Grid and cloud computing: Technologies, applications, market sectors, and workloads.

Altowaijri, Saleh

Grid and Cloud Computing: Technologies, Applications, Market Sectors, and workloads

Saleh Altowaijri

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

College of Engineering
Swansea University
2012

Abstract

Developments in electronics, computing and communication technologies have transformed IT systems from desktop and tightly coupled mainframe computers of the past to modern day highly complex distributed systems. These ICT systems interact with humans at a much advanced level than what was envisaged during the early years of computer development. The ICT systems of today have gone through various phases of developments by absorbing intermediate and modern day concepts such as networked computing, utility, on demand and autonomic computing, virtualisation and so on. We now live in a ubiquitous computing and digital economy era where computing systems have penetrated into the human lives to a degree where these systems are becoming invisible. The price of these developments is in the increased costs, higher risks and higher complexity. There is a compelling need to study these emerging systems, their applications, and the emerging market sectors that they are penetrating into.

Motivated by the challenges and opportunities offered by the modern day ICT technologies, we aim in this thesis to explore the major technological developments that have happened in the ICT systems during this century with a focus on developing techniques to manage applied ICT systems in digital economy. In the process, we wish to also touch on the evolution of ICT systems and discuss these in context of the state of the art technologies and applications. We have identified the two most transformative technologies of this century, grid computing and cloud computing, and two application areas, intelligent healthcare and transportation systems.

The contribution of this thesis is multidisciplinary in four broad areas. Firstly, a workload model of a grid-based ICT system in the healthcare sector is proposed and analysed using multiple healthcare organisations and applications. Secondly, an innovative intelligent system for the management of disasters in urban environments using cloud computing is proposed and analysed. Thirdly, cloud computing market sectors, applications, and workload are analysed using over 200 real life case studies. Fourthly, a detailed background and literature review is provided on grid computing and cloud computing. Finally, directions for future work are given. The work contributes in multidisciplinary fields involving healthcare, transportation, mobile computing, vehicular networking, grid, cloud, and distributed computing.

The discussions presented in this thesis on the historical developments, technology and architectural details of grid computing have served to understand as to how and why grid computing was seen in the past as the global infrastructure of the future. These discussions on grid computing also provided the basis that we subsequently used to explain the background, motivations, technological details, and ongoing developments in cloud computing. The introductory chapters on grid and cloud computing, collectively, have provided an insight into the evolution of ICT systems over the last 50+ years - from mainframes to microcomputers, internet, distributed computing, cluster computing, and computing as a utility and service. The existing and proposed applications of grid and cloud computing in healthcare and transport were used to further elaborate the two technologies and the ongoing ICT developments in the digital economy. The workload models and analyses of grid and cloud computing systems can be used by the practitioners for the design and resource management of ICT systems.

Declarations

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ...07/08/2013......................

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date ...07/08/2013................

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ...07/08/2013................

Table of Contents

Acknowledgements

Writing this PhD thesis would not have been possible without the help, support and encouragement of the kind people around me, to only some of whom it is possible to give particular acknowledgment here.

First and foremost, I am sincerely grateful to my PhD supervisor, Dr. Rashid Mehmood for his patient guidance, encouragement, valuable directions toward this accomplishment, and advice that he has provided throughout my time as his student. I have been fortunate enough to have him as a supervisor who cared so much about my PhD work, and who responded to my questions and queries so promptly. Dr Rashid Mehmood unfortunately moved from Swansea University when I was towards the end of my PhD.

I also would like to thank my current PhD supervisor, Dr Chris Jobling for his valuable input, support, comments, and feedback.

I would like to thank John Williams, Professor of Health Services Research at the College of Medicine, Swansea University, for the collaboration with him which led us to develop a grid-based healthcare system model. This work was published by IEEE Computer Society and formed part of a chapter in this thesis, Chapter 3.

I would also like to thanks Ms Zubaida Alazawi, Dr. Mohmmad Abdljabar, for the collaborations which led us to develop the novel work on intelligent disaster management system. This work has contributed to three peer reviewed papers and forms part of a chapter in this thesis, Chapter 6.

I would also like to thank the external examiner Dr. Behzad Bordbar and the internal examiner Dr. Joseph Kim for their Constructive criticisms and helpful comments that have helped in improving the quality of the thesis.

Endless thanks go to my mother for her encouragement and continuous prayers, my brothers and sisters for their encouragement and support.

Finally, I would like to acknowledge the Ministry of Higher Education in Saudi Arabia, particularly in the award of a Postgraduate Research Scholarship that provided the financial support for this research.

Table of Figures

List of Tables

Abbreviations

AFR	Annual Failure Rate
Amazon EBS	Amazon Elastic Block Store
Amazon EC2	Amazon Elastic Compute Cloud
Amazon ELB	Amazon Elastic Load Balancing
Amazon EMR	Amazon Elastic MapReduce
Amazon FPS	Amazon Flexible Payments Service
Amazon RDS	Amazon Relational Database Service
Amazon S3	Amazon Simple Storage Service
Amazon SDB	Amazon SimpleDB
Amazon SQS	Amazon Simple Queue Service
Amazon VPC	Amazon Virtual Private Cloud
AMT	Amazon Mechanical Turk
APIs	Application Programming Interfaces
ASPs	Application Service Providers
AWS	Amazon Web Services
BPaaS	Business Process as a Service
BSS	Business Support Services
C2C	Car-to-Car
C2I	Car-to-Infrastructure
CCMP	Common Cloud Management Platform
CCRA	Cloud Computing Reference Architecture
CDS	Clinical Decision Support
CORBA	Common Object Request Broker Architecture
CPU	Central Processing Unit
CT	Computerized Tomography
CTMC	Continuous Time Markov Chain
CVS	Concurrent Versions System
DARPA	Defense Advanced Research Projects Agency
DCOM	Distributed Component Object Model
DHS	Department of Homeland Security
DMTF	Distributed Management Task Force

DNS	Domain Name System
DoE	Department of Energy
DoT	department of transportation
DRS	Data Replication Service
EGEE	Enabling Grids for E-science
EHR	Electronic Health Records
EJB	Enterprise Java Beans
ESG	Earth System Grid
ETO	Emergency Transportation Operations
FHWA	Federal Highway Administration
FISMA	Federal Information Security Management Act
FTA	Federal Transit Administration
FTP	File Transfer Protocol
GADU	Genome Analysis Database Update
GPS	Global Positioning System
GRAM	Grid Resource Access and Management
GRIP	Grid Resource Information Protocol
GriPhyN	Grid Physics Network
GRIS	Grid Resource Information Service
GRRP	Grid Resource Registration Protocol
GSI	Grid Security Infrastructure
GT	Globus Toolkit
GTCP	Grid TeleControl Protocol
HCEP	Hydrosult Center for Engineering Planning
HIPAA	Health Insurance Portability and Accountability Act
HIS	Healthcare Information Systems
HIT	Human Intelligence Task
HPC	High Performance Computing
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
ICT	Information and Communication Technologies
IDC	International Data Corporation
IETF	Internet Engineering Task Force

IP	Internet Protocol
ISO	International Organization for Standardization
ISPs	Internet Service Providers
IT	Information Technology
ITS	Intelligent Transport Systems
J2EE	Java 2 Enterprise Edition
LDAP	Lightweight Directory Access Protocol
LIGO	Laser Interferometer Gravitational-Wave Observatory
LWR	Lighthill-Whitham-Richards
MANET	Mobile Ad Hoc Networks
MMPW	Ministry of Municipalities and Public Works
NASA	National Aeronautics and Space Administration
NHS	National Health Service
NHTSA	National Highway Traffic Safety Administration
NIMS	National Incident Management System
NIST	National Institute of Standards and Technology
NSF	National Science Foundation
OASIS	Organization for the Advancement of Structured Information Standards
O-D	Origin-Destination
OGSA	Open Grid Services Architecture
OGSA-DAI	Open Grid Services Architecture Data Access and Integration
ORB	Object Request Broker
OSS	Operational Support Services
PaaS	Platform as a Service
PCI DSS	Payment Card Industry Data Security Standard
PDAs	Portable Digital Assistants
PI	Personal Information
PII	Personally Identifiable Information
PMRM	Privacy Management Reference Model
POA	Portable Object Adaptor
PSEs	Problem Solving Environments
QoS	Quality of Service

RA	Reference Architecture
RFIDs	Radio Frequency Identification Devices
RFT	Reliable File Transfer
RLS	Replica Location Service
RMS	Resource Management System
RoI	Return on Investment
SaaS	Software as a Service
SAS	Statement on Auditing Standards
SCM	Source Configuration Management
SDKs	Software Development Kits
SLAs	Service-Level Agreement
SMA	Simulation Modelling, and Analysis
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SSL	Secure Sockets Layer
SSPs	Storage Service Providers
TCP	Transmission Control Protocol
TIM	Traffic Incident Management
TLS	Transport Layer Security
TRB	Transportation Research Board
VDS	Virtual Data System
VLANs	Virtual Local Area Networks
VO	Virtual Organisation
VPN	Virtual Private Network
W3C	World Wide Web Consortium
WebDAV	Web-based Distributed Authoring and Versioning
WebMDs	Web Medical Doctors
WiFi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WMS	Workspace Management Service
WSDL	Web Services Description Language

Chapter 1: Introduction

The Information Technology (IT) research and practice has changed substantially over the past few decades due to the rapid developments in the field of electronics and computing. The computing devices are becoming more powerful as well as smaller and cheaper. This has resulted into more and more intelligence being embedded into devices and systems.

The term Information and Communication Technologies (ICT) is more common these days, and appropriate, due to the indispensable 'communication' elements of the modern IT systems. The 'communication' aspects of computing systems have also grown tremendously during the past few decades. As computing and communication became cheaper, networked computing system, such as the Internet, became common. The rapid developments in the wireless and mobile computing fields added to the 'networked' element of our society and caused us to begin our journey into the pervasive and ubiquitous computing era.

Developments in sensors and nanotechnologies have also made transformational impacts on the way we use ICT. Over the years, our ability to generate data has grown substantially. Developments in sensor networks have benefitted many areas in science, engineering and Digital Economy. Transportation and healthcare are emerging sectors where data is being generated at a huge pace and the requirements to analyse this data has never been more important. ICT is increasingly penetrating into all aspects of our life. The effect of it is that ICT systems are becoming increasingly complex; managing such systems is becoming difficult and highly risky. The costs for managing these complex ICT systems are on the rise. The result is that the system complexity is becoming a threat to further innovation and hindering the penetration of technology into daily life.

The demands for large computational resources and the complexity problem, as discussed in the paragraph above, have led to the development of two related concepts, utility computing and on demand computing; that is, the provision of computational resources to the users as a utility and as needed. These concepts were first realised into the ICT systems through grid computing which brought together and realised also the concepts of resource sharing and virtual organisations. Grid computing, at times, was seen by many as the global computing infrastructure of the future. Grid computing allowed sharing of large scale data and computational resources as well as sharing of experiments leading to many new discoveries. Grid computing developments also motivated researchers to collaborate and hence it acted as a source of accelerated innovation in computing infrastructure development as well as broadly in all areas of science, engineering, humanities, society and economy. Grid computing also addressed the complexity problems related to the development and management of large scale shared ICT systems.

The concepts and developments in ICT systems enabled by grid computing have been taken up by the ICT industry leading to the birth of the cloud computing era. Cloud computing, however, also adopted the virtualisation technologies which was effectively absent in grid computing. Cloud computing is still in its infancy but is making transformational impacts on the way businesses and ICT interacts with each other.

Another important concept that has emerged in this century to reduce complexity in managing ICT systems is autonomic computing, i.e. computing with self-managing characteristics. Cloud computing is the first technology that has realised some of the characteristics as required in autonomic computing in that the cloud computing users see a system which shrinks and expands in response to the varying levels of user demands.

To summarise, ICT systems have evolved from desktop and tightly coupled mainframe computers to highly complex distributed systems. These ICT systems interact with humans at a much advanced level than what was envisaged during the early years of computer development. We now live in a ubiquitous computing and digital economy era where computing systems have penetrated into the human lives to a degree where these systems are becoming invisible. The price of these developments is in the increased costs, higher risks and higher complexity. There is a need to study these emerging systems, their applications, and the emerging market sectors that they are penetrating into.

