria for performance evaluation are communication, computation, memory including cache, storage, and overall performance. Cost and elasticity are two parameters of economic evaluation. Authentication, data security, and infrastructure are three critical components of cloud security. Customers may compare these features among various cloud offering and select the service which meets their requirements the best.

From security and privacy perspectives a cloud customer should verify (a) the physical location of cloud, (b) security of information, and (c) the trustworthiness of the systems in cloud computing environment (Sun, Chang, Sun, & Wang, 2011). Privacy, security, and trust issues vary in different types of cloud application environments. Sun et al. recommended further research for developing a management framework for meeting security, privacy, and trust requirements when adopting cloud computing.

Garrison et al. (2012) collected data from 314 companies across the globe in various industries for exploring factors behind successful cloud implementations. They found that technical, managerial, and relational capabilities were significant factors in achievement of competitive advantage from cloud computing adoption. Information systems managers and cloud service providers should build a trust based relationship (Garrison et al., 2012). Cloud service providers build trust with clients through effective communication and excellence in service. Information systems manager on the client

side should get the perception that cloud service provider understands the client's needs and works in the best interest of the client.

People, process, and service management issues should be given high priority when transitioning IT services to public, private, or hybrid cloud (Chang, Walters, & Wills, 2013). Many IT leaders focus only the technological aspect of cloud computing, which is only one of the success factors of transition to cloud. The cost of transitioning to the cloud is not just limited to finances; operational, data privacy, legal, and compliance issues are critical for the success of transition and may have larger cost implications. Chang et al. presented the cloud computing business framework (CCBF) for organizations adopting cloud computing. They described the top-down strategic relations between the business models and IT services. They have validated working of CCBF by referencing key benefits of IT service management frameworks PRINCE2 2009, ITIL V3, and IBA SOA.

Han (2011) presented a comparison of the total cost of ownership (TCO) between (a) hosting application by purchasing own hardware and (b) hosting in the public cloud. Total TCO was much lower in cloud hosted applications compared to purchasing own hardware. This comparison was based on two case studies of migration of web applications to the cloud.

Garg, Versteeg, and Buyya (2013) proposed a framework for comparing services of cloud service providers called SMICloud. The SMICloud framework is based on service measurement index (SMI), developed by the cloud services measurement index consortium (CSMIC). One of the challenges of QoS measurement in the cloud is due to

variations in the performance of virtual machines over time. The SMICloud framework overcomes this challenge by using historical performance data of cloud infrastructure for evaluating QoS of cloud service providers (Garg et al. 2013). The SMI framework of CSMIC compares QoS of cloud service providers by comparing seven quality attributes (a) accountability, (b) agility, (c) assurance of service, (d) cost, (e) performance, (f) security and Privacy, and (g) usability. Garg et al. presented a key performance indicator (KPI) metrics for evaluating SMI quality attributes based on the client's needs.

The SMICloud framework contains total 15 quantifiable indexes; these are (a) service response time, (b) sustainability, (c) suitability, (d) accuracy, (e) transparency, (f) interoperability, (g) availability, (h) reliability, (i) stability, (j) cost, (k) adaptability, (l) elasticity, (m) usability, (n) throughput and efficiency, and (m) scalability. The SMICloud framework may be used for generating comprehensive metrics, which could be used for comparing services of different service providers. Organizations may also utilize these metrics for evaluating the performance of their private cloud infrastructure.

Three main categories of costs involved with IT infrastructure are labor cost, infrastructure cost, and administration cost (Singh & Jangwal, 2012). A comparison of the total cumulative costs associated with three types of infrastructure (a) a public cloud, (b) an in-house private cloud, and (c) a private cloud leased from external service provider shows that private cloud costs the least in the long term (Singh & Jangwal, 2012).

Quality of Service

Cloud computing infrastructure should be capable of dynamically managing load changes. It should also meet SLA between cloud services provider and customer (Maurer, Brandic, & Sakellariou, 2013). These SLAs contain QoS goals, and cloud service providers may have contractual obligations to meet these QoS goals. Also, cloud service providers should ensure efficient utilization of resources for keeping their costs down. It is important to maintain a balance between the QoS goals attached to customer SLA and efficient resource utilization (Maurer et al., 2013).

Earlier, for practical purposes QoS was considered important in audio and video applications (Tolosana-Calasanz et al., 2012). A delay in audio/video processing and transmission would significantly reduce quality of audio, video or may make it completely unusable. Another area of computing where QoS is critical is *urgent-computing*, which is about prioritizing compute and data resources for emergency computations such as severe weather prediction (Tolosana-Calasanz et al., 2012). Tolosana-Calasanz et al. proposed a workflow system for enforcing QoS of multiple scientific workflow instances, which may be used for managing QoS in *urgent computing domain* of scientific computing.

Web services are often hosted on cloud infrastructure. Complex solutions could be developed by utilizing multiples web services with *web service composition* (Qi, Dou, Zhang, & Chen, 2012). Although, web service may publish the non-functional characteristics, such as QoS, managing QoS of web service composition in the cloud environment is a major challenge. Qi et al. presented a QoS aware web service

composition method called *local optimization and enumeration method (LOEM)*. The LOEM method is QoS aware and improves the efficiency of web service composition, when several composite solutions are available.

Montero, Moreno-Vozmediano, and Llorente (2011) proposed the design of a high throughput elastic computing cluster that could dynamically change the number of worker loads for meeting varying load. This elastic architecture utilizes virtualization for hiding complexity and isolating workloads. Remote cloud infrastructure, such as Amazon EC2 could be used for extending capabilities of elastic computing cluster (Montero et al., 2011). Use of remote cloud gives flexibility of dynamically changing cluster capacity, according to a given budget and performance policy.

C. Huang et al. (2013) proposed a resource allocation optimization system for cloud computing. This system estimates resource utilization based on the SLA of each process, and optimally allocates virtual machines on physical hardware while meeting performance requirements agreed in the SLA. The proposed system is designed for green computing while also maintain QoS to meet the customer SLA. Green computing reduces emission of greenhouse gases and also reduces cost, due to saving in power consumption.

One of the computing technologies that influenced cloud computing is grid computing. The idea behind grid computing is the utilization of heterogeneous geographically dispersed resources (Albodour, James, & Yaacob, 2012). Traditionally, grid computing was used in scientific computing; with growing computing requirements in the business context domains, grid computing could be useful for complex business

applications. Albodour et al. introduced a new model Business Grid Quality of Service (BGQoS), which could be used for QoS maintenance within the business context domains in grid computing. This model contains a QoS driven resource discovery, resource selection, and resource allocation model meeting high level requirements of grid resource consumers. The model includes a monitoring and reallocation function for detecting resource allocation failures and the QoS violations.

A platform abstraction layer (PAL) such as Aneka provides the user an application development and deployment platform in the cloud environment. Aneka is a middleware technology that provides service-oriented container framework for developing scalable applications in public and private cloud (Calheiros, Vecchiola, Karunamoorthy, & Buyya, 2012). Aneka could utilize desktop grids, which allows using spare compute power of desktops. It could be difficult to guarantee QoS to users just by using desktop grids because firstly, it is not possible to predict aggregate spare compute resources available in desktop, and secondly user load is also unpredictable. Calheiros et al. proposed a solution to guarantee QoS by utilizing capability of Aneka to integrate desktop grids with cloud. Cloud resources are billed on pay-per-use basis. Cloud resources will be used only when resources are not available on desktop grids. This would help in guaranteeing QoS and keeping costs lower, because Aneka would first use the resource with the lowest cost, such as the desktop grid before utilizing more expensive cloud resources.

High performance computing (HPC) is the discipline of IT which deals with compute resource intensive applications such as scientific computations and complex

engineering or business problems. Historically, HPC required building dedicated computing infrastructure (Mauch et al., 2013). The dedicated HPC infrastructures were designed for peak usage; therefore, computing hardware was not fully utilized all the time. Cloud computing can provide HPC solution with highly efficient resource utilization and lower cost. Use of cloud computing for HPC presents three challenges (a) guaranteeing network QoS; (b) dynamic provisioning of virtual machines, including cluster network configuration; and (c) isolation of the individual user or workload in the cluster interconnect network (Mauch et al., 2013). Mauch et al. presented an evaluation of high performance computing cluster in cloud, *HPC2* and proposed using InfiniBand network for cluster interconnect that allows individual network configurations with QoS.

Cloud Computing Adoption Challenges

Cloud computing is movement towards availability of computing services as utility (Armbrust et al., 2010). There are ten obstacles to the adoption of cloud computing – (a) availability of services, (b) data lock-in, (c) data confidentiality and auditability, (d) data transfer bottlenecks, (e) performance unpredictability, (f) scalable storage, (g) bugs in large distributed systems, (h) scaling quickly, (i) reputation fate sharing, and (j) software licenses (Armbrust et al., 2010). As cloud computing keeps on evolving, cloud service providers would continue to find solutions of these obstacles.

Cloud computing adoption challenges affect both cloud service users and cloud service providers (Erdogmus, 2009). Cloud users would not like to give away convenience of deploying and managing their application locally, which gives them a greater control over the environment. For cloud service providers, sustainability could be

a challenge (Erdogmus, 2009). Cloud service providers would have to meet the ever growing demand of resources and deal with the complexity of managing cloud infrastructure.

Cloud computing has transformed computing from product to a service. Cloud computing is not just a technological change; the way compute technology is provisioned and used, has changed. Adoption of cloud computing requires organizational changes and has economic implications (Khajeh-Hosseini, Greenwood, Smith, & Sommerville, 2012). Khajeh-Hosseini et al. reviewed available literature in this area and identified challenges in the area of cloud computing. Some of the important issues to be addresses are the utility billing model, the security, legal, privacy, and compliance.

Adoption of cloud computing based solutions could be beneficial to patients, doctors and other stakeholders of healthcare services such as insurance companies, administration, and the healthcare support staff (Kuo, 2011). Healthcare industry faces same obstacles in adoption of cloud computing as faced by industries in general. Information security, trust, and privacy are the biggest challenges. Subramanian (2012) explored the adoption of cloud computing in the pharmaceutical and life science industry. Subramanian highlighted the problem of low adoption of cloud computing in the pharmaceutical industry. Although, pharmaceutical industry could greatly benefit from adoption of cloud computing, adoption rate remains low. Main reasons for low adoption are the concerns for information security and pharmaceutical industries practice of allocating relatively less funding for IT.

Cloud computing and environmental sustainability

The power consumption in enterprise data-center could be in the range of 25 to 50 percent of their total power consumption (Harmon & Demirkan, 2011). The cost of data-center power consumption in the United States and European Union in the years 2007 was estimated about \$ 10 billion (Harmon & Demirkan, 2011). There are two aspects of electrical power consumption in data-centers, cost of power, and adverse effect on the environment due to the release of greenhouse gases from the power stations. Power stations may burn fossil fuel to fulfill the demand of electrical power. Therefore, reducing power consumption is an important goal of data-center managers. This issue is even more important for cloud data-centers due to their huge hardware installation base.

Cloud computing infrastructure is built by consolidating hardware and virtualizing compute infrastructure. Virtualization provides an opportunity for reducing power consumption by dynamically distributing virtual machines across different hardware that minimizes the power consumption (Lefèvre & Orgerie, 2010). Lefèvre and Orgerie presented the concept of the Green Open Cloud architecture, which is based on the power consumption by virtualized computers.

In a cloud environment, power consumption could be reduced by consolidating running virtual machines on less number of servers during the period of lower workload (Graubner, Schmidt, & Freisleben, 2012). Graubner et al. presented a solution for reducing energy consumption in cloud infrastructure running on Eucalyptus. They developed a method of consolidating load on the virtual machines by live migration of virtual machine and synchronization of storage using Distributed Replicated Block

Device (DRBD). They conducted several long-term and short-term tests using this method and achieved energy savings of up to 16 percent using this method.

Case Studies on Cloud Computing Based Solutions

Cloud computing based solutions are available for the individual users to all types of business and industries. For individual users, solutions are available for online storage, web based document management and collaboration tools, and many more. New cloud base solutions for enterprise are being released frequently. There are many case studies available in literature on cloud-based solutions; two such studied are discussed in this section.

Hexiao, Shiming, and Haijian (2012) developed a cloud-based education service for hosting online courses for universities. They utilized the PaaS services and application development tools provided by Google. The online system was developed with Google App Engine, Google Web Toolkit, and Ext-GWT. Hexiao et al. successfully ported traditional online courses to the cloud. At this time, cloud service providers do not support many frameworks, which are commonly used in traditional application hosting platforms (Hexiao et al., 2012). Hexiao et al. expects that in future cloud service providers will support more traditional frameworks, which may reduce the effort needed in migrating online courses to the cloud.

Cloud computing can be highly beneficial to the scientific computing. One such example in the areas of scientific computing is the migration of the Supernova factory (SNfactory) experiment to Amazon EC2 cloud, from a private Linux cluster (Brandic et al., 2011). The purpose of Supernova factory (SNfactory) experiment is to measure the

expansion history of the Universe. This requires complex calculations and high volume of data processing. In their research paper, Brandic et al. presented the performance and cost analysis of migration of this complex computing application to public cloud.

Cloud Computing Economies

Cloud computing does not require any upfront cost; customers pay a subscription fee or pay per use fees. Enterprises build their own private cloud for security, privacy, performance, compliance, and cost reasons. They may charge back their internal customers like a public cloud service providers. The total cost of ownership will depend on the application hosting requirements.

Total cost of ownership in cloud of two web applications of University of Arizona Libraries was much lower than purchasing own hardware (Han, 2011). Han's research report contains two case studies of migration of web applications to Cloud. The report contains a description of the process used by Han in evaluating various cloud services and cost analysis.

The revenue generated by cloud computing is growing, and it is expected to keep growing in the coming years (Leavitt, 2009). Earliest adopters of public cloud are web 2.0 startups, because these companies do not have a large budget for building datacenters (Leavitt, 2009). Also, web startups will not be able to make an accurate estimate of future customers. It will be expensive and time consuming for them to scale up or scale down infrastructure, cloud is the ideal solution for them.

A major attraction towards cloud computing is the lower cost. Many service providers are offering services at a lower price by using methods such as, over-allocating

compute resources, which may result in inferior performance during peak usage (Durkee, 2010). Also, some service providers are using older and less reliable hardware for reducing costs; this may affect the service level for customers. Durkee advised businesses to carefully evaluate service provider before subscribing to a cloud service.

Multimedia communication research and development require computer simulations, which has very high computational resource requirement. On the traditional infrastructure, simulations require multiple computers and may run for several days, extending the duration of project (Angeli & Masala, 2012). Angeli and Masala proposed a cost-effective framework for reducing the duration of simulation using cloud computing. Simulations may run on servers in the cloud on the pay-per-use basis. This also releases computers and development may continue on these computers while simulation runs in the cloud. Organizations may accelerate development time and reduce costs using this framework.

Cloud Computing Technologies

Cloud computing solutions use same hardware, operating systems, storage, and network as used in the traditional IT infrastructure. The cloud management software transforms infrastructure into the cloud. There are both commercial and open source solutions available for managing cloud. Some service providers have their own solutions.

Eucalyptus is an open source cloud virtualization and cloud management system, which has been designed with functionality similar to Amazon Web Services (AWS) (Milojicic & Wolski, 2011). Open Cirrus is a cloud testbed sponsored by Hewlett-Packard (HP), Intel, and Yahoo (Avetisyan et al., 2010). This project is in collaboration

with the premier research institutions including NSF, and various universities.

Eucalyptus is built on the technologies used in grid computing. Users who are already familiar with grid and high performance computing can explore cloud features of Eucalyptus while maintaining access to grid middleware (Q. Huang et al., 2013).

OpenNebula and CloudStack are other two popular open source cloud solutions (Q. Huang et al., 2013).

Cloud computing gives users a notion of unlimited computing resources. In reality, a cloud data-center has limited resources. There is always a risk that user's requirements of hardware infrastructure, software platform, and software may exceed the available capacity. This may result in the violation of agreed QoS SLAs with customers (Calheiros, Toosi, Vecchiola, & Buyya, 2012). Calheiros et al. presented InterCloud project, a system that dynamically scales applications across multiple cloud data-centers. InterCloud utilizes agents called cloud coordinators, which maintains performance, scalability, and elasticity of cloud applications by extending applications across multiple cloud data-centers.

Cloud computing facilitates ubiquitous application access to multiples devices such as a PC, Tablet, or a mobile phone. One of the challenges with the cloud-based applications is file access and sharing from mobile devices using a consistent interface (Mao, Xiao, Shi, & Lu, 2012). Mao et al. developed a mobile file service, Wukong, which provides a highly-available and user friendly data access method for mobile devices in cloud setting. Wukong utilizes a storage abstraction layer (SAL) and application plugins, which interacts with remote applications using their API available in

plugin. Mobile users can access remote application files like a local file using a user-friendly interface.

Transition and Summary

This study addresses the need of a process framework for managing QoS in private cloud. Section 1 covered the research problem, purpose of research, nature of study, and the literature review. Literature review contained an analysis of literature in the area of cloud computing, technology behind cloud computing, market trends, cloud adoption trends, and challenges of IT organizations. Section 2 will contain the details of the project such as the purpose statement, the role of researcher, details of participants, research methods and design, population and sampling, data collections, data analysis, and the data validation techniques.

Section 2: The Project

The purpose of this qualitative single-case study was to identify a process framework for managing QoS in private clouds. The purpose of Section 2 is to present the research design, design selection rationales, participant selection criteria, participant population, data collection procedures, data analysis methods, and data validation steps. The result of this study added to the body of knowledge in the area of IT service management.

Purpose Statement

The purpose of this research study was to explore ways to establish a process framework for managing QoS in private clouds. I used a qualitative single-case study design by selecting an organization in the state of California, which had successfully implemented private cloud. Data collection included analysis of documents, such as SOPs, policies, and guidelines related to private cloud and interview of 23 staff members involved in private cloud implementation. Participants were selected using purposeful sampling method.

Data collected were analyzed using the qualitative data analysis software Weft QDA, and findings were used for identifying a process framework. Data will be securely stored for 5 years. I triangulated data collected from different sources for the purpose of validation.

A process framework for managing QoS in private cloud may help organizations in adoption of cloud computing. Cloud computing facilitates an agile and flexible IT infrastructure, which results in business agility as confirmed by early cloud adopters

(Narasimhan & Nichols, 2011). The results of this study may also help an organization in taking advantage of latest solutions offered by cloud computing, such as support for mobile devices used by employees, partners, and customers and achieve competitive advantage (Narasimhan & Nichols, 2011).

Role of the Researcher

In a qualitative case study, researchers gather data from multiple resources, such as interviews, record, documentation, and artifacts (Yin, 2009). In qualitative studies, the researcher becomes an instrument for data collection (Houghton, Casey, Shaw, & Murphy, 2013). I have over 23 years of experience in various areas of IT, such as application development, network and system administration, and infrastructure management. I was able to understand technical terms used by participants. However, my experience in IT increased the possibility of bias and prejudice in IT-related studies. I ensured that personal bias and prejudice did not affect research result by objectively analyzing research data as described in data collection and analysis sections.

I had some experience of interviewing people prior to this study. I interviewed people during some of the courses in the master's degree program and doctor of business administration (DBA) program. I also interviewed people as a part of my work for activities such as requirement gathering or feedback. For research purpose, the interviewer must be highly skilled. After approval from the institutional review board (IRB), I practiced with sample questions with a pilot group for improving my interviewing skills.

The pilot study was conducted with three participants with two people experienced in supporting IT infrastructure and one supporting software applications. I followed up with these participants and collected their feedback on the effectiveness of interview questions. Pilot participants found interview questions to be effective. I analyzed pilot data using Weft QDA software and identified themes relevant to the research questions. I did not make any changes to the interview questions after pilot study. The data of the pilot study were saved in digital format on secured optical disks. The data in the computer files were encrypted using TrueCrypt software with AES 256 bit encryption. Optical disks containing encrypted files are stored in a safe and secured place. These disks will be preserved for the next 5 years with other data collected during study. A copy of encrypted files is backed up in a cloud-based storage system for extra protection of research data.

Participants

In this qualitative case study, I relied on information collected from multiple sources including the interviews of information system professionals engaged in activities related to private cloud at the organization studied. These interview participants shared their experience with private cloud implementation by answering the interview questions. Qualitative scholars use a smaller number of information rich cases for in-depth study. These information-rich participants were selected through the purposeful sampling approach (Kipkulei, 2013; Patton, 2001; Shaw, 2012). Information system professionals from selected organization, with experience in managing a private cloud or those responsible for development and maintenance of business applications in a private cloud,

participated in interviews. These information systems professionals had experience with QoS challenges in a private cloud and were able to provide information for this research study.

Twenty three purposefully selected IT professionals participated in this qualitative case study. The number of participants in a qualitative study should be small (Hennick, Hutter, & Bailey, 2011). Patton (2001) suggested selecting the number of participants depending on the type of study and did not suggest any specific number. The participants in a single-case study were from only one organization and were selected using purposeful sampling; therefore, a sample size of about 20 was sufficient (Shaw, 2012). In addition to interviews, I gathered data from documents such as SOPs, policies, and guidelines related to private cloud implementation.

Management of QoS in cloud requires the involvement of IT infrastructure specialists supporting virtualized infrastructure for cloud and clients (Chang et al., 2013). The clients in this case were application architects and application support specialists. Thirteen IT infrastructure specialists and 10 application architects and specialists participated in interviews. These participants were able to provide in-depth information about QoS issues and the type of processes they followed for managing QoS.

The participants selected for interview had at least 5 years of experience of working in IT. These participants shared their own experience with QoS issues they experienced in their respective area. As these participants were experts in their area, they provided a wealth of information. Information collected from 23 participants, in

combination with the documents collected from the organization, was sufficient for this study.

I selected an organization for case study that successfully implemented private cloud. I selected an organization where I was able to get access to IT personnel involved in private cloud implementation and relevant documents. I did not select participants with a close relationship with me, such as peers at work, subordinates, supervisors, family members, and relatives. I provided a five dollar coffee gift card to participants as a token of appreciation for their participation.

After selecting the organization for the case study, the next step was the identification of potential participants for study. I requested the manager of the selected organization to give me permission to contact staff members involved in the implementation of private cloud. After receiving the manager's permission, I contacted prospective participant using e-mail and through phone. I verified a prospective participant's suitability for the interview using the participant selection criteria in Appendix A. I explained the purpose of study and the purpose of the interview to the participants. I also informed each participant that the information collected from the participants will be kept confidential and their identity will not be published in the research report. I made it clear that the participation in the study was voluntary and participants had the option of withdrawing from the study at any time by contacting me in-person, over the phone, or through e-mail. None of the participants requested to withdraw them from study. All of the communication described above was in person or over the phone. I sent all of the information described above to participants through e-

mail, immediately after a phone or in-person conversation. I sent the terms and conditions of the interview to the prospective participants in advance for their review. The terms and conditions for the participants are available in the consent form attached as Appendix C. I explained to the participants that they will receive a copy of the approved study and a brief summary of the findings after the completion of the study.

Research Method

In this study, I addressed the issue of QoS in cloud computing infrastructure by identifying a process framework. I did not find literature that contains such process framework. Most literature on cloud computing QoS was focused on the technical aspect of QoS. Qualitative methods are appropriate for research in areas in which there have been limited or no research in the past (Marshall & Rossman, 2010). Quantitative methodology includes scientific methods for accepting or rejecting one or more hypotheses based on past research (Arghode, 2012; Bryman, 2012). Considering this, I selected qualitative method for this study.