1.1 Aims, Objectives, and Contributions

The aim of this research is to explore the major technological developments that have happened in the ICT systems during this century with a focus on developing techniques to manage applied ICT systems in digital economy. The aim is to also touch on the evolution of ICT systems and discuss these in context of the state of the art technologies and applications. For the study, we have identified the two most transformative technologies of this century and these are grid computing and cloud computing. The application areas of these two ICT technologies that we have chosen in this thesis include intelligent healthcare and transportation systems.

The contribution of this work is as follows.

1. We have proposed and analysed a workload model of a grid-based ICT system in the healthcare sector. This work demonstrates the potential of computational grids for its use in healthcare organisations to deploy diverse medical applications. A number of organisational and application scenarios for grid deployment in the healthcare sector are considered including four different classes of healthcare applications and 3 different types of healthcare organisations. This work has been published as a referred conference paper published by IEEE Computer Society Press [1]. This work was developed in collaboration with John Williams who is a Professor of Health Services Research at the College of Medicine, Swansea University.

2. An intelligent system for the management of disasters in urban environments is proposed by exploiting the advancements in the ICT technologies, including ITS, VANETs, social networks, mobile and Cloud computing technologies. The particular focus of the work is on using distributed computing and telecommunication technologies to improve people and vehicle evacuation from cities in times of disasters. The effectiveness of the proposed intelligent disaster management system is demonstrated through modelling the impact of a disaster on a real city transport environment. The specific contribution of this work is the development of a novel multi-disciplinary cloud computing based system, its architecture and system performance evaluation. Further work on the development and evaluation is in progress. The system is being analysed using

additional cities, environments and scenarios. This work is continuing to make impact and has resulted into developing international collaborations, one invited (refereed) conference paper [2] and another (refereed) book chapter [3].

3. Analysis of Amazon market sectors, applications, and workload. The contribution of this research is in the identification of the major applications and market sectors where cloud computing is being adopted as well as in understanding cloud computing workloads. This research is specific to Amazon but since Amazon is the top and among the largest cloud computing vendors (see e.g. [4], [5]), this study is also representative of the cloud computing landscape in general. This study is of great benefit in studying capacity management, risk management and other interesting aspects of this exciting and rapidly evolving field of cloud computing. Furthermore, modelling and analysing such aspects of cloud computing providers is vital because collapse of a big cloud vendor due to its inability to understand the variations in its applications, market sectors and workloads could lead to severe impacts not only on the cloud provider and its customers but also on the national and global economies.

4. The two technologies (grid and cloud computing) and the related aspects of the two application areas (transportation and healthcare) have been explained in details using over 300 sources (conference and journal papers, books, online articles and news items, and industry reports).

5. The work contributes in multidisciplinary fields involving healthcare, transportation, mobile computing, vehicular networking, grid, cloud, and distributed computing.

The material on the historical developments, technology and architectural details of grid computing (see Chapter 2) serves to understand as to how and why grid computing was seen in the past as the global infrastructure of the future. It explains that grid computing pioneered and helped develop the concepts, science and technologies for dynamic resource sharing, collaborations and multi-institutional virtual organisations (note that these are fundamental ingredients of cloud computing). This material on grid computing (in Chapter 2) forms the basis which we subsequently use to explain the background, motivations, technological and architectural detail, ongoing developments and the future potential of cloud computing (see Chapter 4). The background chapters on grid and cloud computing, collectively, provide an insight into the evolution of ICT systems over the last 50+ years, from mainframes to microcomputers, internet, parallel computing, distributed computing, cluster computing, World Wide Web, and computing as a utility and service.

The existing and proposed applications and realisations of grid and cloud computing in healthcare and transport (see Chapter 3, Chapter 6, and Chapter 7) are used to further elaborate the two technologies (i.e. the state of the art in ICT systems) and the ongoing ICT developments in digital economy. The workload model for the grid-based healthcare ICT system (see Chapter 3) gives an insight into the possibilities for resource sharing, collaborations and virtual organisations enabled through grid computing. This workload model can be used by researcher and practitioners for developing shared resources and collaborations, and for resource management of ICT systems. In the same endeavour to propose innovative applications of ICT systems for digital economy, an application of cloud computing is proposed in the area of intelligent transportation systems (see Chapter 6); it demonstrated an innovative use of cloud computing to provide dynamic decision making in transportation and disaster management situations for traffic control and city evacuation purposes, including the possibilities of moving, in quasi-real-time, a virtual computing infrastructure and decision software out of a disaster zone. Examples of real cloud services available in the market today are given (services from four major vendors are reviewed briefly while the cloud services offered by the top vendor Amazon are described in detail; see Chapter 5). The information about these services is subsequently used in Chapter 7 where we analyse Amazon market sectors, applications, and workload. As mentioned earlier, this analysis of Amazon cloud space is useful for capacity and risk management of ICT systems.

The grid workload modelling study (Chapter 3) is also applicable to cloud computing systems and can be applied to extend the work presented in Chapter 7. During the course of this PhD, we had initially focussed on grid computing because, by the start of the PhD, the concept of modern day cloud computing was not popular and had not really been taken up by the industry. Grid computing, at that time, was the state of the art for the ICT industry developing technologies for dynamic collaborations, large-scale resource sharing and virtual organisations. By 2010, cloud computing demonstrated high potential to become the future of computing infrastructure and consequently the industry shifted its focus toward cloud computing. Accordingly, we had also shifted the focus of this PhD toward cloud computing technology and applications. We intended to apply the grid-based healthcare model to the Amazon cloud study of Chapter 7 using the real data collected from Amazon but were unable to due to the time limitations.

1.2 Publications

This PhD research has resulted into the following.

1. Saleh Altowaijri, Rashid Mehmood, John Williams, "A Quantitative Model of Grid Systems Performance in Healthcare Organisations," isms, pp.431-436, 2010 International Conference on Intelligent Systems, Modelling and Simulation, 2010. DOI: 10.1109/ISMS.2010.84. [1].
2. Zubaida Alazawi, Saleh Altowaijri, Mohmmad B. Abdljabar, Rashid Mehmood "Intelligent Disaster Management System based on Cloud-enabled Vehicular Networks", ITST, pp. 361-368, 2011 11th International Conference on ITS Telecommunications. DOI: 10.1109/ITST.2011.6060083. [2].
3. Zubaida Alazawi, Mohmmad B. Abdljabar, Saleh Altowaijri, Anna Maria Vegni and Rashid Mehmood, "ICDMS: An Intelligent Cloud Based Disaster Management System for Vehicular Networks", *Communication Technologies for Vehicles*, Lecture Notes in Computer Science, Volume 7266, pp. 40–56, Springer Berlin / Heidelberg, 2012 . DOI: 10.1007/978-3-642-29667-3_4. [3]
4. Zubaida Al-azawi, Mohmmad B. Abdljabar, Saleh Altowaijri, Omar Alani, and Rashid Mehmood, "An Intelligent Disaster Management System with Cloud Computing and Vehicular Networks", 3rd CSE Doctoral School Postgraduate Research Conference, University of Salford, Salford, UK, November 2012, to appear. [6]

The work that I have produced during the PhD, I believe, can easily produce three or more descent quality journal publications. However, this research involved collaborations with multiple individuals, and therefore further research and the publication process had been disrupted due to my supervisor and other collaborators moving out of Swansea unexpectedly causing the research network links to be broken. I am very hopeful that the research contributed towards this thesis will continue to improve resulting in a number of high quality publications in the near future (6 to 12 months).

1.3 Thesis Organisation

The organisation of this thesis is as following.

Chapter 2 serves to introduce grid computing and its related technologies and concepts. It describes in details the grid computing architecture. It also describes the architecture of Grid Toolkit, the technology that enables you to develop and use grid computing systems. Finally it discusses the future of grid computing.

Chapter 3 presents our work on the workload model of a grid-based ICT system in the healthcare sector. Multiple organisational and application scenarios for grid deployment in the healthcare area are considered. The changing healthcare landscape due to the rapid and continued developments in ICT is explored and key issues in grid computing based healthcare are discussed.

Chapter 4 provides a detailed introduction to cloud computing including a discussion of its definitions and scope, its history along with a discussion of its relationship with grid computing, its drivers, current and future prospects, its reference architecture, Service and Deployment models, issues related to its regulations and data protection, and guidelines on migration to this paradigm from traditional IT.

Chapter 5 provides you a look at specific cloud services available in the market. The cloud services offered by IBM, Google, and Microsoft are reviewed briefly. Amazon Cloud services are introduced in detail.

Chapter 6 presents our research on the intelligent disaster management system. The background technologies are introduced. A literature review to establish the motivation for this work is provided. The system architecture is described in detail followed by system evaluation.

Chapter 7 introduces our modelling and analysis work of cloud computing applications, market sectors, and workload.

Chapter 8 concludes the thesis with a number of directions for future work.

Chapter 2: Grid Computing: Dawn of the Utility Computing Era

In 1994, Rick Stevens, Tom DeFanti and Ian Foster proposed and established temporary links among 11 high-speed research networks to create a national grid for a period of two weeks, before and during the Supercomputing 1995 conference. New protocols were created that allowed the national grid users to run applications on computers across the country. This experiment attracted additional funding from the Defense Advanced Research Projects Agency (DARPA), the Department of Energy (DoE), and the National Science Foundation (NSF), and led to the development of the first version (1997) of the Globus Toolkit (GT), the technology that enables grid computing across organisations and allows formation of virtual organisations (VOs). The GT was soon deployed across 80 sites worldwide. Since then, many new projects were initiated by the NSF and the European commission leading to significant commercial interest in grid computing. By 2002, GT version 2.0 was released and the number of its downloads per month from ftp.globus.org grew to over a thousand; reaching to over 25,000 downloads per month by 2003. GT 4.0 version saw major enhancements through new open-standard grid services leading The New York Times to call it "the de facto standard" for grid computing, as well as earning a prestigious R&D 100 award given by R&D Magazine in a ceremony where the GT was named the "Most Promising New Technology" among the top 100 innovations of 2002. In 2005, a group of companies with an interest in supporting GT enhancements for enterprise use formed the Globus Consortium (http://www.globusconsortium.org/) [7].

Since 2005, grid computing has continued to gain many awards for technology innovation to the present day with its latest release GT 5.2.1 having released in April this year. This Chapter serves to introduce grid computing and its related technologies/concepts, describe in details the architecture of grid Computing and the GT (the technology that enables you to develop and use grid computing systems), and discuss the future grid computing.

The Chapter is organised into four sections. Section 2.1 introduces grid computing, taking the reader through its various definitions that have been used, since its inception, over the last two decades or so. Various technologies and concepts related to grid computing have been defined and it is compared with distributed systems and other technologies in order to disambiguate the confusion about grid computing and its relationship with other technologies. Section 2.2 takes a requirements engineering approach and establishes the goals, architecture and functionality of a grid computing platform/framework. The architectural and protocol requirements of Grid framework are discussed along with a detailed discussion of the five architectural layers of grid framework. Section 2.3 describes the GT architecture. GT is the software that allows one to build a grid computing system and VOs (see Section 2.1.1 for definition of VOs). The architectural and functional components (such as service and resource discovery and management, data and execution management and security controls) are described in detail. A discussion of the quality of the software engineering process of the GT is also given. Finally, Section 2.4 discusses the future of grid Computing.

2.1 Grid Computing: A Revolution in IT Systems

Gird computing started off in the early 1990s as an analogy in order to make the computer power easy to access as an electric power grid. Ian Foster, Carl Kesselman, and Steve Tuecke, were the masterminds behind the ideas of grid, and they are known as the fathers of the grid. They have started the mission to create the GT, which is used to integrate computation management, storage management, security provisioning, data movement, monitoring. Furthermore, they have created a toolkit which enables users to develop additional services based on the same infrastructure, which consists of agreement negotiation, notification mechanisms, trigger services, and information aggregation [8], [9].