A researcher using qualitative methods seeks meaning by exploring and analyzing information collected from the participants (Arghode, 2012). Qualitative researchers are the research instruments. They gather information through interviews by asking open ended questions, or through observations, analyzing documented data, or recorded audio-visual data (Bryman, 2012; Willig, 2008). Some qualitative studies may result in developing new theories, which could be validated using quantitative methods.

A scholar starts quantitative research with one or more hypotheses, which are proved or rejected using survey of a large sample compared to qualitative study sample.

Quantitative studies are deductive in nature where qualitative studies are inductive (Bryman, 2012). In quantitative studies, scholars rely heavily on the past research and established theories. In this study, I focused on exploring a process framework for managing QoS in cloud infrastructure. I could not find past research on this topic; therefore, qualitative methods were best suited for this study.

Research Design

The research problem was about identifying a process framework for managing QoS in private cloud. The qualitative method of inquiry was appropriate as discussed in the previous section. Five approaches used for qualitative study are (a) narrative, (b) phenomenological, (c) grounded theory, (d) ethnographic, and (e) case study (Lopez, 2012). Narrative approach is based on lived and told stories of individuals. Researcher is required to understand the situation and context; participants' work, home, ethnicity and other relevant information and extract meaningful information from those stories (Marshall & Rossman, 2010). In narrative approach, the scholar focuses mostly on individuals; goals of this research design were to extract information from experienced information systems professionals about their experience of managing QoS in cloud based infrastructure. This required specific information from participants, not a detailed story; therefore, narrative approach was not the best approach for this study.

Grounded theory approach is used for developing a new theory. Scholars use, grounded theory approach, when current literature does not adequately describe a phenomenon (Leedy & Ormrod, 2009). The goal of this research was to provide a better insight to a business problem, not necessary developing a new theory. In ethnographic

studies, scholars study cultural groups and use case study approach for in depth analysis of a specific business case. These two approaches did not align with this research problem, which was about identifying a process framework for managing the QoS in private cloud. Ethnography did not align to this research problem, because this study could not be generalized by studying culture of a specific group.

Phenomenological approach is about studying experiences of individuals, who experienced a particular phenomenon (Polkinghorne, 1989). The researcher conducts in depth interviews of individuals who experienced this phenomenon and learns from their lived experience (Marshall & Rossman, 2010). This study was not about participants' lived experience; therefore, I did not select phenomenological approach for this research.

Selection criteria for research design depend on (a) the research question, (b) the extent of control the researcher may have over actual behavioral event, and (c) the degree of focus on contemporary events (Yin, 2009). A case study is appropriate when (a) the research questions are of type how and why, (b) the researcher does not have control over the behavioral events, and (c) the focus is on contemporary events (Yin, 2009). The overarching research question for this study was: How organizations can successfully manage QoS in private cloud? The adoption of private cloud is a contemporary issue and I had no option to control implementation of private cloud in an organization. Therefore, case study design aligned well with the objectives of this study.

A single-case study design is appropriate when a critical case is selected for the study. Selection of a critical case is a valid rationale for selecting the single-case design over multi-case design (Yin, 2009). I selected an organization which had successfully

implemented private cloud for the case study. Therefore, a qualitative single-case study design was appropriate for this study.

Population and Sampling

Selection of the population sample should align with the goals and research question of this study. In this qualitative single-case study, I focused on the issue of managing QoS in cloud infrastructure. This single-case study was conducted in an organization selected using purposeful sampling. The selected organization was located in the state of California; it had successfully implemented a private cloud within last two years, and hosted more than 50 computer applications in the private cloud. The research participants were IT professionals working with the selected organization. Thirteen participants were selected with background in IT infrastructure management and 10 participants with background in business application support. Total number of participants was 23.

Qualitative studies are done with a small number of information rich participants (VanderStoep & Johnson, 2009; Willig, 2008). In this research study, I used the single-case study method. I used two data collection methods, interviews and document review. I collected documents such as SOPs, policies, and guidelines related to private cloud implementation. Information systems professionals involved in the private cloud implementation in the organization participated in the interviews, contributing to another set of data for this study. Participants of this study were highly experienced information systems professional with experience in IT infrastructure management and application

development and support. The participation selection criteria are listed in Appendix A, and a detailed description is provided later in this section.

I conducted unstructured interviews, which helped me in extracting detailed information from each participant. Considering the depth of information received from participants, a sample size of 23 was sufficient for this study. With my experience in IT, I was able to steer direction of the interviews towards answers of research questions.

Data collected in qualitative studies should be kept in the raw format for analysis. Each interview resulted in the huge amount of data. I recorded interviews with an audio device and later transcribed for analysis. Due to the huge amount of data to be analyzed, the number of participants in a qualitative study should be small (Willig, 2008); otherwise, it becomes difficult to manage and analyze data. I analyzed data collected from 23 participants using Weft QDA software and drew conclusions towards the answers of research questions. The data collected for the study were sufficient because, the information provided by participants was based on their extensive experience in managing enterprise infrastructure and applications.

Participants were selected using purposeful sampling method. Purposeful sampling method is about selecting information rich sources (VanderStoep & Johnson, 2009), who could provide valuable information for the study. Purposeful sampling was appropriate for this study; because, all the IT professionals may not have experience with cloud technology or shared resources. Information provided by such participants may not be useful for study. I used criterion sampling; participants were selected using criterion

relevant to the research questions (Willig, 2008). This ensured that the participants provided information relevant to research questions.

I requested information systems managers in the selected organization to suggest people who could be the participants for this study. I contacted the potential participants and completed an evaluation form by asking them questions about their experience. Data in these forms were used for evaluating suitability of the participants for the research. Participants, who did not meet the above criteria, were excluded from the study. A clear selection criterion was developed for selection of the participants.

Application support professional selection criteria:

1. Participants had a bachelors, masters, or doctoral degree in the computer science, electrical engineering, IT, or equivalent.
2. Participants had at least 5 years of experience supporting business applications as developer, analyst, architect, project manager, manger, director or similar role. These participants were rich source of information due to their extensive experience.
3. Participants supported applications hosted in virtualized shared infrastructure for at least 2 years. Application specialists, who developed or supported application on virtualized infrastructure, provided relevant information for the study.
4. Participants had at least 2 years' experience in managing application service level agreements. Those who managed application service level agreements understood challenges of managing service level agreement

and shared strategies to overcome those challenges. These participants provided information relevant to research questions.

IT infrastructure professional selection criteria:

1. Participants had a bachelors, masters, or doctoral degree in computer science, electrical engineering, IT, or equivalent.
2. Participants had at least 5 years of experience supporting IT infrastructure. The IT infrastructure professionals shared relevant information about management of QoS in shared infrastructure.
3. Participants supported virtualized infrastructure for two or more years.
4. Participants supported shared IT infrastructure with multi-tenancy for 2 years or more. Infrastructure specialists with experience in multi-tenancy provided information about managing QoS in cloud; because multi-tenancy is one the prime features of cloud computing.
5. Participants supported IT infrastructure with established service level agreements to clients. The main reason behind QoS is ensuring that service level agreements with clients are met effectively.
6. Participants had at least 2 years' experience of supporting production IT infrastructure.

Ethical Research

By participating in this study, the participants and their employer were not exposed to any risk. The information collected from the participants was not sensitive private information and is not likely to result in any controversies. For the purpose of

protecting the participants names, employer details and any other personally identifiable information of participants was kept confidential. None of the personally identifiable information about participants and their employer was published in the report.

Participants were contacted and interviewed only after getting approval from IRB.

Participants received the declaration form attached in Appendix C prior to the interview. The declaration form contains the purpose of research and how data collected during research will be handled. I also explained the contents of declaration to participants and interview was conducted only after their acceptance. These interviews were recorded using an audio device with prior consent of the participants.

Data Collection

Instruments

The researcher in the qualitative studies is the primary instrument. Participants attended interviews in person. All the interview questions were open-ended. Prior to interviewing participants, I tested research questions with a pilot group of three participants for determining the effectiveness of the interview questions. The participants of the pilot group were excluded from the main study. I audio recorded interviews with the consent of participants.

The research questions are listed in Appendix B. I conducted unstructured interviews for eliciting responses relevant to the research questions. This is an appropriate interviewing strategy when the researcher and the participants are highly knowledgeable about the research topic (Cooper & Schindler, 2010). Each interview question listed in Appendix B aligns to one or more research questions. During

interviews, I re-worded some of the questions or asked follow up questions based on the participant's knowledge and experience.

Participants were sent consent forms (see Appendix C) and the research questions, one week before the interview through email. I contacted the participant over the phone three to five days before the interview and explained the purpose of interview and confidentiality. I sent an email reminder and a Microsoft Outlook meeting invitation, one day before the interview. Interviews were about one hour in duration; I conducted interviews in the private rooms, free of distraction. Other than audio recording of the interviews, I also made notes of my observations, such as context in which participant made a statement, participant's body language, and any other non-verbal cues. During the interview, I made sure that the participants clearly understood questions by explaining in detail the context, in which question has been asked. Depending on the participant's knowledge and experience and the flow of the interview, I asked only subsets of questions to some participants.

Conducting a pilot testing with three volunteers aided in validating interview questions; I took feedback from volunteers doing pilot testing on the effectiveness of questions and also analyzed their responses. This helped in ensuring that interview responses were providing answers to research questions. I received a positive feedback from the pilot participants on the effectiveness of interview questions. The analysis of pilot interview data using Weft QDA provided five different themes, relevant to research questions. I used the interview questions for further study without any modification. I

also took necessary steps for ensuring reliability and validity of data collection instruments.

Two categories of information systems professionals - infrastructure specialists and application specialists participated in the interviews. These two participants groups viewed research problem from different perspectives. The analysis of data from these two groups resulted in a coherent theme; this ensured reliability and validity of data. Yin (2009) recommended triangulating different data sources of information for the purpose of validation of data. In addition to interviews, I collected documents such as SOPs, policies, and guidelines and analyzed with Weft QDA. I cross-verified themes from documents with interview data and they were coherent. Above approach of cross-verifying data collected from participants of different perspectives helped in ensuring validity and reliability of data.

I selected a peer-reviewer for reviewing the study and compared reviewer's interpretation from my own assessment. Two accounts did match; this further confirmed validity of the study. I did not share any personally identifiable information of participants and identity of the organization being studied with peer-reviewer.

Researcher's bias in some cases may distort findings of study; the researcher should clarify his own bias in study (Yin, 2009). Researcher's reflection on bias would help readers in correctly interpreting the findings of the study. I have experience in managing shared IT infrastructure and private cloud; therefore, I may have focused more in infrastructure compared to applications. I made a conscious effort, for remaining unbiased throughout the study.

Audio recordings of interviews were stored on optical disks and I kept them in a safe and secured place for 5 years. All the computer files were encrypted using TrueCrypt software with AES 256 bit encryption strength. An extra copy of all encrypted files was uploaded in a secured cloud base storage and will be retained for 5 years. The recording will be made available to the research chair, URR, or other relevant parties on demand. The transcripts do not contain names or any other personal identifiable information about the participants.

Data Collection Technique

Data collected through the interviews were transcribed as text. The details of the interviewees were coded for hiding their identities. I have saved all the information in secured computer files. Computer files were automatically backed up every six hours for avoiding loss of data due to computer malfunction or accidental file deletion. All the documents collected for the case study were also saved and secured like transcribed interviews text.

A pilot study with three participants was conducted after the IRB approval for validating data collection, organization, and analysis techniques. I followed all of the data collection steps during the pilot study for identifying any weaknesses if present in the process. With the help of pilot study, I validated that the data handling process was appropriate.

Data Organization Techniques

I analyzed transcribed data for identifying the common themes among responses. I used Weft QDA qualitative analysis software for identifying themes in the transcribed

data. I selected Weft QDA software after reviewing few recent doctoral studies completed by Walden students. Two Walden graduates completed their PhD dissertation using Weft QDA in recent years (Diffin, 2011; Scott, 2012), and another Walden graduate successfully completed Ed.D study (McCoy-Wilson, 2011) by using Weft QDA software for qualitative analysis. For familiarizing myself with Weft QDA, I installed software on a personal computer; after that, searched through the collection of scholarly article and assessed its processing capabilities. Qualitative analysis capabilities of Weft QDA were sufficient for this study.

Transcribed data were stored in a secure computer within encrypted files. I used a strong password, which was eight characters long and a mix of lowercase, uppercase letters, numbers, and punctuations. There was an additional layer of data security, because data were kept in the computer files encrypted with TrueCrypt software with AES 256 bit encryption strength. Encrypted files on the computer were automatically backed up to an Apple Time Capsule backup device every 6 hours. Also, a weekly backup of encrypted data was copied on secured cloud-based storage on Mozy.com. These data will be securely retained for at least 5 years from the dates of study approval.

Any document on paper such as field notes, printed documents collected from the organization, and consent forms were locked in a secured file cabinet at my home. These documents will be retained for 5 years and will be cross-shredded after that time. All the electronic documents will be permanently erased after 5 year period.

Data Analysis

I listened to all the interview audio recordings and transcribed in Microsoft word. Weft DBA software was used for the analysis of transcribed data. Weft QDA software provided output in the comma separated values (CSV) file format that were imported in the Microsoft Excel worksheets. Weft QDA helped in identifying themes in the interview data. The goal of theme identification was to find the answers of the research questions. Each identified theme was matched to one or more of the research questions. Answers pertaining to the interview questions in this study addressed the research questions and the overarching research question: How organizations can successfully manage QoS in private cloud?

Research questions are listed below:

1. What processes should be implemented in private cloud environment for guaranteeing QoS?
2. How can information systems managers identify QoS requirements of private cloud service users?
3. How can information systems managers proactively identify QoS issues, which may cause application performance degradations?
4. What strategy should be followed for prioritizing allocation of computing resources in private cloud?

Following were the interview questions:

1. What is your experience in application development, application support, infrastructure design, or infrastructure support? Please explain.

2. What processes you followed for ensuring consistent application performance in the private cloud?
3. How do you determine performance requirements for applications and ensure that application performance requirements are consistently met?
4. How do you determine hardware resource requirements for applications?
5. What challenges did you face in migrating applications to private cloud?
6. How do your determine performance, availability, and security SLAs for applications? How do you document these SLAs?
7. What proactive actions you take for ensuring that applications hosted in private cloud consistently meet established performance SLAs?
8. What information do you periodically receive and share with different IT groups in your organization?
9. How do you ensure that applications hosted in private cloud continue to get adequate system resources?
10. What additional information would you like to share that you may not have had the opportunity to address?

While separating data in the themes, I started writing narratives that were later included in the study. The qualitative research is a continuous process; a researcher should not wait for completing all the interviews before analyzing data. I started data analysis right after the completion of interview. Once data were identified in themes; further analysis led to the answers of the research questions. Another goal of data

analysis was tying results to the conceptual framework the theory of disruptive innovation.

The next step was to read the documents collected from the organization and look for the themes that may give answers to research questions. I analyzed documents using Weft QDA software and identified common themes among all the textual data. Textual information with each theme was stored in separate Excel worksheets. I read data categorized in various themes; interpreted the information provided by participants and wrote narratives that provided answers to the research questions. Notes and annotations made during analysis were used for making logical conclusions. I created a diagram and included in the findings based on the data analysis.

Reliability and Validity

A researcher must ensure that the results of study are reliable and valid. There are several steps that must be completed for ensuring reliability and validity. The first step towards reliability and validity is making sure that data collected are handled and stored properly (Gibbs, 2008). I made sure that transcripts were recorded without any error. I wrote all the interview transcripts and analyzed the data. I knew the context in which the data were gathered; this ensured that there was no drift in the meaning of coding when data were tagged with codes in various themes.

Murphy and Yelder (2010) suggested six strategies for strengthening validity of the qualitative data - (a) triangulate, (b) use member checking, (c) use rich, thick, (d) clarify the bias, (e) spend prolonged time in the field, and (g) use peer debriefing. For this research study triangulation, clarifying the bias, and peer debriefing were appropriate

as described in the data collections instrument section. I made conscious effort to not influence findings with my personal bias. Few participants' response did not align with any of the themes and were filtered out.

Transition and Summary

This qualitative single-case study explored ways to develop a process framework for managing QoS in cloud computing. Research method and design were selected after carefully evaluating various research methods and design. Twenty three information systems professionals participated in the interview for this study. Their responses were recorded, transcribed, analyzed, and validated. After analysis of data, I presented the process framework for managing QoS in private cloud.

Section 3: Application to Professional Practice and Implications for Change

Overview of Study

The purpose of this study was to identify the processes followed for managing QoS in successful private cloud implementations. I obtained data by interviewing 23 information systems professionals working in an organization in California that successfully implemented private cloud within 2 years of this study. Other sources of data were technical and process-related documents provided by the organization. The analysis of the data provided answers to the research questions. The overarching research question for this study was the following: How can organizations successfully manage QoS in a private cloud platform? The four major research questions for this study were the following:

1. What processes should be implemented in private cloud environment for guaranteeing QoS (Liu et al., 2012)?
2. How can information systems managers identify QoS requirements of private cloud service users (Sunyaev & Schneider, 2013)?
3. How can information systems managers proactively identify QoS issues, which may cause application performance degradations (Faniyi et al., 2012)?
4. What strategy should be followed for prioritizing allocation of computing resources in private cloud (Chung et al., 2013)?

I selected a qualitative single-case study design. For the sample, I selected an organization in California that successfully implemented private cloud within last 2 years.

Data collection for the study included the interview of 23 IT professionals and documents related to private cloud implementation at the organization. I uncovered six different themes after analyzing the data using Weft QDA software. These themes were (a) end user expectations, (b) SLAs, (c) application architecture, (d) infrastructure sizing, (e) monitoring, and (f) trending analysis. These themes were used to identify the process framework used by the organization for managing QoS in private cloud.

According to the analysis of the data collected during this study, there is a need to align application response time with end users' expectations. The QoS is measured by the degree of user satisfaction with the performance of service (Kourtesis, Alvarez-Rodríguez, & Paraskakis, 2014). Respondents mentioned that the management of QoS in the cloud is challenging due to the distributed nature of infrastructure and because the IT staff needs tools for the dynamic mapping and monitoring of the application execution environment.

Presentation of the Findings

The result of data analysis by Weft QDA software helped me in identifying six major themes. These themes cover different processes used by the organization for managing QoS in private cloud. The primary goal of managing QoS is meeting the application performance expectations of end users. In the context of cloud computing, QoS is the ability to meet service performance requirements of users while delivering cloud-based services (García, Espert, & García, 2014). QoS data are the key-enablers in the design and identification of application performance SLA (Kourtesis et al., 2014). For end users, QoS should be measured in the terms of UX, such as response time and the

reliability of service. I found that QoS should be focused on UX, which may require the proper allocation of computational and network resources in the private cloud. This aligns with the body of literature on this topic.

Fifty-six percent of the respondents (13 out of 23) mentioned that the adoption of cloud computing was disruptive for organizations as many IT processes have changed. The role of the IT department has changed from IT infrastructure owner to a service provider. This conforms to the theory of disruptive innovation, the conceptual framework behind this study.

Eighty-three percent of the respondents (19 out of 23) stressed the need of managing application performance in cloud. This same need was evident by the analysis of documents collected from the participating organization. This aligns with the existing body of literature, as there are many articles on QoS being major issues in the cloud. However, the focus of research in most of the QoS management-related scholarly articles is on technical solutions, not on a process-based approach.

Theme 1: End User Expectations

The participants of this study suggested that the goals of QoS management should be meeting the expectations of end users, who are often referred to as customers. They measured UX in terms of application response time, availability of applications, errors encountered by users, and advanced notification of application downtime due to planned maintenance. The participants suggested that maintenance activities that may result in QoS degradations should be planned when there is low or no usage of applications by end users. The maintenance window selection criteria in the maintenance and upgrade

guidelines confirmed the same. About response time, the organization had a recommended minimum response time of 7 seconds; however, the response time varied depending on the type of application. Application users do not expect an inferior QoS after migrating applications to the private cloud. In fact, users expect superior QoS after migration to the private cloud because of newer, faster hardware and the newer version of the software. Any deterioration in service after the migration to private cloud may result in end user dissatisfaction.

Theme 2: Service Level Agreements

SLAs are the criteria for determining the achievement of QoS goals. SLA documents contained maximum response time for user transactions, agreed maintenance windows, and availability requirements as shown in Table 1. An important document to complement SLA was the application profile document. The application profile document contained several QoS related items including (a) maximum number of concurrent users, (b) total number of users, (c) peak usage period, (d) external services used by application, (e) expected annual user growth, (f) expected annual data growth, (g) application-related batch job details, and (h) minimum hardware requirements (see Table 2 and Table 3). Participants noted that the various support groups within organizations have operational level agreements (OLAs) among them. These OLAs defined the level of service expected from the internal support groups. There is a dependency between SLAs and OLAs. The terms of SLAs should not conflict with OLA. For example, if SLA requires a 99.9% availability of applications, OLAs should ensure a 99.9% or higher

availability of underlying infrastructure. Table 4 contains SLA and OLAs of one of the applications that were present in the application profile document.

Table 1

Information from SLA documents

Application	Maximum Response Time	Maintenance Window	Availability Requirements
Application 1	7 Seconds	Saturday, 5 PM to 7 PM Pacific Time	99.99%
Application 2	7 Seconds	Sunday, Midnight to 4 AM Pacific Time	99.99%
Application 3	4 Seconds	Friday, 6 PM to 8 PM Pacific Time	99.99%
Application 4	4 Seconds	Last Saturday of month, 5 PM to Sunday, 5 AM Pacific Time	99.99%

Table 2

Information from Application Profiles

App.	Max Number of Concurrent Users	Total Number of Users	Peak Usage Period (Pacific Time)	External Service	Annual User Growth	Annual Data Growth	Number of Batch Jobs
1	200	1000	5 AM to 5 PM Mon to Fri	None	5%	10%	1
2	50	350	6 AM to 5 PM Mon to Fri	Interface to ERP	10%	15%	2
3	20	500	6 AM to 5 PM Mon to Fri	Interface to SRM	10%	20%	3
4	70	5000	6 AM to 9 PM Mon to Sun	Interface to CRM	10%	20%	2

Table 3

Information about Minimum System Requirement from Application Profile

<u>Minimum System Requirement</u>						
App	<u>Web/Application Server (2 Nodes)</u>			<u>Database Server (2 Nodes)</u>		
	CPU	RAM	Storage	CPU	RAM	Storage
1	2 x 1 Ghz or faster, 64 bit	8 GB	10 GB	4 x 1 GHz or faster, 64 bit	32 GB	100 GB
2	2 x 1 Ghz or faster, 64 bit	4 GB	6 GB	2 x 1 GHz or faster, 64 bit	16 GB	40 GB
3	2 x 1 Ghz or faster, 32/64 bit	4 GB	8 GB	2 x 1 GHz or faster, 64 bit	16 GB	50 GB
4	2 x 1 GHz or faster 64 bit	8 GB	20 GB	4 x 1 GHz or faster, 64 bit	32 GB	60 GB

Table 4

Service Availability Requirement

Service	Availability Requirement
Application	99.99%
Web Server	99.995%
Database Server	99.995%
Storage Server	99.999%

Theme 3: Architecture

One of the advantages of cloud is elasticity. Respondents mentioned that cloud computing provided limited vertical scalability. Vertical scalability of the infrastructure is the ability to increase system resources, such as the number of CPUs and the amount of memory. Applications can take advantage of cloud elasticity through horizontal scalability. The horizontal scalability of the infrastructure is the ability to provide multiple server instances for hosting applications. One of the inherent features of cloud computing is dynamic elasticity, which is achieved by adding more instances of web servers, application servers, and database servers and with the use of load-balancers. In response to Question 5, one of the challenges that respondents identified was the ability of some of the applications to scale across multiple load-balanced virtual instances. One respondent mentioned that developers enhanced applications code for one of the applications to make it horizontally scalable. Another respondent explained that few applications had little workload and they did not expect an increase in workload in future. Such applications were able to meet QoS requirements after deploying on a single virtual server. There was no need to make these applications horizontally scalable.