2.1.1 What is Grid Computing?

As is true of many terms for concepts and innovations, the meaning behind the term grid computing can vary significantly from one user of the phrase to another, and can be the same with terms as diverse as High Performance Computing (HPC), cluster computing, peer-to-peer computing, or even utility computing. In 2004, Ian Foster and Carl

Kesselman, in their book "The Grid 2: Blueprint for a New Computing Infrastructure", described the grid in the following words: "The Grid is an emerging infrastructure that will fundamentally change the way we think – and use – computing. The word Grid is used by analogy with the electric power grid, which provides pervasive access to electricity and, like the computer and a small number of other advances has had a dramatic impact on human capabilities and society. Many believe that by allowing all components of our information technology infrastructure – computational capabilities, databases, sensors, and people – to be shared flexibly as true collaborative tools, the Grid will have a similar transforming effect, allowing new classes of application to emerge" [10].

Coming into its own existence in the 1990s, grid Computing has developed out of the electronic sciences in general, and specifically parallel, distributed, and high performance forms of computing as part of the drive to achieve high speeds and capabilities for data intensive applications, such as in scientific research. One of the first and most widely used definitions focuses on these roots is by Foster and Kesselman: "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities" p.5 [11].

While the sharing of resources had its earliest motivation and successful implementation with the research related applications described above, it soon revealed itself to have significant potential advantages for applications in other spheres of endeavour. It was a logical step from this point that the goal in development should be extending the properties of resource sharing generically to all types of applications which a system might be supporting and not strictly those in obvious need of high performance capability. This goal became the core problem in the IT challenge of developing grid computing, and illustrated by Foster et al as: "The real and specific problem that underlies the Grid concept is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem-solving and resource brokering strategies emerging in industry, science, and engineering" p.2 [12].

The quotation above uses the term Virtual Organisation (VO) to mean the group, whether made up of persons, organisations, or any other conceivable collection of entities, which bands together to set the circumstances for and limits on resource sharing in a given network context [9].

The concept of virtualisation has led to a configuration which allows one physical computer, i.e, one machine, to host multiple virtual machines or virtual operating systems, by having a control program create any number of simulated environments in which multiple virtual computers employ what is called guest software- often as extensive as an entire operating system. The only limit to the number of virtual machines that can run on one physical host is the host computer's resources in terms of hardware. The virtualised platform which the physical computer creates must be able to support guest interfaces in order to deal with the peripherals which the guest software needs. Virtualisation is the biggest key to the cloud's ability to slash expenditures on hardware and the physical components of infrastructure [13].

There are diverse functions, size, duration, configuration, community in VOs, however; they all have similar needs and requirements. For example, there is a need for sharing relationships that are very flexible in client server, and peer to peer modes. Such sharing relationships are needed in order to ensure that well defined and accurate levels of control over how users gain access to shared resources [12].

The VOs described here involve the sharing of resources between either a homogeneous or heterogeneous variety of independent entities known as participants, a group which may be either large or small in number, set up on anything from a strictly short-term to long-term or even indefinitely continuing basis. What is more, they can be either institution-internal or multi-institutional in makeup. The individual VO may be organised out of participants' smaller scale systems, which may or may not have overlapping members. For those who develop the applications designed especially for the VO, providing Quality of Service (QoS) is high priority, no matter what the characteristics (described here) of the individual VO in question, nor whether the service deals with distributed workflow and resource management, common security semantics, problem determination services, coordinated fail-over, or any type of metrics being delivered.

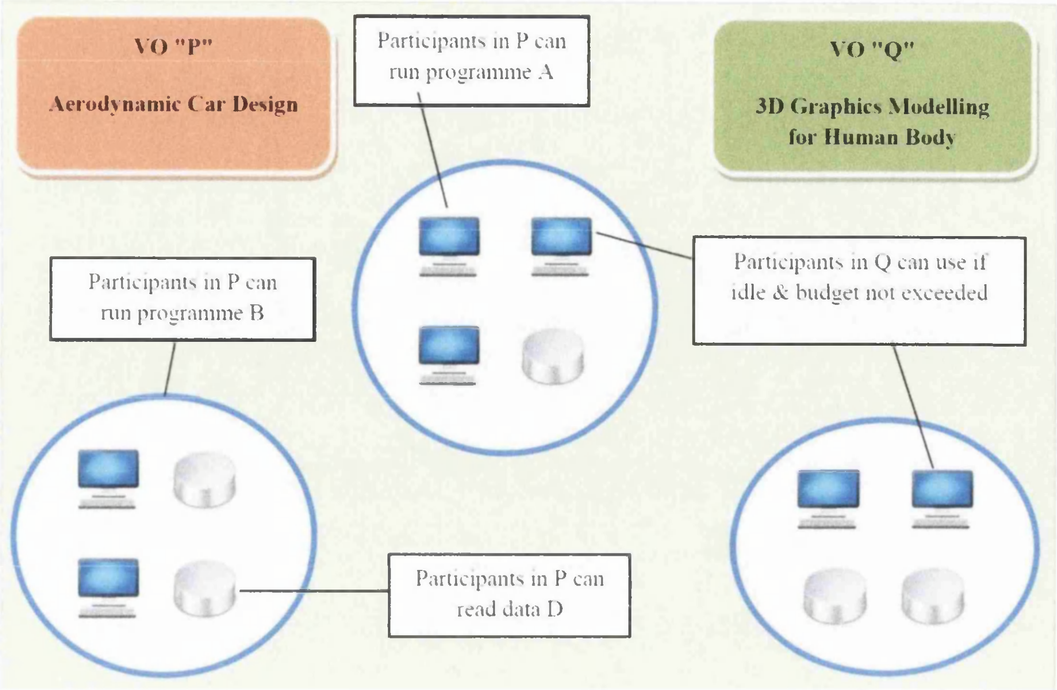

VO "P"
Aerodynamic Car Design
Participants in P can run programme A
VO "Q"
3D Graphics Modelling for Human Body
Participants in Q can use if idle & budget not exceeded
Participants in P can run programme B
Participants in P can read data D

Figure 2.1: Participation of an organisation in one or more Virtual Organisations (VOs), adapted from [10]

The complex and challenging nature of the issues discussed above are illustrated in Figure 2.1, using three hypothetical physical organisations all of which are involved in different overlapping VOs. These three, to be called "A", "B", and "C", are all involved, in many cases mutually so, in different VOs which feature shared computing, data handling, and storage resources. The first two of these companies collaborate in designing an advanced aerodynamic car design model vehicle through VO "P", a hypothetical, internationally based VO, even though they are actively and highly competitive in the larger aerodynamic car design industry. In a totally separate area of its business, organisation "B" is active in VO "Q", a hypothetical consortium that enables regional participants to pool the unused cycles of a locally-based provider, organisation "C", in order to carry out rendering tasks that make heavy use of computational resources. It is normal for there to be conditions governing which resources participants are willing to share, as well as when, where, and with what constraints or limitations, a natural consequence of each resource owner being an independent entity. In the scenario presented here, VO "P" partners of organisation "A" might be permitted to use its simulation service only it the use falls within Organisation A's definition of a "simple" operation.

Those who have extra resources are not the only participants to set the rules; users may also specify what resources they will accept and under what conditions or with what stipulations. For instance, in the VO "Q" part of the example from Figure 2.1 a participant limits on the pooled resources accepted into that participant's internal network to those resources that have been certified as secure according to a predetermined standard. In order for any IT set-up to comply with such constraints from participants, mechanisms to establish the identity of each participant and resource (known as authentication) must be in place, in addition to verifying that any operation being requested is in compliance with all applicable constraints and conditions from all involved participants (known as authorisation).

As time progresses and sharing relationships continue, they may evolve or even transform suddenly in terms of who is participating in them, which participants among as well as within organisations are allowed access, what the nature and conditions of that access are, and what resources are involved in the sharing. These variables may be defined either overtly by naming those involved or by setting the governing access rules and policies as to determine participation by implication.

In the hypothetical scenario, organisation "C" may choose to grant access to all those who can legitimately claim to be customers according to the parameters it sets. Therefore, given all this variability, not only must the key categories such as customer be defined, but mechanisms or protocols for revealing relationships and the characteristics which enable the definition and categorisation of the relationship at any given point in time must also be in place and operational. This necessity is equally for the benefit of the participant, such as a first-time participant in the VO "Q", who must be able to figure out what resources it may access and under what rules and policies, as well as what the conditions of access and characteristics of those resources are.

Sharing relationships differ from their standard cloud-based counterparts in that the latter feature customer/user to vendor/provider relationships, while the former are normally relationships among peers. In this arrangement, providers are in a position to be consumers of others' resources; furthermore, sharing relationships may occur among any subset of the participants in a larger network. The coordination of computational resource involving the many different organisations through their respective, independently owned resources is the essence of sharing relationships. Using the VO

"Q" in Figure 2.1 as an example, an operation may involve starting the computing work on one of the participant's pooled resources and then shifting it elsewhere for sub-operational computations or to retrieve and process data. The delegating of authority in such enterprises becomes a necessary prerequisite for these VOs to function in order to bring the resources of various participants into coordination. Additional complexity and the need for coordination comes from the shifting ways in which given resource may operate. In terms of the example a participant's computer may only run a specified software application in VO "P", while doing general computation cycles for VO "Q".

Not being able to know beforehand what use will be requested at any given time means that, in order to achieve desired qualities of service (QoS), everything from performance metrics to expectations to usage limitations, security measures, and policies must be spelled out and implemented prior to operation. Increasingly, these needs must be applied to individual enterprises operations because unlike in the past, when host-centric internally integrated systems were the norm, today the organisation typically conducts it IT/computing operations in a distributed mix between external networks, services, and the resources they provide, together with company internal infrastructure, itself becoming more heterogeneous, all this complexity driven primarily by e-commerce and the Internet in general.

In the face of all these changes, it becomes vital for enterprises to reintegrate all these components of their computing operations, always mindful of the priority of maintaining QoS. This makes the QoS on the traditional company owned and administered data centre [14] just as necessary in the current distributed computing environment, both internal and external, for engaging effectively in e-business. In order to achieve all this controlled, yet efficient resource sharing, in a grid environment, well planned and set-up infrastructure is essential.

Three key characteristics of grid systems are enumerated in [15], as follows

- ✓ Those resources outside of the scope of centralised control are coordinated by the grid
- ✓ The interfaces and their accompanying protocols are non-application specific and accessible to all
- ✓ The qualities which the service delivers are significant and substantial

Grid computing is ideal for examination and research which involves large numbers of parallel computations, and projects that require growing processing needs, which could be tremendously complex to meet. Grid computing is a powerful tool which could reduce the time of processing applications and healthcare needs. Work that could take up to several months to process or extract, takes only hours with grid. For example, in labs, the use of grid has helped doctors and researchers to identify new viruses. The use of grid computing has made image processing very possible and quick. It is a computer intensive to produce images, however; this is very essential task because it assists doctors and specialists to easily visualise the organs in the patient's body and look for disease, damages or abnormalities. For more details see [8], [9].

Resources which the grid can effectively provide the means for sharing range from the specific and directly accessed, such as actual scientific instruments and application software, through the increasingly more abstract, such as networking of file systems and data storage, to infrastructure related features, such as communications, bandwidth, and processing power.

Moreover, as explained in [16], the grid concept extends to the currently proliferating world of embedded computing as it shows up in Portable Digital Assistants (PDAs) and smart phones, in home appliances and Radio Frequency Identification Devices (RFIDs).

Since the first days of grid computing, the scientific community has embraced a more focused, as well as specified definition as that of a layer of middleware, functioning to make it possible for distinct and independently operating groups or entities to share a set of resources for computation and data storage in a manner that is efficient, reliable, and secure [17]. As mentioned earlier, grid computing moved beyond the realm of eScience, in particular to that of industry, which embraced the concept, although interpreted differently. The quotation below indicates the features which IBM has found most essential to defining grid computing: "Grid computing allows you to unite pools of servers, storage systems, and networks into a single large system so you can deliver the power of multiple-systems resources to a single user point for a specific purpose. To a user, data file, or an application, the system appears to be a single enormous virtual computing system" (Kourpas cited in Stanoevska-Slabeva et al) p.24 [18].

The sharing of computer resources across various networks in a distributed system is the central denotative characteristic by which Insight Research identified grid computing based on their study of the market [19]. According to Rappa [20] , Seti@Home and similar open initiatives, along with research initiatives in the eScience field were the earliest demonstrations of the feasibility of international grids. As a logical extension of the worldwide web, grid computing makes accessible a host of computing resources beyond data to include applications, as well as instruments, sheer computing power, and the like, all through the medium of the Internet. As recorded by [21], with the power grid of electricity-as-a-utility as their model, computer scientists envisioned as early as the 1990s a similar grid system which would make computing power and capacity, along with all necessary support structure, available anywhere and all the time, and most significantly, without the user having to rely on his or her own infrastructure.