Theme 4: Sizing

Another theme that emerged after the analysis of interview data and documentation was infrastructure sizing. Virtualization allows the sharing of computational resources in cloud. Virtualizations provides agility to IT infrastructure; new virtual machines are quickly built and made available to cloud service users. These virtual machines share hardware resources of host computers, such as CPU, memory,

storage, and network. Since all the virtual machines do not always run at full capacity, cloud administrators often overprovision computational resources. For example, virtual machines are assigned virtual CPUs also known as vCPUs. These vCPUs could either be mapped to a single physical CPU core on the host computer, or multiple vCPUs could be carved from a single CPU core using overprovisioning. In case of overprovisioning, a single CPU core shares processing power among multiple virtual machines, which may result in resource contention. Sharing CPU cores may be acceptable for applications in which intermittent performance delays are acceptable.

In response to the interview Questions 2, 3, and 9 participants mentioned the resource reservations guidelines they followed. For CPU intensive applications such as database servers, they assigned dedicated vCPUs; this ensures that the guest operating systems and the underlying processes continue to get sufficient processing power. The organization also had a policy of selecting dedicated hardware for critical applications where hardware resource requirements were high. The criterion for selecting dedicated hardware in the documents was more than 4 dedicated vCPUs or 16 GB RAM, this criterion was used in the year 2013. The criterion for selecting dedicated hardware is revised every year because of increasing capabilities of hardware and virtualization software.

The organization under study used a mixture of approaches for determining the size of infrastructure in private cloud. Respondents noted that for applications that are migrated from conventional infrastructure to private cloud, historical capacity utilization data of hardware were used for determining the size of new infrastructure. For the

applications purchased from external vendors, the infrastructure specifications provided by the vendor were used for sizing. One important step in sizing was a pre-production performance-testing of applications by simulating the peak workload. One problem of sizing was associated with the agile application development methodology; developers build and deploy applications in iterations that made it difficult to make size estimations. This issue was addressed by designing applications that were horizontally scalable. More virtual servers were added to the application pool when application's resource requirements increased.

Theme 5: Monitoring

Monitoring of infrastructure and application components was a major theme in both interview transcripts and documentations. The physical location of an application hosted in cloud may not be static. Therefore, traditional monitoring of hardware performance may not quickly identify QoS issues with the applications. Respondents discussed about monitoring in response to interview Questions 2, 3, and 7. Respondents mentioned several layers of monitoring; at the top-most layer, monitors measure UX by simulating a web browser which connects to the application periodically, performs a series of critical business transactions, and records response time or errors. The monitoring software records the application performance data in a database and sends alerts when the performance is slower than the established baseline or when errors are encountered.

Another layer of monitoring identified by the respondents was the application-transactions at different tiers of applications. A typical application-transaction uses

multiple tiers such as a web server, application server, database server, and a network-based file system. The organization under study used an application performance management (APM) system for dynamics discovery and the mapping of application transactions in the cloud infrastructure. The APM system was able to do end-to-end monitoring of system; it could identify the processing time and delays at each application tier. This helped application and infrastructure support teams in identifying precisely which tier in cloud was responsible for the QoS issue in the application.

Theme 6: Trending

The trending analysis of operational data was another theme that emerged in the research data. In response to interview Question 7, 8, and 9 respondents emphasized the importance of capturing and analyzing the operational data related to the application performance by establishing baselines and identifying trends in order to proactively identify QoS issues before they start affecting performance of applications in the cloud. In many cases, the performance of application slowly deteriorates due to reasons such as data growth, increase in application usage, and higher utilization of private cloud infrastructure. Information systems managers review the trending data and take corrective actions to prevent future violations of the QoS SLAs.

I reviewed 15 trending reports, including reports on web traffic and the use of system resources (CPU, memory, swap, I/O, storage, and network). Sixty percent of the applications (9 out of 15) had no significant variations in system resource utilization. The system resource utilization increased linearly in 3 applications. For one application, the system resource utilization decreased. There was a sudden increase in system

resource utilizations for 2 applications due to a major software upgrade, as noted in the resource utilization report. According to one respondent, IT team members are developing a report that will show historical information on application response time. The application performance report will contain the performance trends at different application tiers. The application support specialist will review trends on a monthly basis and identify slowness before it becomes a problem. Early identification will help in preventing future QoS issues with the applications.

The six themes that emerged from data analysis helped in exploring the answers to four research questions. I found a many-to-many relationship between themes and research questions. Each theme helped me in exploring the answers to two or more of the research questions. Similarly, I found that each research question was mapped to two or more themes. The answers to four research questions helped me in identifying an enterprise process framework for managing QoS in private cloud.

Research Question 1

The first research question was about the processes for guaranteeing QoS in private cloud environment. The major themes that emerged from the analysis of interview responses and various documents are mapped to a series of processes. The participating organization used these processes for managing QoS in private cloud. These processes are shown in Figure 3. The goal of QoS is to consistently meet the performance requirements of the end user. The first step is to establish SLAs that accurately document the end user's performance requirements. The next step is to architect and develop horizontally scalable applications. Cloud computing offers

dynamic elasticity; this is accomplished by launching new server instances when the workload increases. Load-balancers distribute the workload among available server instances in the cloud. An application designed for a standalone server cannot take advantage of multiple instances and load balancing. The interview respondents emphasized the need of either building or enhancing applications that are horizontally scalable, can take advantage of dynamic elasticity, and consistently meet QoS SLAs.

Proper infrastructure sizing and resource reservation is another step in guaranteeing QoS in cloud. The next step for guaranteeing QoS in cloud is the monitoring of application components. The respondents mentioned about using APM software for mapping and monitoring critical business transactions. This is different from traditional infrastructure because the execution environment for critical transactions is not attached to a specific hardware. The monitoring system should be capable of accurately measuring the execution time of transactions irrespective of its execution environment, detecting the QoS violations, and alerting when established thresholds are exceeded. Most respondents mentioned that the monitoring system should simulate and measure the UX. The ultimate goal of QoS is meeting the end user's application performance needs; the monitoring of UX is an essential step to accomplish this.

I reviewed the last 12 months reports on the application response time; these reports presented QoS trends over time. There were variations in application response times at certain hours of the day or week. I observed that the application response was slow during maintenance activities such as data backups and batch job executions. Also, there were variations in the system resource utilization over the time. Respondents

mentioned that the applications resource requirements may change over time due to factors such as an increase in user population, the growth of data, or increased complexity of application logic. Reports on QoS trends contained the analysis of data collected from the monitoring systems. The application and infrastructure support staff review these reports periodically and take corrective actions to avoid QoS violations.

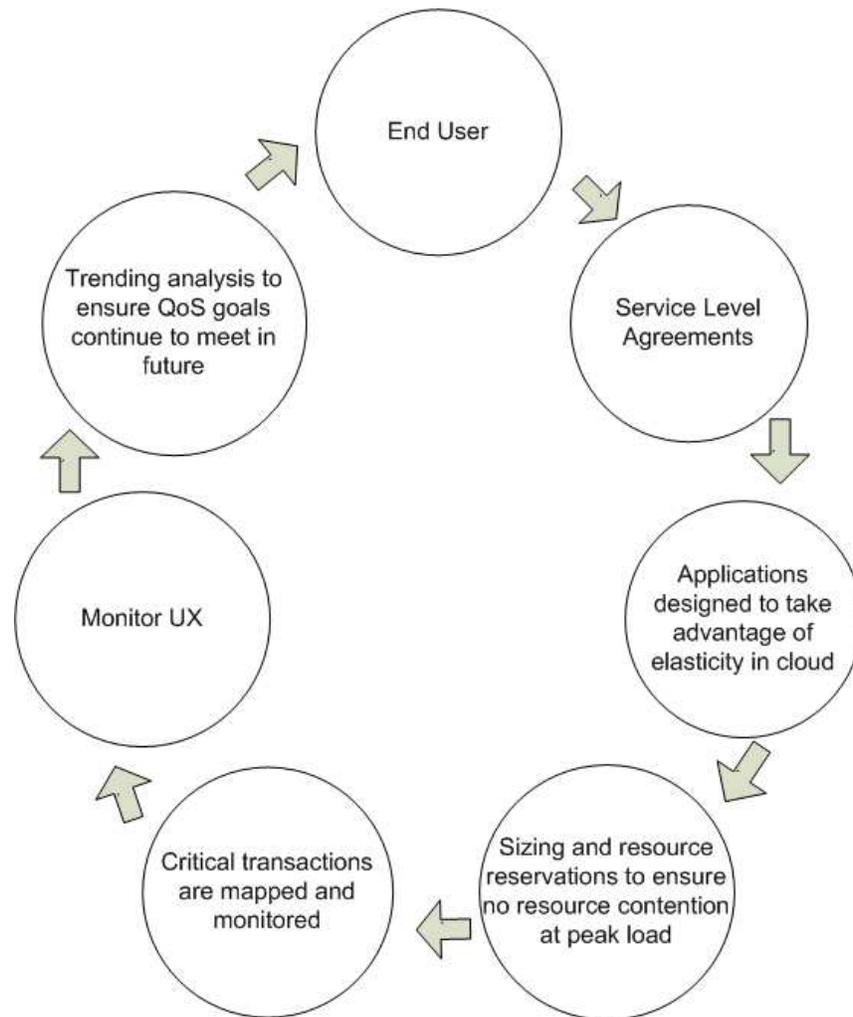


Figure 3. Processes for managing QoS in private cloud.

Research Question 2

The second research question was about identifying QoS requirements of private cloud service users. Three of the identified themes (a) end use expectations, (b) SLA, and (c) sizing were related to this research questions. Private cloud QoS requirements are tied with the end users' application performance expectations. The primary source for the QoS requirement was the SLA document. Processes related to the sizing of the application environment were also useful in identifying QoS requirements.

Research Question 3

The third research question was about proactively identifying QoS issues before they start affecting the application performance. The participating organization used a short-term and a long-term approach to handle this. The short-term approach was application transaction monitoring and the long-term approach was the trending analysis of monitoring data. Monitoring and trending themes are related to this research question.

Research Question 4

The fourth research question was about strategy for the prioritization of computing resources in the private cloud. Based on the analysis of responses categorized under theme sizing, I found four approaches for the prioritization of computing resources based on the application design and workload. Applications that require considerably large computing resources should be installed on a physical server instead of a virtual server. The participating organization had a guideline of using a dedicated server, when the requirements exceeded 4 dedicated CPUs or 16 GB RAM. These guidelines were for

the year 2014 and will be revised when the organization upgrades its hardware and virtualization software.

The respondents mentioned that they keep the size (CPU and memory) of virtual machines small because this policy allows them to deploy a large number of instances on the host servers. Having large instances will restrict the number of instances on the server; it also becomes difficult to dynamically move the images of large virtual machines to other hosts because other hosts will require sufficient free resources to accommodate the large size virtual machines.

Another method for prioritizing the computing resources was resource reservation. Hypervisors have the ability to over-provision computing resources such as CPUs. The concept of over-provisioning is based on the assumption that computing resources on all the virtual machines will not be utilized to their full capacity simultaneously. This assumption may not remain valid in all situations. Resource reservations were used for guaranteeing sufficient computing resources to virtual machines.

The *virtual server build standards* document listed three templates for virtual machines: (a) small, (b) medium, and (c) large. These templates ensured that pre-defined computing resources and storage were assigned to applications. Respondents mentioned that they selected templates based on the QoS requirements of applications. The fourth approach for prioritizing the allocation of system resources was developing applications that are horizontally scalable and deploying on multiple load-balanced virtual machines. The load-balancer intercepts all the incoming traffic and redirects traffic to a virtual

server based on its current workload. Load-balancers were configured to calculate the workload on the virtual instance using a predefined formula and redirected traffic to virtual servers that could meet QoS requirements.

Application to Professional Practice

A cloud-based infrastructure provides an agile application-hosting environment due to rapid provisioning, scalability, and elasticity. The technology behind cloud is complex and the management of QoS in cloud is challenging. The management of QoS in cloud requires ensuring that adequate computing resources are available for the workload and QoS requirements established in SLAs (Kousiouris, Menychtas, Kyriazis, Gogouvitis, & Varvarigou, 2014). By analyzing data collected during this study, I identified a set of processes for managing QoS in private cloud, for the various phases of application lifecycle.

The findings of this study may be used as a guideline by information systems leaders for establishing or validating processes, used in their organizations, for consistently meeting the established application performance SLAs. The process framework identified in this study may help in migration of applications from traditional infrastructure to cloud. This study identified processes used by the participating organization for meeting the end user application performance requirements. This may help in increasing level of customer satisfaction; a high level of customer satisfaction may result in a higher retention rate of customers and an increase in customer population. Information systems managers may use these processes for improving coordination between the application development and infrastructure teams.

Implication for Social Change

The implications for positive social change include the potential to increase the profitability of businesses and support a sustainable environment. An increase in business profitability will make businesses sustainable and make positive change to society. Businesses make a positive contribution to society by increasing wealth of shareholders, providing employment opportunities, and contributing to the government's tax revenue.

The adoption of cloud computing may reduce consumption of electricity due to a reduction in hardware in computer data-centers. This may reduce the emission of greenhouse gases and help in environmental sustainability. Cloud computing helps in the consolidation of IT infrastructure and may reduce the total hardware footprint in data-centers. The hardware reduction will decrease power consumption by computer servers and other related equipment in the data-center, such as network switches, routers, and storage devices. Computer data-centers are maintained at a constant temperature for ensuring reliability and performance of devices. Server and other hardware in data-centers generate heat and require air conditioning for cooling. Hardware reduction may result in additional savings due to the electricity saved in cooling (Graubner et al., 2012).

The adoption of cloud computing reduces hardware requirements in developing nation. This may reduce electronic waste in the developing countries like Ghana (Oteng-Ababio, Amankwaa, & Chama, 2014) and lower the impact to these countries' local environment. Also, a reduction of hardware will result in reduced machines available that can be used to steal personal data off and target.

Recommendations for Action

Recommendations for action include establishing guidelines for IT support managers, application architects, and IT operations managers. Based on the findings of this study, I recommend that IT support managers determine the application performance SLAs in consultation with the end user. Organizations must ensure that applications deployed in cloud provide consistent superior performance to the users. There should be no performance degradation after deploying applications in cloud. Cloud computing offers dynamic elasticity, application architects should design applications that could scale horizontally and take advantage of the elasticity offered by cloud.

The application and infrastructure support managers should present a correct estimate of the system resource requirements. This resource estimation should be based on the delivery of desired application performance at the peak workload. This will help in selecting the right size of instances in cloud and setting up necessary resource reservations. The application resource requirements may change over time due to complexity added during the enhancements and increases in the workload. Periodical evaluation of resource utilization trends by the IT operations managers will help in forecasting and planning future resource requirements.

The performance monitoring of cloud-based applications requires mapping application tiers and monitoring the response time of individual tiers. Some of the participants mentioned that they used APM software for mapping application tiers, monitoring, and trending analysis. Based on the analysis of data, I recommend using the APM software for important applications hosted in cloud.

Recommendations for Further Study

Recommendations for further study include exploring processes used for managing QoS in PaaS and SaaS cloud deployment models. The management of QoS in PaaS and SaaS presents challenges different from the IaaS model (Wu, Garg, & Buyya, 2012); a study of these processes will complement this study. An enterprise process framework for managing QoS for PaaS and SaaS will help organizations in successfully hosting applications in private and public cloud.

In this study, I explored processes used for managing QoS in private cloud. The process framework is a good reference model for managing QoS in private cloud with IaaS deployment model. Two other deployment models PaaS and SaaS, function as additional layers above the IaaS model. Some processes identified in this study will also be applicable to the PaaS and SaaS service. The management of QoS in PaaS and SaaS may require some additional processes and some of the IaaS processes may be enhanced.

Reflections

During this study, I had the opportunity to get insight from IT professionals working with private cloud. I reviewed various documents such as SOPs, technical reports, and guidelines used for managing applications hosted in private cloud infrastructure. This helped me learn about cloud computing service management. This process was rigorous and time consuming, but ultimately rewarding.

I manage web and database infrastructure for a large organization. Some of the web and database servers that I manage are hosted in private cloud. Due to this, I had the experience of working with cloud computing infrastructure. However, I did not design or

build cloud infrastructure. Also, I did not define the processes related to management of cloud infrastructure as part of my job responsibilities. The discovery of QoS management processes in cloud was a good learning experience for me. I remained very objective and reported information only on the basis of data collected. I made a conscious effort to ensure my personal experience and preconceived ideas were not reflected in the study. My background in IT and cloud computing made it easier to understand the terminology and context in which participants spoke. My background also helped me analyze the technical documents collected during research.

The interview participants were enthusiastic during interviews. Many participants even gave me the option of calling them in future for any further questions and clarifications. They were looking forward to seeing results of the study. They were using these processes, but knowledge was not documented in one place. This report has consolidated QoS management-related processes in one place, which may become a useful reference for them.

This research work was different from my work at the place of my employment. In my workplace, I have never collected and analyzed qualitative data of this volume. I have a very good understanding of the research process now. I may use these skills at my place of employment in future projects.

Summary and Study Conclusions

QoS in cloud is a major concern for information systems leaders (Kaur & Chana, 2014). This is a major obstacle in the adoption of cloud computing. Cloud computing offers agility and flexibility to organizations due to its features such as on demand

availability of computational resources, dynamic elasticity, broad network access, and pay-as-you-go cost model. The findings of this study may help information systems leaders in managing QoS in private cloud and successfully adopt cloud computing.

The overarching research question for this study was: How can organizations successfully manage QoS in a private cloud platform? I analyzed data using Weft QDA software and explored processes used by the participating organization for managing QoS in private cloud. Six major themes emerged from the analysis of data: (a) end user expectations, (b) SLAs, (c) application architecture, (d) infrastructure sizing, (e) monitoring, and (f) trending analysis. These themes were aligned to one or more of the research questions. Each theme contained a set of processes, which helped me in documenting the process framework used by the participating organization. I reported the process framework in the findings section of this report.

The focus of QoS should be on meeting the end user's expectation of application response time. The application performance SLAs must reflect customer expectations accurately and clearly. The process framework for management of QoS spans throughout the application lifecycle. The application architects and developers should design and develop applications that are horizontally scalable. This allows applications to utilize dynamics elasticity offered by cloud. Infrastructure architects and engineers should correctly size and use resource reservations based on applications peak workload and performance SLAs. Monitoring setup for applications in cloud should include dynamic mapping of the application component execution environments and response time monitoring at each application-tier. The application performance data and infrastructure

capacity utilization data should be stored in a database. Information systems managers should periodically review and analyze performance and capacity utilization data, and take necessary actions to mitigate risks of future QoS SLA violations.

References

- Ahlstrom, D. (2010). Innovation and growth: How business contributes to society. *Academy of Management Perspectives*, 24, 11–24.
doi:10.5465/AMP.2010.52842948
- Albodour, R., James, A., & Yaacob, N. (2012). High level QoS-driven model for grid applications in a simulated environment. *Future Generation Computer Systems*, 28, 1133–1144. doi:10.1016/j.future.2011.06.013
- Aljabre, A. (2012). Cloud computing for increased business value. *International Journal of Business & Social Science*, 3, 234–239. Retrieved from <http://www.ijbssnet.com/>
- Amazon.com Inc. (2013). *Amazon Elastic Compute Cloud (Amazon EC2)*. Retrieved from <http://aws.amazon.com/ec2/>
- Angeli, D., & Masala, E. (2012). A cost-effective cloud computing framework for accelerating multimedia communication simulations. *Journal of Parallel and Distributed Computing*, 72, 1373–1385. doi:10.1016/j.jpdc.2012.06.005
- Arghode, V. (2012). Qualitative and quantitative research: Paradigmatic differences. *Global Education Journal*, 2012, 155–163. Retrieved from <http://www.franklinpublishing.net/globaleducation.html>
- Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, . . . Zaharia, M. (2010). A view of cloud computing. *Communication of the ACM*, 53(4), 50–58.
doi:10.1145/1721654.1721672

- Avetisyan, A. I., Campbell, R., Gupta, I., Heath, M. T., Ko, S. Y., Ganger, G. R., . . . Kwan, T. (2010). Open Cirrus: A global cloud computing testbed. *Computer (New York)*, 43(4), 35–43. doi:10.1109/MC.2010.111
- Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Generation Computer Systems*, 28, 755–768. doi:10.1016/j.future.2011.04.017
- Bernardes, E. S., & Hanna, M. D. (2009). A theoretical review of flexibility, agility and responsiveness in the operations management literature: Toward a conceptual definition of customer responsiveness. *International Journal of Operations & Production Management*, 29, 30–53. doi:10.1108/01443570910925352
- Bhatt, G., Emdad, A., Roberts, N., & Grover, V. (2010). Building and leveraging information in dynamic environments: The role of IT infrastructure flexibility as enabler of organizational responsiveness and competitive advantage. *Information & Management*, 47, 341–349. doi:10.1016/j.im.2010.08.001
- Boniface, M., Nasser, B., Papay, J., Phillips, S. C., Servin, A., Yang, X., . . . Kyriazis, D. (2010). *Platform-as-a-service architecture for real-time quality of service management in Clouds*. Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services (ICIW), Barcelona, 2010, 155-160. doi:10.1109/ICIW.2010.91

- Bossche, R. V. den, Vanmechelen, K., & Broeckhove, J. (2013). Online cost-efficient scheduling of deadline-constrained workloads on hybrid clouds. *Future Generation Computer Systems*, *29*, 973–985. doi:10.1016/j.future.2012.12.012
- Brandic, I., Raicu, I., Jackson, K. R., Muriki, K., Ramakrishnan, L., Runge, K. J., & Thomas, R. C. (2011). Performance and cost analysis of the Supernova factory on the Amazon AWS cloud. *Scientific Programming*, *19*, 107–119.
doi:10.3233/SPR-2011-0324
- Bryman, A. (2012). *Social research methods*. Oxford; U.K.: Oxford University Press.
- Budrienė, D., & Zalieckaitė, L. (2012). Cloud computing application in small and medium-sized enterprises. *Issues of Business & Law*, *4*, 199–130.
doi:10.520/ibl.2012.11
- Calheiros, R. N., Toosi, A. N., Vecchiola, C., & Buyya, R. (2012). A coordinator for scaling elastic applications across multiple clouds. *Future Generation Computer Systems*, *28*, 1350–1362. doi:10.1016/j.future.2012.03.010
- Calheiros, R. N., Vecchiola, C., Karunamoorthy, D., & Buyya, R. (2012). The Aneka platform and QoS-driven resource provisioning for elastic applications on hybrid Clouds. *Future Generation Computer Systems*, *28*, 861–870.
doi:10.1016/j.future.2011.07.005
- Carrillo de Gea, J. M., Nicolás, J., Fernández Alemán, J. L., Toval, A., Ebert, C., & Vizcaíno, A. (2012). Requirements engineering tools: Capabilities, survey and assessment. *Information and Software Technology*, *54*, 1142–1157.
doi:10.1016/j.infsof.2012.04.005