Given that analysts, the business community and those in academics and research all have posited terminology and definitions based on their individual purposes, confusion over what constitutes the grid and grid computing is to be expected. The following explanations serve to disambiguate [12]. **Grid middleware** refers to that software which is designed for the specific purpose of functioning to make it possible to pool and share among diverse types of resources, in the process creating VOs. Offered commercially subject to licensing stipulations, grid middleware most often gets installed into previously set-up business infrastructure and integrated as a special layer of virtualisation for the precise purpose of making the sharing of heterogeneous infrastructural components not only possible, but glitch-free.

Grid computing occurs anytime a potentially diverse assortment of groups, businesses, organisations, institutions, and even individuals are utilising what is likely to be an equally assorted grouping of networks, data and application storage systems, and servers, presented to each user as a single unified computational resource through the coordination of grid middleware. While in one respect, grid computing refers to the diverse group of resources, pooled together and virtualised it also refers to the programming that makes the pool function as integrated infrastructure, and also to the applications configured to that infrastructure.

Grid infrastructure means the grid middleware together with the hardware for which it is designed working in conjunction to create the virtual integrated infrastructure, which while in reality consisting of heterogeneous components, still presents to the client or end user as a unified entity just as a single stand-alone computer would.

Utility computing refers to grid computing, along with the applications which it supports, when they are made available on a pay-for-what-you-use basis in one of two ways- first, via an open grid utility service to multiple users or second, in the form of a hosting solution to benefit and individual group, business, organisation or VO.

2.1.2 Grids versus Distributed Systems and other Technologies

As documented in [12], issues involving HPC, the sharing of resources, as well as attendant issues of manageability and coordination have coalesced into what is called the "grid problem," which has become the focus that distinguishes grid computing from distributed computing. Grid Computing is a form of distributed computing, however, the key differentiating factor is that components, such as networks and their attached storage devices, clusters, scientific instruments, etc, are managed as shared resources in a true grid even though they may be individually central to the operation. While the way a grid administers resource sharing may lead it to being also classified as HPC, being HPC does not necessarily mean the system can be called grid computing. These together with other factors and circumstances have made grid computing legitimately a new paradigm in IT and computer operations, analogous to the grid which supplies the electricity that keeps the physical hardware running. Grid middleware, whether commercially available as open source or packaged software, is the linchpin which enables the grid to carry out its quintessential resource sharing functions.

In VOs, the ability to share and communicate with other technologies is very essential part, and it is anticipated that other technologies which are similar to grid are sharing the same idea. However, when it comes to the requirements of the VOs the existing technologies do not have the ability to offer a framework which address requirements. Additionally, grid technologies are differentiated by allowing for resource sharing, and such approaches provide various useful opportunities for the development of grid application. Web technologies such as the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C) standard protocols such as the Transmission Control Protocol (TCP), the Internet Protocol (IP), and the Hypertext Transfer Protocol (HTTP) provide powerful support for the interaction between the browser, the client and the web server. However, such technologies lack features which are required models with richer communication which occur in VOs. An example would be the support of single sign-on, the Web technologies use Transport Layer Security (TLS) authentication, and offer no support to the single sign-on. However, in order to have single sign-on to several Web servers, it is important to ensure that the extensions of the Grid Security Infrastructure (GSI) provide capabilities to the TLS. If this done accurately then the delegation and allocation capabilities of the GSI will ensure that any browser client would have the permission to delegate and assign capabilities to a Web server, which will enable the server to operate on the behalf of the client.

Providers of hosting companies, as well as providers of application service and storage service operate by offering to outsource and subcontract specific types of business and engineering applications especially in the case of the Application Service Providers (ASPs). Additionally, such providers offer to outsource storage capabilities, specifically in the case of the Storage Service Providers (SSPs). In such way, customers and clients can negotiate and agree on defined service level agreement, which classify accesses of specific combination of hardware and software.

Furthermore, Virtual Private Network (VPN) technology is being used to take care of the security aspect which will ensure the extension of the customer's intranet to cover all the resources which the ASP or SSP operate of the behalf of the customer. In addition, with user ids, passwords, and access control list, services of file sharing are provided by other SSPs where the access is provided via the use of the HTTP, the File Transfer Protocol (FTP), or the use of the Web-based Distributed Authoring and Versioning (WebDAV) [22].

There are several enterprise development technologies which are considered to be systems intended and designed to ensure that the construction of distributed applications is easily enabled. Such technologies would include; the Enterprise Java Beans (EJB), the Java 2 Enterprise Edition (J2EE), the Common Object Request Broker Architecture (CORBA), and the Distributed Component Object Model (DCOM). These enterprise technologies provide the clients with the ability to have resource interfaces which are standards, as well as isolated invocation mechanisms. Furthermore, the technologies provide trading services for invention and discovery, which single organisation could benefit from by having the ability to share resources within the organisation easily.

Nevertheless, when it comes to the requirements of the VO, such technologies are unable to meet the requirements due to the fact that the arrangements of the resources sharing are restricted and limited to single organisation where the main structure or form of the interaction is client-server, and not the organised use of multiple resources. Therefore, is it vitally important to allow for the use of a role of grid technologies within enterprise computing and development technologies. E.g., when it comes to the use of CORBA, then it would be beneficial to construct and implement an Object Request Broker (ORB) with the use of GSI mechanisms as this will enable clients to address security issues cross-organisations. Additionally, in order to gain access to resources which are spread across a VO, an implementation of a Portable Object Adaptor (POA) is required, and will be used to communicate with the grid resource management protocol to allow such access to the VO.

Furthermore, Grid-enabled Naming and Trading services could be constructed to use the protocols of the grid information service which will enable the query of the information sources that are distributed across large VOs. By using any of the examples mentioned above, grid protocols will provides enhanced capability as well as enables interoperability with other (non-CORBA) clients.

2.2 Grid Architecture and Protocols

This Section discusses grid architecture and its constituent protocols. The architectural and protocol requirements of Grid framework are discussed first, followed by a detailed discussion of the five architectural layers of grid framework.

2.2.1 Architecting a Grid Enabling Framework

The purpose of this section is to identify and establish the architecture of a grid, i.e. what is required to build a grid enabling platform or framework. The requirement is to have an adaptable open architectural structure which can provide solutions to requirements and needs of the VOs. Figure 2.2 (see [10]) depicts grid layered architecture in terms of the functions and purposes of its components, in addition to revealing their interactions, which are described in more detail in [9]. The most significant force in this architectural design is the need to achieve interoperability among resource providers and users, the backbone of the sharing interactions and relationships, along with the protocols that make this communication possible. Figure 2.2 also compares the relation of the layered grid architecture and the IP architecture; it gives particular attention to the organisation of the required protocols.

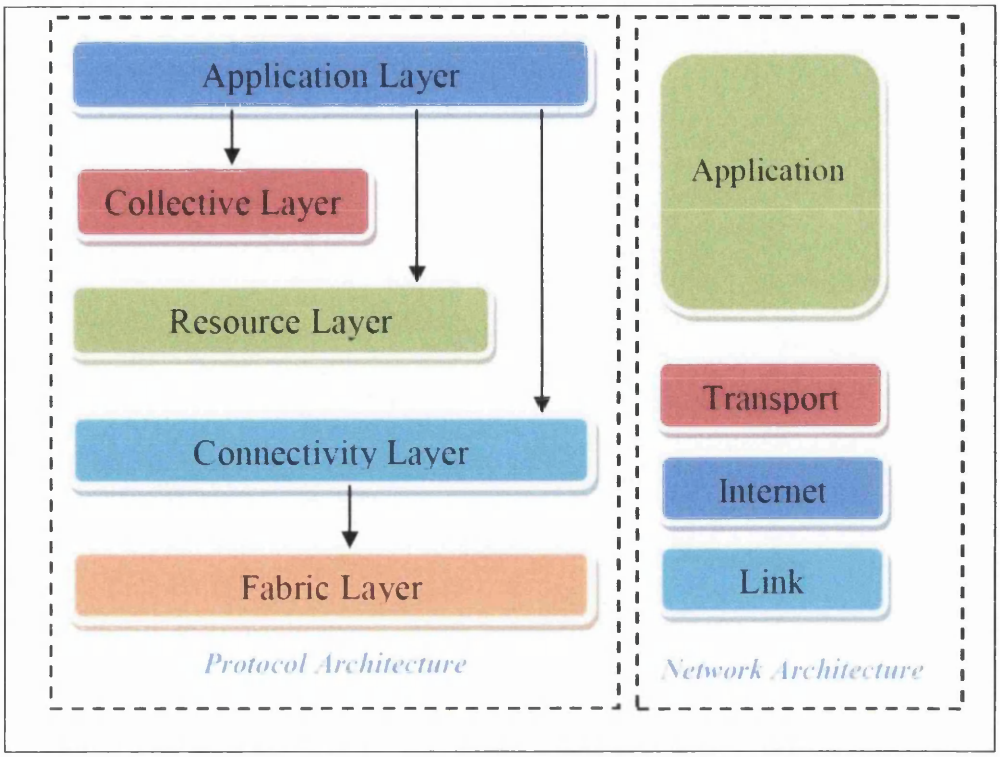


Figure 2.2: Grid layered architecture and the Internet protocol architecture, adapted from [10]

In the architecture, the components of each individual layer contribute to common aspects. Additionally, the components can build on capacities, performances, and behaviours which are provided by other lower layer. The way that protocols are designed at the layers allows them to be implemented on top of a various resource types. These protocols are classified at the Fabric layer, and this enable them to be used to build a broad range of services as well as application-specific behaviours at the Collective layer due to the fact that there is an involvement of coordinated use of numerous type of resources.

When computing resources are distributed, as opposed to centralised, there needs to be a mechanism to bring them together to accomplish the single operation; this is what the computing grids like TeraGrid and Enabling grids for E-science (EGEE) enable [23], [24]. The complex task of finding, accessing, allocating, monitoring, in addition to managing the accounting and billing of all these component activities requires standardised, service-based protocols for the Web. Open Grid Services Architecture (OGSA) fills this function [25].

For some activities at least, standardised protocols have been created and have succeeded to some degree in making possible the delivery of on-demand computing via the Internet. However, according to Buyya et al., [26], there have been various obstacles to achieving this goal across the board for all sorts of applications and operations. The result has been the creation of an impediment to portability with the majority of grid infrastructures, holding users back from turning grids into a computing utility. These problems have found their solution in virtualisation technology [27], [28] (see Section 2.1 for a definition of virtualisation).

2.2.2 The Grid Layered Architecture

2.2.2.1 Fabric Layer

In the Grid Architecture, the Fabric layer consists of all the resources which are part of the grid and have a physical instantiation. Among the specifics in this layer are storage systems, both computational and network resources, sensors, software modules, and catalogues, as well as any other resources of the system [10]. The Fabric layer operates by supplying the resources where shared access is arbitrated by Protocols of the grid. Additionally, any resource could be rational entity such as a distributed file system, distributed computer pool, or computer cluster. Therefore, the implementation of any resource may possibly involve internal protocols such as NFS storage access protocol or a cluster resource management system's process management protocol. In the Fabric layer, the components execute and implement resource-specific operations. Such operations occur on specific resources due to the sharing of the operations at higher levels. This has resulted in a tight association among the functions which are implemented at the Fabric level and the sharing operations which are supported. In order for the sharing of the operations to be more effective and well defined, the functionality of the Fabric must be well implemented. An example would be in the advance reservation, the resource level support makes it achievable for higher-level services to aggregate resources in attractive behaviour that would otherwise be impossible to achieve.

Grid services have provided considerable new capabilities which allow organisations to have large and integrated systems, just-in-time by building up and aggregation. In grid, it is recommended to enable resources to implement enquiry mechanisms which allow for the discovery of their structure, state, ability, and capabilities. Additionally, the resources should be able to implement resource management mechanisms in order to provide quality controls for the delivered services. In grid, each resource has a specific classification of capabilities. For example, in the computational resources, the classification would require mechanisms which are needed to start programmes as well to monitor, to manage and to control the implementation of the resulting processes. Additionally, there is a need for enquiry functions in order to establish the characteristics of the hardware and software characteristics and to furthermore establish the appropriate state information such as queue and current load state if there is a need for them in the scheduler-managed resources.

The classification in the Storage resources requires mechanisms to set and acquire files, also, it is important to ensure that the mechanisms have high-performance and transfers ability in order to perform remote data collection as well as to read and write file's subsets. Additionally, in the storage resources, the management mechanisms can be useful if they have power and can provide control over the resources such as space, disk bandwidth, network bandwidth, and central processing unit (CPU), which are allocated to data transfers. Last but not least, one of the main classifications in the storage resources is the enquiry functions, such functions are very essential for establishing hardware and software characteristics, and to further more determine the load information which are relevant such as the availability space and bandwidth deployment.

In the network resources, resources allocated to network transfers such as assigning priorities and reservation should be controlled and this could be provided by the management mechanisms. In addition, in order to establish the characteristics and load of the network in the network resources, Enquiry functions should be presented. Furthermore, code repositories are specific forms of storage resource which necessitate mechanisms to manage and administrate source which are versioned and object code. This could be achieved by the use of Concurrent Versions System (CVS) which is a version control system, and it allows users to record the sources files' history as well as documenting the CVS. CVS is an essential component of Source Configuration Management (SCM).

We will see later in this Chapter that the GT, which is an implementation of the grid architecture, is designed in a way which allows the use of any existing fabric components such as vendor-supplied protocols and interfaces. Additionally, GT takes account of any missing functionality if they are not being provided by the vendor. An example would be enquiry and analysis software which are used to determine the structure and state of the information for a variety of resource types such as the version of the operating system, the configuration of the hardware in computers, as well as the space which is available in the storage systems, and finally in networks to establish the present and future load of the networks. For more details please see [29], [30].

2.2.2.2 Connectivity Layer

The second layer in the Grid Architecture is the Connectivity layer, which is the heart of the protocols for the communication and authentication, such protocols are very essential for the network transactions which are specified in grid. In the Connectivity layer, data can be exchanged among the resources in the Fabric layer by the use of the communication protocols. On the other hand, the role of the authentication protocols is to build on communication services which ensure that the identity of users and resources are confirmed as well as verified by providing cryptographically secure mechanisms. The Connectivity layer is layer which handles the needed authentication protocols and other core communication for the network transactions specific to the grid. Among the most crucial tasks which this layer handles are the data exchanges between fabric layer resources and tasks to support security, such as supporting single sign on, providing accessibility to all and only those resources to which the individual user has permission to use, and coordinating with local security solutions and stipulations so as not to hinder legitimate access while not granting any prohibited access [10].

In the Connectivity layer, security aspects are very crucial, and therefore; it is very important when it is possible to ensure that the security solutions provided are based upon security standards which exist. Furthermore, most of the security standards which are developed for the IP suite are appropriate to be used in the communication protocols.

When developing VOs, it is fundamental to ensure that the authentication solutions have certain characteristics, such as, single sign on, which allows users to have access to various grid resources which are classified in the Fabric layer. In order for the users to have access to multiple grid resources, they must first log on and be verified. Delegation is another characteristic which is required as part of authentication solutions. In the delegation, the user should have the ability to establish a programme, and such programme should be capable to run on the behalf of the user. This will ensure that the programme will be proficient enough to access any resources that the user is allowed and verified to access.

Integration with various local security solutions such as Kerberos and Unix security is another characteristic which is essential for the authentication solutions in VO. Kerberos is known to be an authentication protocol for computer network, and it works by permitting connection points known as nodes to establish communication between themselves over a network which is not secured. in addition, Kerberos allows the nodes to verify their identity to each other in secure behaviour .When implementing security solutions for grid, it is very critical to assure that such solutions can be easily interoperate with these various local solutions.

The last characteristic that authentication solutions in VO require is the User-based trust relationships. Such characteristic is very fundamental in order to ensure that users can access and use multiple resources without the need for collaborating or interacting between the providers of the recourse to configure and establish the security environment. For example, if a user is granted the right to access resource A and resource B, then the characteristic of User-based trust relationships should allow the user to work on both resources without the need for communication or interaction between the security administrators of the two resources.

In order to ensure that the communication protection is well defined in grid systems, it is vital to have flexible security solutions which provide supports to the communication protection. The security solutions should allow for total management and control over the level of protection as well as providing support for dependable and reliable transport protocols and unreliable protocols. Additional, the security solutions should be able to provide clients and stakeholders with the rights to manage, authorise and restrict access to the resources in ways that meet their requirements.

We will see later in this Chapter that the GT provides its users with all of the above authentication and communication solutions. For example, in the communication part, the IPs are used. On the other hands for authentication, communication protection, and authorisation, GT has used the public-key based GSI in order to certify that the communication among the components of grid systems are secured, and to furthermore provide supports to security across organisational limits. For more details please see e.g. [31]–[34].

2.2.2.3 Resource Layer

The third layer in the Grid Architecture is the Resource layer which is builds on communication and authentication protocols in the Connectivity layer in order to classify protocols, as well as identifying Application Programming Interfaces (APIs) and Software Development Kits (SDKs). The Resource layer consists primarily of management and information protocols which serve to enable the connectivity layer's communication and security protocols to do their job of securely negotiating, initiating, monitoring, as well as the accounting of and paying for what is employed from among individual resources [10]. The management protocols are involved specifically in the negotiating of access to the resources in the fabric layer while staying within the boundaries already prescribed for the conditions of such sharing of access as it related to the individual request for access or usage. The information protocols are essential to the monitoring of the current status of all resources in terms of both structure and availability. Furthermore, the layer classifies and defines protocols and applications to establish secured negotiation and initiation, as well as to provide monitoring, control, and accounting, and finally to ensure that the payment of sharing operations on individual resources is secured. In order to gain the right to access and to manage as well as to control the local resources, the implementations of the protocols in the Resource layer have to call the functions in the Fabric layer. Additionally, the protocols in the Resource layers consider individual resources only and pay no attention to any issues which are related to universal condition or minute actions that are across distributed collections [18].

In the protocols of the Resource layer there are two main classes, and they are; information protocols which are mainly used to get hold of information which are related to the structure and condition of a resource such as the present load, the usage guidelines and policy as well as the configuration. The second class is called the management protocols, such protocols are used for negotiation purposes to gain access to shared resource. Furthermore, the management protocols are accountable for the representation of sharing relationships, and therefore; the protocols should perform and act as policy application point in order to confirm and establish that all of the protocol's operations that are requested are Compatible with the policy of the shared resource. In addition, management protocols may possibly have the ability to provide support to monitor and control any preformed operation.

We will see later that in the GT, there are protocols which are adopted to perform specific functionality; such protocols are mainly standards-based. The protocols include; Grid Resource Information Protocol (GRIP), this protocol is primarily use for identifying the resource information protocol standard as well as the model of the associated information. Another protocol is the HTTP-based Grid Resource Access and Management, which is known as (GRAM) the functionalities of this protocol include the distribution and allocation of computational resources as well as controlling and monitoring such resources. Additionally, GT has a management protocol which is an extended version of the FTP. This protocol is called the GridFTP, and it mainly used as a management tool to manage the data access. In GT, client-side C and Java APIs and SDKs are classified and defined for each one of these protocols. In addition, the protocols are being provided with Server-side SDKs and servers, which are used to smooth the progress of the integration and incorporation of a variety of resources such are storage, network, and computational into the grid. For example, the functionality of the server-side Lightweight Directory Access Protocol (LDAP) functionality is being put into operation and implemented by the Grid Resource Information Service (GRIS), this allow for random resource information to be published.

One of the most important and fundamental elements in the server-side in the Toolkit is called the gatekeeper, such elements allow for GSI- authenticated to communicate with the protocol of GRAM, as well as sending out a range of local operations. Furthermore, in GT, the Generic Security Services are used for obtaining, forwarding, and confirming the authentication and verification credentials, as well providing the reliability and privacy of the transport layer within the SDKs and servers. This will result in providing the ability to have replacement of alternative security services at the Connectivity layer.

2.2.2.4 Collective Layer

The fourth layer in the Grid Architecture is called the Collective layer. In this layer there are Protocols, services as well as APIs and SDKs which function globally due to the fact that they allow for interactions and communication across collections of resources. Additionally, the Collective layer handles resource management on a global or system-wide level, particularly the interaction between distinct resource collections or groupings. Directory, collocation, scheduling, monitoring, diagnostic, and brokering services, in addition to data replication are all significant tasks carried out at this layer. Programming systems make functional through the grid, workflow systems, software discovery and collaboration services, as well as community authorisation and the accounting and payment services, are examples of the tools and programming models are the activities which most frequently call the Collective layer into action [10]. Furthermore, in the collective layer, components that are collective are based on the fine resources and on the neck of the protocol in the Connectivity layer. Therefore, the components are able to implement and put into action a range of sharing behaviours devoid of the need to place any new requirements on the shared resources. The following example services demonstrate how the above is done:

- ✓ Directory services: this enables the participants of the VO to determine the VO resources existence or properties. Additionally, the users of the directory service are able to enquire about any type of resources in different ways such as name, or attributes by their type, availability, or load. Also, in order to build and implement directories, Resource-levels Grid Resource Registration Protocol (GRRP) and GRIP are being deployed to perform such operations.
- ✓ Co-allocation, scheduling, and brokering services, enable the participants of the VO to demand the allocation of resources for purposes that are specific. Furthermore, the services allow the participants to arrange for tasks on the resources which are seen to be appropriate.
- ✓ Monitoring and diagnostics services: the services here are being deployed to provide the users with the ability to monitor the failure, attack and overload of the resources in the VO.
- ✓ Data replication services (DRS); the services here are being used to enable the users to manage the VO storage resources as well as network and computing.

This will allow for maximising the access of the data, while keeping in mind the measurement to be taken such as the response time, consistency, and cost.

- ✓ Grid-enabled programming systems; the benefit of such systems is the ability to provide support any programming models which are common to be used in grid environments. Additionally, the models can benefit from using a range of grid services which will enable them to closely deal with resource discovery, security, and resource allocation.
- ✓ Workload management systems and collaboration frameworks: enable users and organisations to have a better management and control systems over multi-step and multi-component workflows. Such systems are also known as Problem Solving Environments (PSEs).
- ✓ Software discovery services: the role of such services is to provide the users with the best working environment by the determination and selection of the software implementation and execution platform which are known to be best and this is done based on the parameters and limitations of the problem which is being solved
- ✓ Community authorisation servers: in order to implement and make the community policies lead the resource access, the community authorisation servers are being deployed. The servers allow users to generate abilities as well as facilities which enable the member of the community to gain access to the available resources. Furthermore, such servers ensure that enforcement service of the global policy is in place. This is done based on the use of protocols such as resource information, and resource management which are parts of the protocols in the Resource layer, and the security protocols which are found in the Connectivity layer.

The functions of the Collective layer may possibly be implemented and deployed as services which are believed to be constant. The services can be implemented with protocols which they are linked to, or as SDKs that are designed to be in connected and correlated with applications. Additionally, the implementation of the functions can be built on either the Resource layer or on protocols and application programming interfaces in the Collective layer. In Figure 2.3 in the top tier, the figure shows a demonstration of co-reservation service protocol, and there is an implementation of co-reservation service for the purpose of speaking to the protocol. Furthermore, the implemented service call the co-allocation API in the middle tier in order to deploy and put into action co-allocation operations which deal with authorisation, error tolerance, and logging. Moreover, Figure 2.3 shows a demonstration of how API and SDK of Collective co-allocation in the middle tier use a management protocol in the Resource layer in order to control the resources which are underlying.