- Chang, V., Walters, R. J., & Wills, G. (2013). The development that leads to the Cloud Computing Business Framework. *International Journal of Information Management, 33*, 524–538. doi:10.1016/j.ijinfomgt.2013.01.005
- Che, J., Duan, Y., Zhang, T., & Fan, J. (2011). Study on the security models and strategies of cloud computing. *Procedia Engineering, 23*, 586–593. doi:10.1016/j.proeng.2011.11.2551
- Christensen, C. M. (2011). *The innovator's dilemma: The revolutionary book that will change the way you do Business*. New York, NY: HarperBusiness.
- Chung, L., Hill, T., Legunsen, O., Sun, Z., Dsouza, A., & Supakkul, S. (2013). A goal-oriented simulation approach for obtaining good private cloud-based system architectures. *Journal of Systems and Software, 86*, 2242–2262. doi:10.1016/j.jss.2012.10.028
- Ciznicki, M., Kierzyńska, M., Kopta, P., Kurowski, K., & Gepner, P. (2012). Benchmarking data and compute intensive applications on modern CPU and GPU architectures. *Procedia Computer Science, 9*, 1900–1909. doi:10.1016/j.procs.2012.04.208
- Cooper, D., & Schindler, P. (2010). *Business research methods*. New York, NY: McGraw-Hill/Irwin.
- Crockett, D. R., McGee, J. E., & Payne, G. T. (2013). Employing new business divisions to exploit disruptive innovations: the interplay between characteristics of the corporation and those of the venture management team. *Journal of Product Innovation Management, 30*, 856–879. doi:10.1111/jpim.12034

- Curino, C. A., Jones, E. P. C., Popa, R. A., Malviya, N., Wu, E., Madden, S. R., ...
Zeldovic, N. (2011). *Relational cloud: A database-as-a-service for the cloud*.
Retrieved from Massachusetts Institute of Technology. Dept. of Electrical
Engineering and Computer Science website: <http://hdl.handle.net/1721.1/62241>
- Cusumano, M. (2010). Cloud computing and SaaS as new computing platforms.
Communications of the ACM, 53(4), 27–29. doi:10.1145/1721654.1721667
- Daniels, J. (2011). *Assured Identity for the cloud* (Doctoral dissertation). Retrieved from
<http://hdl.handle.net/10484/2031>.
- Dawes, N. P., & Larson, R. (2011). How youth get engaged: Grounded-theory research
on motivational development in organized youth programs. *Developmental
Psychology*, 47, 259–269. doi:10.1037/a0020729
- Diffin, K. (2011). *Clinician competency in treating transgender individuals*. (Doctoral
Dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No.
3439917).
- Dikaiakos, M. D., Katsaros, D., Mehra, P., Pallis, G., & Vakali, A. (2009). Cloud
computing: Distributed Internet computing for IT and scientific research. *IEEE
Internet Computing*, 13, 10–13. doi:10.1109/MIC.2009.103
- Doelitzscher, F., Sulistio, A., Reich, C., Kuijs, H., & Wolf, D. (2011). Private cloud for
collaboration and e-Learning services: From IaaS to SaaS. *Computing*, 91, 23–42.
doi:10.1007/s00607-010-0106-z
- Durkee, D. (2010). Why cloud computing will never be free. *Communication of the
ACM*, 53(5), 62–69. doi:10.1145/1735223.1735242

- Dwivedi, Y. K., & Mustafee, N. (2010). It's unwritten in the Cloud: the technology enablers for realising the promise of Cloud Computing. *Journal of Enterprise Information Management*, 23, 673–679. doi:10.1108/17410391011088583
- Erdogmus, H. (2009). Cloud computing: Does nirvana hide behind the Nebula? *IEEE Software*, 26(2), 4–6. doi:10.1109/MS.2009.31
- Espadas, J., Molina, A., Jiménez, G., Molina, M., Ramírez, R., & Concha, D. (2013). A tenant-based resource allocation model for scaling Software-as-a-Service applications over cloud computing infrastructures. *Future Generation Computer Systems*, 29, 273–286. doi:10.1016/j.future.2011.10.013
- Faniyi, F., Bahsoon, R., & Theodoropoulos, G. (2012). A dynamic data-driven simulation approach for preventing service level agreement violations in cloud federation. *Procedia Computer Science*, 9, 1167–1176. doi:10.1016/j.procs.2012.04.126
- Fernando, N., Loke, S. W., & Rahayu, W. (2013). Mobile cloud computing: A survey. *Future Generation Computer Systems*, 29, 84–106. doi:10.1016/j.future.2012.05.023
- Ferrer, A. J., Hernández, F., Tordsson, J., Elmroth, E., Ali-Eldin, A., Zsigri, C., Sirvent, R., et al. (2012). OPTIMIS: A holistic approach to cloud service provisioning. *Future Generation Computer Systems*, 28, 66–77. doi:10.1016/j.future.2011.05.022
- Gagnon, A. J., Carnevale, F., Mehta, P., Rousseau, H., & Stewart, D. E. (2013). Developing population interventions with migrant women for maternal-child

health: a focused ethnography. *BMC Public Health*, *13*, 1–14. doi:10.1186/1471-2458-13-471

García, A., Espert, I., & García, V. (2014). SLA-driven dynamic cloud resource management. *Future Generation Computer Systems*, *31*, 1–11. doi:10.1016/j.future.2013.10.005

Garg, S. K., Versteeg, S., & Buyya, R. (2013). A framework for ranking of cloud computing services. *Future Generation Computer Systems*, *29*, 1012–1023. doi:10.1016/j.future.2012.06.006

Garrison, G., Kim, S., & Wakefield, R. L. (2012). Success factors for deploying cloud computing. *Communications of the ACM*, *55*(9), 62–68. doi:10.1145/2330667.2330685

Gibbs, G. R. (2008). *Analysing qualitative data*. Thousand Oaks, CA: SAGE.

Gong, Y., & Janssen, M. (2012). From policy implementation to business process management: Principles for creating flexibility and agility. *Government Information Quarterly*, *29*, 61–71. doi:10.1016/j.giq.2011.08.004

Goscinski, A., & Brock, M. (2010). Toward dynamic and attribute based publication, discovery and selection for cloud computing. *Future Generation Computer Systems*, *26*, 947–970. doi:10.1016/j.future.2010.03.009

Graubner, P., Schmidt, M., & Freisleben, B. (2012). Energy-efficient Virtual Machine Consolidation for Cloud Computing. *IT Professional*, *15*(2), 18-34. doi:10.1109/MITP.2012.48

- Grobauer, B., Walloschek, T., & Stocker, E. (2011). Understanding cloud computing vulnerabilities. *IEEE Security Privacy*, 9(2), 50–57. doi:10.1109/MSP.2010.115
- Gupta, P., Seetharaman, A., & Raj, J. R. (2013). The usage and adoption of cloud computing by small and medium businesses. *International Journal of Information Management*, 33, 861–874. doi:10.1016/j.ijinfomgt.2013.07.001
- Han, Y. (2010). On the clouds: a new way of computing. *Information Technology & Libraries*, 29, 87–92. Retrieved from <http://www.ala.org>
- Han, Y. (2011). Cloud computing: Case studies and total costs of ownership. *Information Technology & Libraries*, 30, 198–206. Retrieved from <http://www.ala.org>
- Harmon, R. R., & Demirkan, H. (2011). The next wave of sustainable IT. *IT Professional*, 13(1), 19–25. doi:10.1109/MITP.2010.140
- Henfridsson, O., & Bygstad, B. (2013). The generative mechanisms of digital infrastructure evolution. *MIS Quarterly*, 37, 907–930. Retrieved from <http://www.misq.org/>
- Hennick, M., Hutter, I., & Bailey, A. (2011). *Qualitative research methods*. Thousand Oaks, CA: Sage.
- Hexiao, H., Shiming, Z., & Haijian, C. (2012). Reengineering from tradition to cloud: A case study. *Procedia Engineering*, 29, 2638–2643. doi:10.1016/j.proeng.2012.01.364
- Hofmann, P., & Woods, D. (2010). Cloud computing: The limits of public clouds for business applications. *IEEE Internet Computing*, 14(6), 90–93. doi:10.1109/MIC.2010.136

- Horsewood, N. (2011). Demystifying quantitative methods in comparative housing research: dispelling the myth of black magic. *International Journal of Housing Policy*, *11*, 375–393. doi:10.1080/14616718.2011.626601
- Houghton, C., Casey, D., Shaw, D., & Murphy, K. (2013). Rigour in qualitative case-study research. *Nurse Researcher*, *20*(4), 12–17.
doi:10.7748/nr2013.03.20.4.12.e326
- Huang, C., Guan, C.T., Chen, H.M., Wang, Y.W., Chang, S.C., Li, C.Y., & Weng, C.H. (2013). An adaptive resource management scheme in cloud computing. *Engineering Applications of Artificial Intelligence*, *26*, 382–389.
doi:10.1016/j.engappai.2012.10.004
- Huang, Q., Yang, C., Liu, K., Xia, J., Xu, C., Li, J., ... Li, Z. (2013). Evaluating open-source cloud computing solutions for geosciences. *Computers & Geosciences*, *59*, 41–52. doi:10.1016/j.cageo.2013.05.001
- Kaur, P. D., & Chana, I. (2014). A resource elasticity framework for QoS-aware execution of cloud applications. *Future Generation Computer Systems*, *37*, 14–25.
doi:10.1016/j.future.2014.02.018
- Khajeh-Hosseini, A., Greenwood, D., Smith, J. W., & Sommerville, I. (2012). The cloud adoption toolkit: Supporting cloud adoption decisions in the enterprise. *Software: Practice and Experience*, *42*, 447–465. doi:10.1002/spe.1072
- Khorshed, M. T., Ali, A. B. M. S., & Wasimi, S. A. (2012). A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud

computing. *Future Generation Computer Systems*, 28, 833–851.

doi:10.1016/j.future.2012.01.006

Kipkulei, K. (2013). *Effects of information technology on reducing perishable waste in supermarkets* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3560427)

Kourtesis, D., Alvarez-Rodríguez, J. M., & Paraskakis, I. (2014). Semantic-based QoS management in cloud systems: Current status and future challenges. *Future Generation Computer Systems*, 32, 307–323. doi:10.1016/j.future.2013.10.015

Kousiouris, G., Menychtas, A., Kyriazis, D., Gogouvitis, S., & Varvarigou, T. (2014). Dynamic, behavioral-based estimation of resource provisioning based on high-level application terms in Cloud platforms. *Future Generation Computer Systems*, 32, 27–40. doi:10.1016/j.future.2012.05.009

Kuo, A. M.-H. (2011). Opportunities and challenges of cloud computing to improve health care services. *Journal of Medical Internet Research*, 13, e67. doi:10.2196/jmir.1867

Laili, Y., Tao, F., Zhang, L., Cheng, Y., Luo, Y., & Sarker, B. R. (2013). A ranking chaos algorithm for dual scheduling of cloud service and computing resource in private cloud. *Computers in Industry*, 64, 448–463. doi:10.1016/j.compind.2013.02.008

Lawrance, H., & Silas, S. (2013). Efficient Qos based resource scheduling using PAPRIKA method for cloud computing. *International Journal of Engineering Science & Technology*, 5, 638–643. Retrieved from <http://www.ijest.info/>