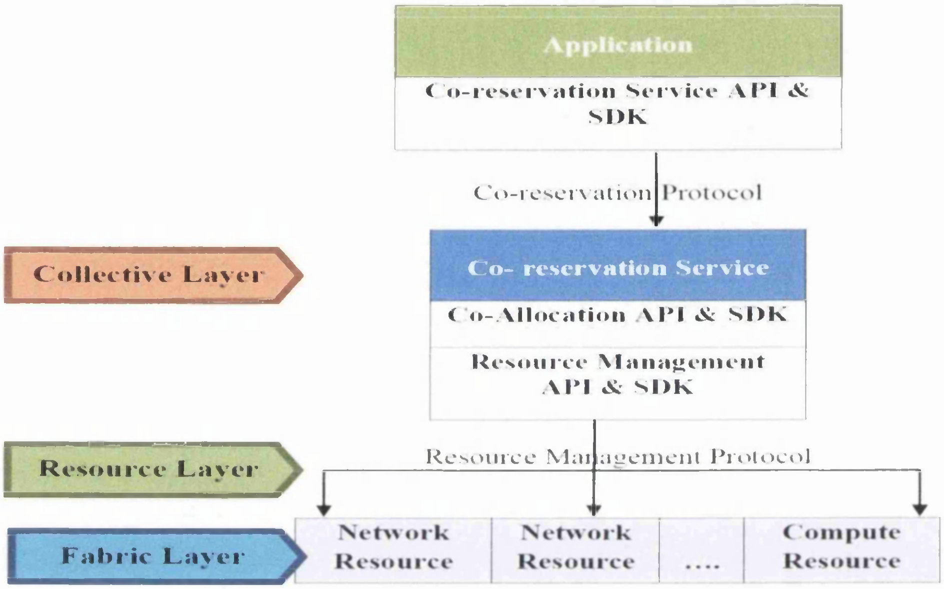


Figure 2.3: Illustration of the Collective and Resource layers' protocols, services, APIs, and SDKs, adapted from [12]

2.2.2.5 Applications Layer

The final layer in the Grid Architecture is called Applications, and it is what is visible to the end user since here reside the applications that the user creates, deploys, reconfigures, and operates, depending on the service. According to Berstis [35], in order for an application to make use of grid infrastructure, it must be either initially designed or else reconfigured to operate in parallel processing employing multiple processors or on a group of distinct heterogeneously configured machines. This layer includes several applications which function within a VO environment.

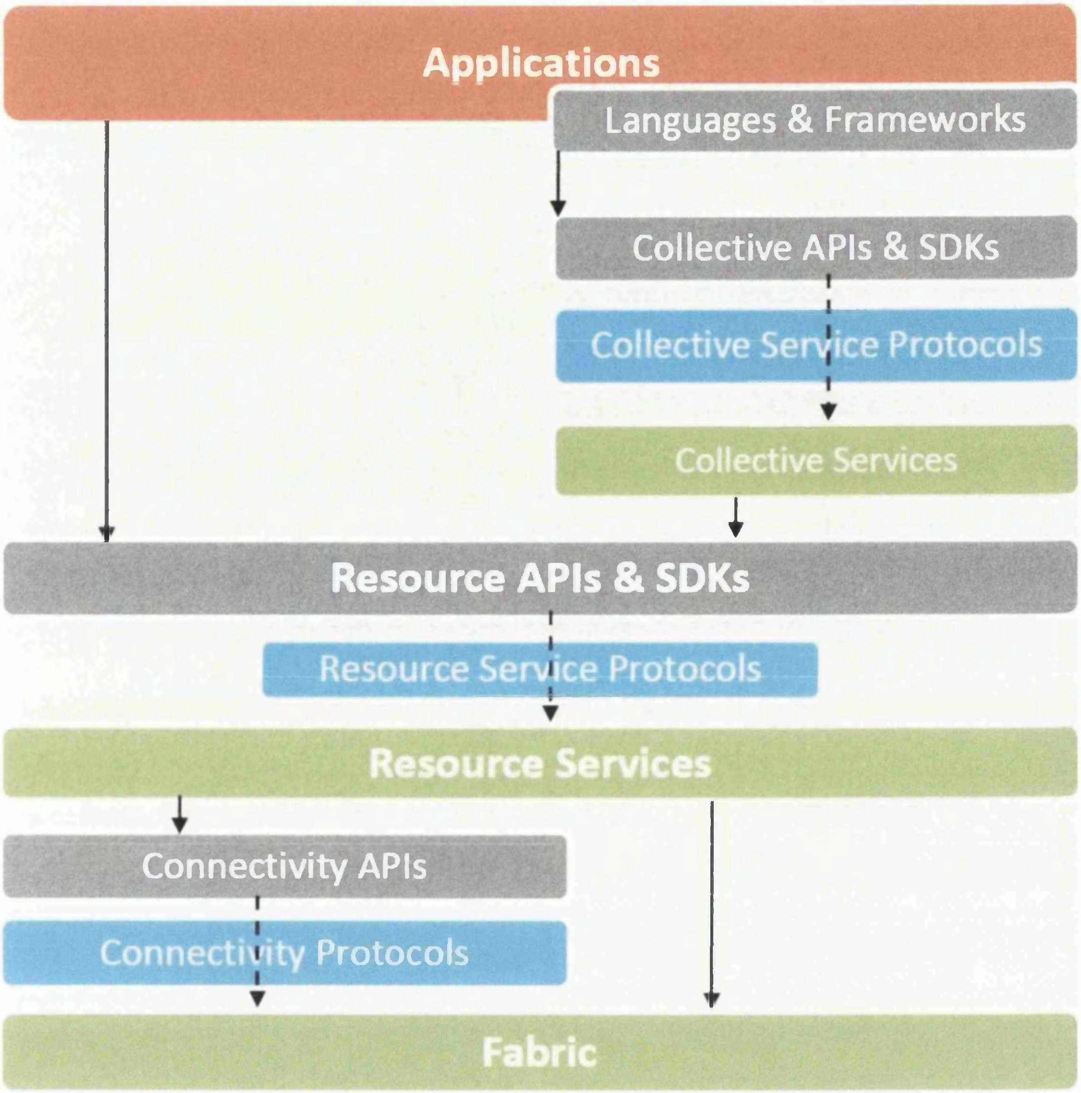


Figure 2.4: Illustration of the Application programmer's view of Grid architecture, adapted from [12]

Figure 2.4 shows how the applications are created based on classifying and calling services that are identified at any layer. Additionally, there are protocols which are used to enable users to gain access to services such as resource Management, data access, resource discovery, etc. those protocols are appropriately defined to ensure that the services they offer are to highest quality possible. Furthermore, application programming interfaces that are provided by SDKs may be used to identify the implementation of how protocols can interact with the right services in order to carry out actions which are required.

As Figure 2.2 on page 39 suggests, all these layers are integrated into a cohesive whole, with each of these five layers relying particularly on those below it through their respective interfaces, all of them interrelating to comprise the grid middleware. The result is that independent entities may share resources in a reliable, efficient (in terms of computers, data, etc.) and secure manner. The functionality runs the gamut of services from the lower level directory, information, resource management, and security to the high level scheduling, resource management, and application development [35]. In between these, grid middleware must handle the mundane functions of accounting and billing, as well as the brokering of resources. As presented in detail in [36], its principal functionalities include:

- ✓ The integration of heterogeneous autonomous resources through virtualisation.
- ✓ The supplying of information concerning the status and availability of resources.
- ✓ The allocation and management of resources in a dynamic, flexible manner.
- ✓ The brokering of resources in accordance with company policy.
- ✓ The maintenance of security and trust, the former including both the authorisation (i.e., ascertaining user's right of access to specific services or data) and the authentication (i.e., ascertaining and confirming the user's identity), the latter relating to accountability in the system
- ✓ License management.
- ✓ The functions of accounting, such as billing and payment.
- ✓ The provision of a significant level of QoS.

The complexity of functionalities in the above list should make it clear that assembling a grid computing system is no simple or easily accomplished matter. Typical grid middleware alone, such as Globus Toolkit 4.0 (GT4), Tomcat 5.5 and Axis, consist of a level of complexity and sophisticated integration that runs into Java classes by the thousand [37]; moreover, then they must be integrated with components from multiple software and middleware providers in order to achieve a fully functioning grid system, which will analogous in complexity to an ecosystem in the natural world. While individual components of a grid computing system may be available commercially, according to Castro-Leon and Munter [16] one cannot go out and acquired an entire grid system with one purchase.

2.3 The Globus Toolkit

Globus Toolkit (GT) is an open source technology which is essentially allowing the use of the technology for grid. GT enables users across businesses, organisations, and geographic boundaries to share computing power, databases, and other tools securely online. The GT consists of software services and libraries which are used to monitor, determine, and manage resource, as well as security and file management. The GT includes software for protection and security, information infrastructure, communication, error detection, data and resource management, and portability. It is put together as a group of components with the aim of using them either separately or jointly to develop applications. Due to the fact that every organisation has unique modes of process, and cooperation between multiple organisations is delayed by the incompatibility of the resources such as data archives, computers, and networks, the GT was considered to eliminate difficulties that prevent flawless collaboration. The central services, interfaces and protocols of the GT allow users and clients to gain access to resources as if they were located within their own computers, whereas at the same time protecting limited control over who can use the resources and when. The material in this section on grids has been taken from a number of references, for more details please see [29], [38].

The development of GT has started since the late 1990s with the aim to provide support to the applications and infrastructures development of the service-oriented distributed computing. The main reason behind GT components is to be able to take in hand, and address issues which are related to security, resource access, resource management, data movement, and resource discovery within a common framework. Additionally, the components of the GT allow for greater Globus environment which contains tools and components that build on, or interact with GT functionality in order to give a broad variety of application-level functions which are useful. Furthermore, such tools have been used in the development of a wide range of the infrastructures for grid and distributed applications as well.

Globus can be defined as a community of users and developers who work together on the utilisation and development of open source software, and associated documentation, for distributed computing and resource confederation. In addition, GT enables developers to have a range of components and capabilities which include service implementations which focus on infrastructure management, tools which allow developer to build new Web services, in Java, C, and Python. Further components and capabilities which GT provides are powerful standards-based security infrastructure and detailed documentation of the variety of components and their interface as well as and how such components can be used to build applications [29].

The main aim of Globus software is to provide applications which associate distributed resources such computers, storage, data, services, or networks. The demands of VOs was the primarily the motivation to start working on Globus. However, in the recent years and due to the fact that commerce and science have similar concerns and interests, the business and commercial applications have become progressively more important [29].

Organisations and federations and businesses are generally inspired and motivated by the ability to access services and resources which cannot be locally replicated easily. For example: an engineer who needs to design and control experiments and researches on remote equipment, as well as connecting and evaluating numerical and physical simulations. Another example would be a business which requires the allocation of computing, storage, and network resources dynamically for various reasons such as supporting workload of physical data or time-varying e-commerce. Business analysts or scientists have the requirements to be able to access data located in different databases across enterprises or scientific associations.

Even though there are unique requirements for every application, there are certain types of functions which recur regularly such as the need to discover resources which are available, the need for configuration of a computing resource to run an application. Addition functions which recur regularly, would include the need move data consistently from one site to another, the need for monitoring system components, as well as controlling what users can do, and managing the users credentials

If the functions mentioned above are implemented with high quality, then the development cost will be reduced. In addition, interoperability can be effectively improved if the implementations of the functions are adopted widely. Globus software does it all as it addresses both goals, using an open source model which encourages both contributions and adoption.

In GT4 Web services mechanisms are being used widely in order to classify the interfaces of it as well as to structure the components. In GT4 the Web services enable users to have widely adopted XML-based mechanisms which are flexible, extensible with the aim to describe, discover, and invoke network services. Moreover, the document-oriented protocols in GT4 are finely suited to the interactions and communications which are believed to be very preferable for distributed systems that are robust.

Services mechanisms in GT4 smooth the progress of the development of Service-Oriented Architecture (SOA) systems and applications structured as communicating services, where the interfaces of the service are described, the access is secured and the operations are invoked in very consistent ways. GT 4 provides set of grid infrastructure services which allow users to be able to implement interfaces to manage resources [39].

2.3.1 Open Grid Services Architecture

The Open Grid Services Architecture (OGSA) illustrates structural design for the environment of a service-oriented grid computing for business and scientific requirements. The OGSA is based on a number of other Web service technologies. In brief, OGSA can be described as distributed communication and computing architecture based in the region of services, with the aim to assure the interoperability on heterogeneous systems so that diverse kinds of resources can communicate and share information effortlessly. OGSA has been illustrated as a modification of the rising Web Services architecture, which is particularly designed to support the requirements grid. The classifications and criteria of OGSA are related to hardware, platforms and software in standards-based grid computing. The OGSA is an extension and modification of the SOA, it is concentrated on the fragmentary issues and challenges such as verification, authorisation, negotiation and enforcement of the policy, management of VOs and customer data integration, and finally the administration of service-level agreements (SLAs). In order for a Web service to be regarded as a grid service, it ought to allow clients and users to without difficulty determine, update, amend and delete information which are related to the service's state. Additionally, it must allow them to identify how the service progresses and to ensure ongoing compatibility with other services. The main purpose of grid here is to ensure the optimisation of the communication and interoperability among resources of all types [29], [38].

Grid computing is a new technology which meets the requirements of setting up, organising, and utilising dynamic and cross-organisational VOs' sharing relationships. Grid architecture is defined as the ability to establish and set up sharing relationships between participants, and keeping in mind that the main issue to be taken into consideration here is the Interoperability, which is defined in a networked environment as common protocols. Therefore, the architecture of grid is protocol architecture where there are set of protocols in which the sharing relationships are set up, administered, and utilised by the users and resources of the VOs. Moreover, protocols can be used to identify how the components of distributed systems can collaborate and interact with each other to ensure that the structure of exchanged information, and particular behaviours are accomplished [40].

In grid architecture, services are essential elements and they are exclusively classified by the protocols which they communicate and the behaviours which they put into operation. Additionally, standard services such as computation and data access, resource discovery, and data replication enable the improvement of the services which are offered to participants of the VOs, and allow for taking away unnecessary details which could result in delaying the development of VO applications.

In order to ensure that the users of the VOs are running the applications effectively, it is very important for the developers to be aware of the fact that the applications should be developed to meet the needs and requirements of such complex and dynamic execution environments. Therefore, the developers should consider APIs and SDKs due to the fact that VOs require more than just protocols, services, and interoperability. Also, other areas which must be considered when developing applications are the strength and accuracy of application performance, as well as the costs of the development and maintenance. Grid architecture can be seen as highlights which classify and describe protocols and services, as well as APIs and SDKs.

2.3.2 The Globus Toolkit Architecture

Figure 2.5 illustrates several set of components of the Globus Architecture such as service implementations that are used to put into operation valuable communication and infrastructure services. The services are deployed to address concerns such as execution management (GRAM), data access and movement (GridFTP).

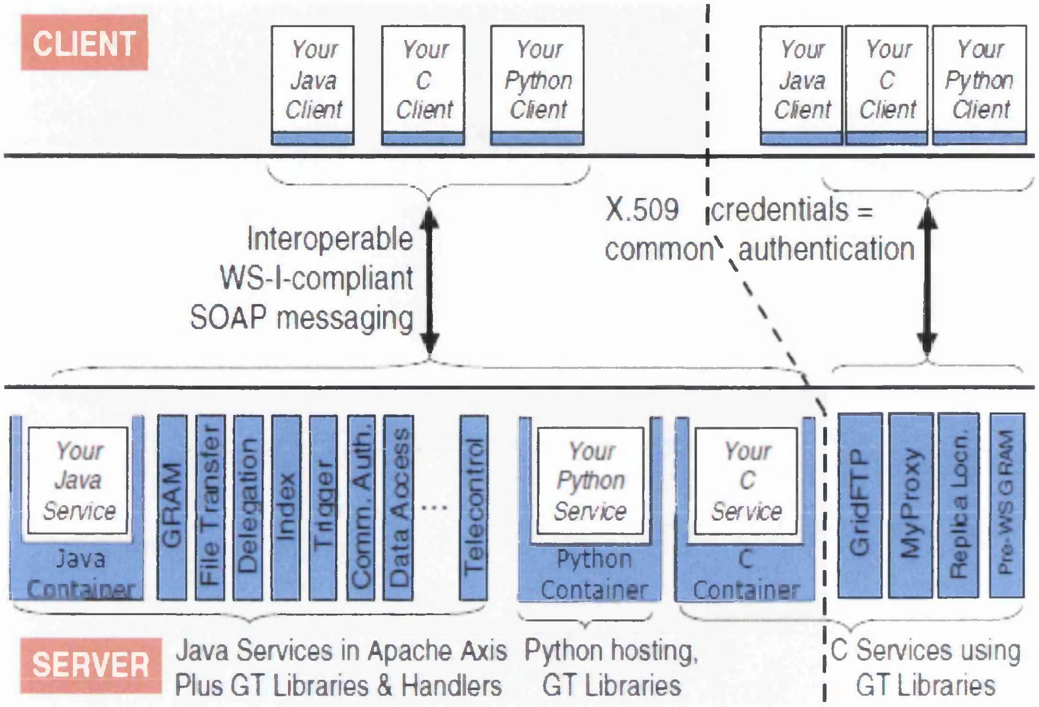


Figure 2.5: Architecture Components of Globus Toolkit 4, adapted from [29]

Other components which are being used in the GT4 are the three containers which are used to host user-developed services that are written in Java, Python, and C, respectively. Such containers are known for the ability to provide the implementations of security, management, discovery, state management. Moreover, the Globus Architecture provides the users with a set of client libraries which enable the clients programming in Java, C, and Python to invoke operations on both GT4 and other user-developed services [41].

GT 4 is powerful software which does not only provide useful services but much more. In GT4 clients can have the ability to interact and communicate with different set of services in similar ways. This is done by the use of the consistent and uniform abstractions and mechanisms. Such operations allow clients to facilitate as well as smoothing the progress of the construction of complex, and interoperable systems as well as encouraging code reuse. Such uniformity in GT4 take places at several levels:

- ✓ A Web Services Interoperability which compliant Simple Object Access Protocol (SOAP) messaging among Web services and their clients.
- ✓ A security and messaging infrastructure enables interoperability among different applications and services.
- ✓ An authorisation framework which is powerful and extensible enough to supports a range of different authorisation mechanisms.

Figure 2.6 demonstrates another viewpoint on GT4 structure as it shows on the right side of the figure the main components which are provided for basic runtime, from left to right it shows the security, the execution management, the data management, and the information services. These components can be used to perform various tasks.

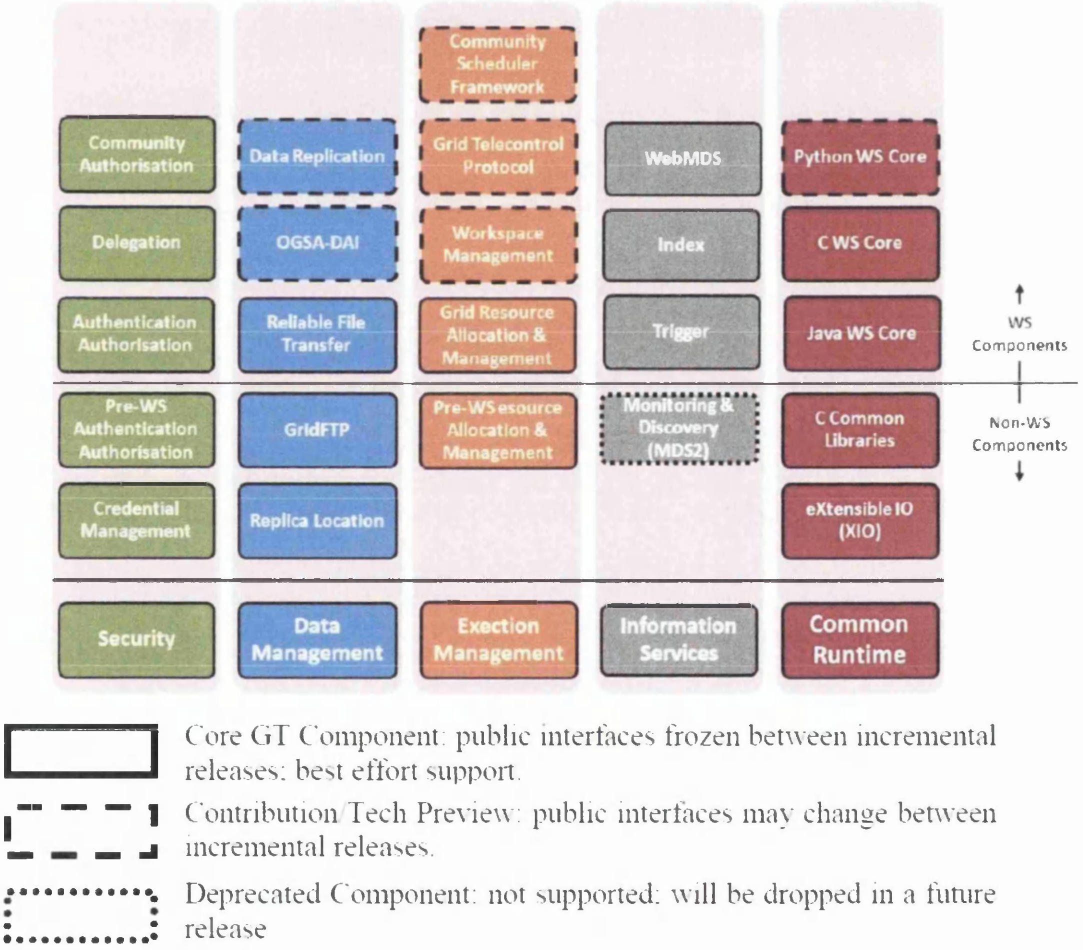

Community Scheduler Framework
Community Authorisation
Data Replication
Grid Telecontrol Protocol
WebMDS
Python WS Core
Delegation
OGSA-DAI
Workspace Management
Index
C WS Core
Authentication Authorisation
Reliable File Transfer
Grid Resource Allocation & Management
Trigger
Java WS Core
WS Components
Pre-WS Authentication Authorisation
GridFTP
Pre-WS esource Allocation & Management
Monitoring & Discovery (MDS2)
C Common Libraries
Non-WS Components
Credential Management
Replica Location
eXtensible IO (XIO)
Security
Data Management
Exection Management
Information Services
Common Runtime
Core GT Component: public interfaces frozen between incremental releases; best effort support.
Contribution/Tech Preview: public interfaces may change between incremental releases.
Deprecated Component: not supported; will be dropped in a future release

Figure 2.6: Globus Toolkit 4 Components, adapted from [29]

2.3.2.1 Execution Management

If a user wants to manage an execution, for example, a user would like to run a task on a computer, or to deploy and manage a service that provides some capability to an organisation. Then the user firstly needs to obtain access to a computer, and then the user would need to make sure that the computer is configured to meet the needs. Additionally, the user needs to ensure that the computer is configured to stage an executable, to start execution of a program, and to monitor and manage the resulting computation.

In the GT4 the Grid Resource Allocation and Management (GRAM) service addresses these issues, and provides users with a Web services interface which allow them to initiate, monitor, and manage the execution of random computations on remote computers. The interfaces of GRAM provide the clients with the ability to express such things as type and quantity of resources required, as well as to have the data staged to and from the execution site. Additionally, it allows for the execution and its arguments, qualifications to be used, and job determination requirements. Moreover, other operations in GRAM allow clients to monitor the status of both the computational resource and individual tasks, to subscribe to notifications relating to their status, and control a task's execution.

In Globus TK4, GRAM service can be used for many different purposes, some of these purposes are; the Grid Physics Network (GriPhyN), the Virtual Data System (VDS), Ninf-G, and Nimrod-G are all tools which use interfaces of GRAM to dispatch large numbers of individual tasks to computational clusters. For example, the Genome Analysis Database Update (GADU) service regularly uses VDS to dispatch several million BLAST and BLOCKS runs as it updates its proteomics knowledge base. Additionally, there are a range of applications that use GRAM as a service deployment and management service. The way the applications use GRAM is by firstly using a GRAM request in order to start the service and then to control the consumption of its resource and provide for restart in the event of resource or service failure [42].

However, there are execution management components which are provided in the GTK4 as tech previews, and the reason for that is because such components are not fully tested like other components and they are more likely to change in the future. Some of these components are; Workspace Management Service (WMS) which is used for the dynamic allocation of Unix accounts as a simple form of sandbox. Another component is the Grid TeleControl Protocol (GTCP) which is a service that is deployed in managing instrumentation. Such service has been used for engineering facilities of earthquake and microscopes.