- Leavitt, N. (2009). Is cloud computing really ready for prime time? *Computer (New York)*, *42(1)*, 15–20. doi:10.1109/MC.2009.20
- Leedy, P. D., & Ormrod, J. E. (2009). *Practical research: planning and design* (9th Ed.). Boston, MA: Pearson.
- Lefèvre, L., & Orgerie, A. C. (2010). Designing and evaluating an energy efficient Cloud. *Journal of Supercomputing*, *51*, 352–373. doi:10.1007/s11227-010-0414-2
- Lengnick-Hall, C. A., & Wolff, J. A. (1999). Similarities and contradictions in the core logic of three strategy research streams. *Strategic Management Journal*, *20*, 1109–1132. doi:10.1002/(SICI)1097-0266(199912)20:12<1109::AID-SMJ65>3.0.CO;2-8
- Li, Z., Zhang, H., O'Brien, L., Cai, R., & Flint, S. (2013). On evaluating commercial Cloud services: A systematic review. *Journal of Systems and Software*, *86*, 2371–2393. doi:10.1016/j.jss.2013.04.021
- Lin, Q., Qi, Z., Wu, J., Dong, Y., & Guan, H. (2012). Optimizing virtual machines using hybrid virtualization. *Journal of Systems and Software*, *85*, 2593–2603. doi:10.1016/j.jss.2012.05.093
- Liu, K., & Dong, L. (2012). Research on cloud data storage technology and its architecture implementation. *Procedia Engineering*, *29*, 133–137. doi:10.1016/j.proeng.2011.12.682
- Liu, T., Lu, T., Wang, W., Wang, Q., Liu, Z., Gu, N., & Ding, X. (2012). SDMS-O: A service deployment management system for optimization in clouds while

- guaranteeing users' QoS requirements. *Future Generation Computer Systems*, 28, 1100–1109. doi:10.1016/j.future.2011.10.015
- Lopez, R. H. (2012). *Information data security specialists' and business leaders' experiences regarding communication challenges*. Retrieved from ProQuest Dissertations and Theses database. (UMI NO. 3503982)
- Low, C., Chen, Y., & Wu, M. (2011). Understanding the determinants of cloud computing adoption. *Industrial Management & Data Systems*, 111, 1006–1023. doi:10.1108/02635571111161262
- Lu, Y., & Ramamurthy, K. (2011). Understanding the link between Information Technology capability and organizational agility: an empirical examination. *MIS Quarterly*, 35, 931–954. Retrieved from <http://www.misq.org/>
- Luftman, J., Zadeh, H. S., Derksen, B., Santana, M., Rigoni, E. H., & Huang, Z. (2012). Key information technology and management issues 2011-2012: an international study. *Journal of Information Technology*, 27, 198–212. doi:10.1057/jit.2012.14
- Madhavaiah, C., Bashir, I., & Shafi, S. I. (2012). Defining cloud computing in business perspective: A review of research. *Vision*, 16, 163–173. doi:10.1177/0972262912460153
- Mao, H., Xiao, N., Shi, W., & Lu, Y. (2012). Wukong: A cloud-oriented file service for mobile Internet devices. *Journal of Parallel and Distributed Computing*, 72, 171–184. doi:10.1016/j.jpdc.2011.10.017
- Marshall, C., & Rossman, G. B. (2010). *Designing qualitative research* (5th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

- Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing — The business perspective. *Decision Support Systems*, *51*, 176–189. doi:10.1016/j.dss.2010.12.006
- Mauch, V., Kunze, M., & Hillenbrand, M. (2013). High performance cloud computing. *Future Generation Computer Systems*, *29*, 1408–1416. doi:10.1016/j.future.2012.03.011
- Maurer, M., Brandic, I., & Sakellariou, R. (2013). Adaptive resource configuration for Cloud infrastructure management. *Future Generation Computer Systems*, *29*, 472–487. doi:10.1016/j.future.2012.07.004
- McCoy-Wilson, K. (2011). *Teacher perceptions of job satisfaction in an economically depressed rural school district*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3443645).
- Milojicic, D., & Wolski, R. (2011). Eucalyptus: Delivering a private cloud. *Computer (New York)*, *44(4)*, 102–104. doi:10.1109/MC.2011.109
- Mondal, B., Dasgupta, K., & Dutta, P. (2012). Load balancing in cloud computing using stochastic hill climbing - a soft computing approach. *Procedia Technology*, *4*, 783–789. doi:10.1016/j.protcy.2012.05.128
- Montero, R. S., Moreno-Vozmediano, R., & Llorente, I. M. (2011). An elasticity model for high throughput computing clusters. *Journal of Parallel and Distributed Computing*, *71*, 750–757. doi:10.1016/j.jpdc.2010.05.005
- Murphy, F. J., & Yelder, J. (2010). Establishing rigour in qualitative radiography research. *Radiography*, *16(1)*, 62–67. doi:10.1016/j.radi.2009.07.003

- Narasimhan, B., & Nichols, R. (2011). State of cloud applications and platforms: The cloud adopters' view. *Computer (New York)*, *44*(3), 24–28.
doi:10.1109/MC.2011.66
- National Institute of Standards & Technology. (2012). *Cloud computing synopsis and recommendations* (NIST Publication No. SP 800-146). Retrieved from National Institute of Standards & Technology website: <http://csrc.nist.gov/>
- Norlyk, A., & Harder, I. (2010). What makes a phenomenological study phenomenological? An analysis of peer-reviewed empirical nursing studies. *Qualitative Health Research*, *20*, 420–431. doi:10.1177/1049732309357435
- Oteng-Ababio, M., Amankwaa, E. F., & Chama, M. A. (2014). The local contours of scavenging for e-waste and higher-valued constituent parts in Accra, Ghana. *Habitat International*, *43*, 163–171. doi:10.1016/j.habitatint.2014.03.003
- Pal, S., & Pattnaik, P. K. (2013). Classification of Virtualization Environment for Cloud Computing. *Indian Journal of Science & Technology*, *6*, 3965–3971. Retrieved from <http://www.indjst.org/>
- Pallis, G. (2010). Cloud Computing: The new frontier of Internet computing. *IEEE Internet Computing*, *14*(5), 70–73. doi:10.1109/MIC.2010.113
- Patton, M. (2001). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.
- Polkinghorne, D. E. (1989). Phenomenological research methods. In R. S. Valle & S. Halling (Eds.), *Existential-phenomenological perspectives in psychology:*

Exploring the breadth of human experience (pp. 41–60). New York, NY: Plenum Press.

- Qi, L., Dou, W., Zhang, X., & Chen, J. (2012). A QoS-aware composition method supporting cross-platform service invocation in cloud environment. *Journal of Computer and System Sciences*, 78, 1316–1329. doi:10.1016/j.jcss.2011.12.016
- Quarati, A., Clematis, A., Galizia, A., & D'Agostino, D. (2013). Hybrid clouds brokering: Business opportunities, QoS and energy-saving issues. *Simulation Modelling Practice and Theory* 39, 121-134. doi:10.1016/j.simpat.2013.01.004
- Rogers, O., & Cliff, D. (2012). A financial brokerage model for cloud computing. *Journal of Cloud Computing*, 1, 2-12. doi:10.1186/2192-113X-1-2
- Ross, V. W. (2010). *Factors influencing the adoption of cloud computing by decision making managers* (Doctoral dissertation). Retrieved from <http://gradworks.umi.com/33/91/3391308.html>
- Ruiz-López, T., Noguera, M., Rodríguez, M. J., Garrido, J. L., & Chung, L. (2013). REUBI: A Requirements Engineering method for ubiquitous systems. *Science of Computer Programming*, 78, 1895–1911. doi:10.1016/j.scico.2012.07.021
- Ryan, M. D. (2011). Cloud computing privacy concerns on our doorstep. *Communication of the ACM*, 54(1), 36–38. doi:10.1145/1866739.1866751
- Scott, L. G. (2012). *Influence of provider perception of health literacy on 30 day readmission rates of heart failure patients*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3545581).

- Shaw, B. T. (2012). Exploring *the factors of an enterprise resource planning system in a local government organization*. (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3510463).
- Shin, D. (2013). User centric cloud service model in public sectors: policy implications of cloud services. *Government Information Quarterly*, 30, 194–203.
doi:10.1016/j.giq.2012.06.012
- Singh, S., & Jangwal, T. (2012). Cost breakdown of public cloud computing and private cloud computing and security issues. *International Journal of Computer Science and Information Technology*, 4(2), 17–31. doi:10.5121/ijcsit.2012.4202
- Skiba, D. J. (2011). Are you computing in the clouds? Understanding cloud computing. *Nursing Education Perspectives*, 32, 266–268. doi:10.5480/1536-5026-32.4.266
- Škrinjar, R., & Trkman, P. (2013). Increasing process orientation with business process management: Critical practices. *International Journal of Information Management*, 33, 48–60. doi:10.1016/j.ijinfomgt.2012.05.011
- Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, 34, 1–11. doi:10.1016/j.jnca.2010.07.006
- Subramanian, B. (2012). The disruptive influence of cloud computing and its implications for adoption in the pharmaceutical and life sciences industry. *Journal of Medical Marketing*, 12, 192-203. doi:10.1177/1745790412450171

- Sultan, N. (2013). Knowledge management in the age of cloud computing and Web 2.0: Experiencing the power of disruptive innovations. *International Journal of Information Management*, 33, 160–165. doi:10.1016/j.ijinfomgt.2012.08.006
- Sun, D., Chang, G., Sun, L., & Wang, X. (2011). Surveying and analyzing security, privacy and trust Issues in cloud computing Environments. *Procedia Engineering*, 15, 2852–2856. doi:10.1016/j.proeng.2011.08.537
- Sunyaev, A., & Schneider, S. (2013). Cloud services certification. *Communications of the ACM*, 56(2), 33–36. doi:10.1145/2408776.2408789
- Takabi, H., Joshi, J. B. D., & Ahn, G. (2010). Security and privacy challenges in cloud computing environments. *IEEE Security & Privacy*, 8(6), 24–31. doi:10.1109/MSP.2010.186
- Tolosana-Calasanz, R., Bañares, J. Á., Pham, C., & Rana, O. F. (2012). Enforcing QoS in scientific workflow systems enacted over cloud infrastructures. *Journal of Computer and System Sciences*, 78, 1300–1315. doi:10.1016/j.jcss.2011.12.015
- Ullah, A., & Lai, R. (2011). Modeling business goal for business/IT alignment using requirements engineering. *Journal of Computer Information Systems*, 51(3), 21–28.
- Ullah, A., & Lai, R. (2013). A systematic review of business and information technology alignment. *ACM Transactions on Management Information Systems*, 4(1), 1–30. doi:10.1145/2445560.2445564

- VanderStoep, S. W., & Johnson, D. D. (2009). *Research methods for everyday life: Blending qualitative and quantitative approaches* (1st ed.). San Francisco, CA: Jossey-Bass.
- Vázquez, C., Huedo, E., Montero, R. S., & Llorente, I. M. (2011). On the use of clouds for grid resource provisioning. *Future Generation Computer Systems*, 27, 600–605. doi:10.1016/j.future.2010.10.003
- Wang, N., Liang, H., Zhong, W., Xue, Y., & Xiao, J. (2012). Resource structuring or capability building? An empirical study of the business value of Information Technology. *Journal of Management Information Systems*, 29, 325–367.
- Willig, C. (2008). *Introducing qualitative research in psychology*. New York, NY: McGraw-Hill Education.
- Wu, L., Garg, S., & Buyya, R. (2012). SLA-based admission control for a Software-as-a-Service provider in Cloud computing environments. *Journal of Computer and System Sciences*, 78, 1280–1299. doi:10.1016/j.jcss.2011.12.014
- Yang, J., Yu, T., Jian, L. R., Qiu, J., & Li, Y. (2011). An extreme automation framework for scaling cloud applications. *IBM Journal of Research and Development*, 55(8), 1–12. doi:10.1147/JRD.2011.2170929
- Yin, R. K. (2009). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.