In grid, resources can be from different providers and therefore, Resource Management System (RMS) is required to ensure that all providers of the resources are trustable [43], [44].

2.3.2.2 Data Access and Movement

Applications of Globus regularly need to manage, to integrate, and as well as to provide access to large quantities of data at one or many sites. However, the issue here is that the data is very complex, and there is no software which can easily deal and sort such complex data on its own. Nevertheless, GT4 has components which can provide clients with the ability to implement and apply functional mechanisms that can be used individually and in combination with other components to develop attractive solutions. An example would be GridFTP which is a component of the Globus. The GridFTP specification provides clients with powerful libraries and tools, which can be used to reliably and securely move memory to memory and disk to disk data.

Furthermore, there is the Reliable File Transfer (RFT) service which allows clients to have reliable management of multiple GridFTP transfers. The RFT has been used arrange and organise the transfer of one million files from one astronomy archive to another. Moreover, there is the Replica Location Service (RLS) which is a scalable system that is used to provide maintain access to information about the location of replicated files and datasets.

RLS has been used by the Laser Interferometer Gravitational-Wave Observatory experiment (LIGO) to manage more than 40 million file replicas. GridFTP and RLS can be combined by the DRS in order to provide the management of data replication. Finally there is the Open Grid Services Architecture Data Access and Integration (OGSA-DAI) tools. Such tools have been developed by the UK eScience program with the aim to provide access to relational and XML data [45].

2.3.2.3 Services and Resources Monitoring and Discovering

In distributing systems there are several fundamental functions which includes monitoring and discovery, especially in systems which cover different locations due to the fact that no single person could have detailed knowledge of all of the components. With the monitoring function clients and users can become aware of and identify problems which could arise in such frameworks. Additionally, the discovery function provides clients and users with the ability to classify resources or services with required properties. In order to acknowledge the significance of such functions, the monitoring and discovery mechanisms are built at a fundamental level in GT4.

In GT4 there are mechanisms which are standardises to associate XML-based resource properties with network entities and to provide the access to the properties. Such mechanisms which are built into every service and container of GT4, and clients could integrate them easily into any user-developed service are fundamentally implementations of the Web Services Resource Framework (WSRF) and WS-Notification specifications.

The services in GT4 can be constructed to register with their container. Additionally, the containers could easily be configured to register with other containers, which as a result allow the creation of hierarchical organisations.

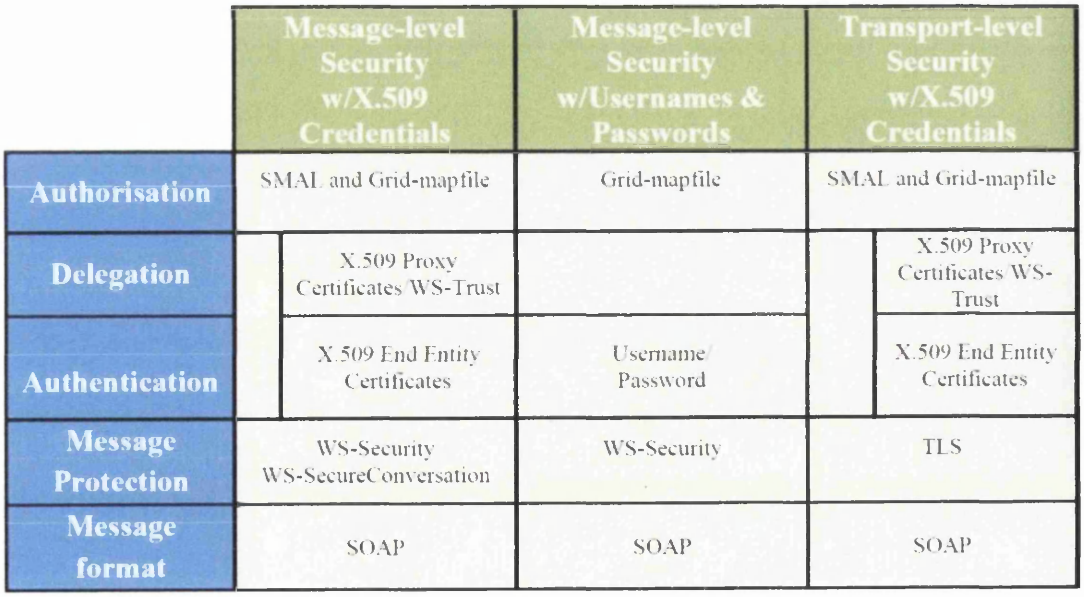
Such mechanisms in GT4 allow users to have a powerful framework for monitoring various collections of distributed components as well as obtaining information about components for discovery purposes. For instance, the Earth System Grid (ESG) uses these mechanisms to monitor the status of a variety of services which the ESG uses to distribute and provide access to more than 100 TB of climate model data.

Users and clients who use The GT4 are provided with two aggregator services which allow for the collection of the recent state information from registered information sources. Due to the fact that not all of the information sources support WSRF/WS-notification interfaces, the aggregators which are provided by the GT4 can be configured to collect data from any information source, whether XML-based or otherwise. Additionally, the GT4 allow users to have a range of browser-based interfaces, command line tools, and Web service interfaces which allow them to query and access the collected information, especially the Web Medical Doctors (WebMDs) service which can be configured via the Extensible Stylesheet Language Transformations (XSLT) transformations to create specialised views of Index data [46].

2.3.2.4 Security Controls

The area of security is above all important and challenging especially when it comes to resources and/or users spans multiple locations. A range of players may want to apply control over who can do what, and this could include the owners of individual resources, as well as the users who initiate computations and the VOs which are established to manage resource sharing.

Applying control may include variously enforcing policy and auditing behaviour. Therefore, it is vitally important to address these requirements when designing mechanisms, the aim should not only be the protection of the communications but also to limit the impact of break in at end systems. Moreover, in order to have a complete security solution, the systems must combines components concerned with establishing identity, applying policy, tracking actions, etc., to meet specific security goals. GT4 and related tools provide users and clients with powerful building blocks that can be used to construct a range of such systems.



	Message-level Security w/X.509 Credentials	Message-level Security w/Usernames & Passwords	Transport-level Security w/X.509 Credentials
Authorisation	SMAL and Grid-mapfile	Grid-mapfile	SMAL and Grid-mapfile
Delegation	X.509 Proxy Certificates/WS-Trust		X.509 Proxy Certificates/WS-Trust
Authentication	X.509 End Entity Certificates	Username/Password	X.509 End Entity Certificates
Message Protection	WS-Security WS-SecureConversation	WS-Security	TLS
Message format	SOAP	SOAP	SOAP

Figure 2.7: Globus Toolkit 4 security protocols, adapted from [29]

At the lowest level, GT4's highly standards-based security components implement credential formats and protocols that address message protection, authentication, delegation, and authorisation. As shown in Figure 2.7 support is provided for (a) WS-Security- compliant message-level security with X.509 credentials (slow) and (b) with usernames/passwords (insecure, but WS-I Base Security Profile compliant) and for (c) transport-level security with X.509 credentials which is (fast and thus the default). In GT4's default configuration, each user and resource is assumed to have a X.509 public key credential. Additionally, protocols are implemented in order to allow two entities to validate each other's credentials, to use those credentials to establish a secure channel for purposes of message protection, and to create and transport delegated credentials that allow a remote component to act on a user's behalf for a limited period of time.

[47] Has Illustrated three types of solutions for grid security; the first solution is the system-based solution with the aim to protect and secure the environments of grid computing. System-based solution' centre of attention is to control the software and grid system directly to ensure that security is achieved. The second solution is the behavioural solution, which is intangible and perceptive, and they accentuate the management and policy controls over the system-based solutions to ensure that the security of grid is maintained. The final solution is the Hybrid Solution which focuses on the authentication and authorisation.

Due to the decentralised nature of gird commuting authentication and authorisation is very sensitive areas. Although they are related to each other, they both have diverse specifications and purposes. Whilst the purpose of authentication is to verify the identity of a process or a person, authorisation' purpose has a process that establishes what rights and permissions a person or a server supposed to have [48].

2.3.2.5 Building New Services

GT4 has a wide range of software that are enabled to support the development of components that implement Web services interfaces. Furthermore, GT4 deals with several issues such as message handling, resource management, and security, which as a result provide the developers with the ability to focus their attention on implementing application logic. Additionally, GT4 also put together additional GT4-specific components to provide GT4 Web services containers for deploying and managing services written in programming languages such as Java, C, and Python. Figure 2.8

illustrates the capabilities of GT4 containers which can host several types of different services.

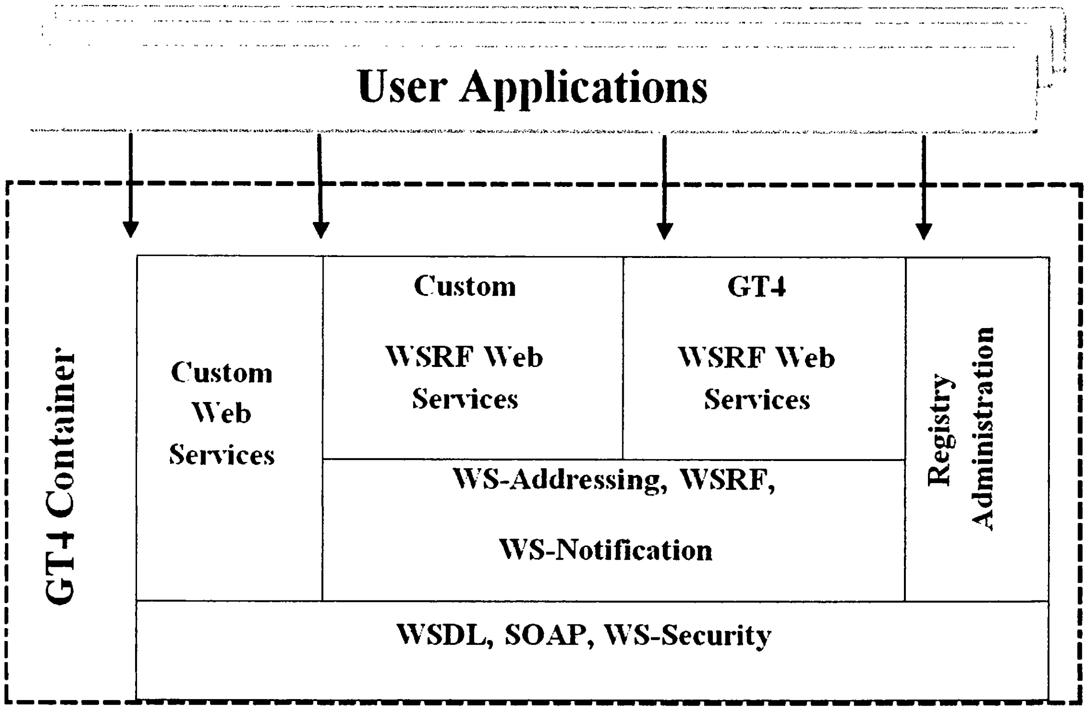


Figure 2.8: Illustration of GT4 Containers, adapted from [49]

From Figure 2.8, we can see that GT4 has several containers such as the implementations of basic WS specifications such as the Web Services Description Language (WSDL), SOAP, and WS-Security support services that make use of such specifications to implement basic Web services functionality. Additionally, there are the implementations of other specifications, particularly the WS-Addressing, the WSRF, and the WS-Notification, support services that want to expose and manage state which are associated with services, as well as back-end resources, or application activities [50]. Furthermore, there is the enhanced registry and management capabilities container, which is notably known as the representation of information about services running in a container as WS-Resources, facilitate the creation of distributed registries and system monitoring tools.

2.3.2.6 The Quality of Software Engineering

The engineering processes of the GT have been improving increasingly over the past years. Such improvements have been made potential by increasing the engineering resources of the software as well as having more users available for further testing. The engineering processes now include; extensive unit test suites and the use of test coverage tools to which allows users to evaluate coverage. Regular automated execution of build and test suites on more than 20 platforms via both local systems and the distributed build and test facility of GRIDS. Additionally, there are processes such as the extensive performance test suites which are used to evaluate various aspects of component performance that include latency, throughput, scalability, and reliability. Documentation plan of the GT, it is managed by a dedicated documentation specialist with the aim of ensuring complete coverage and uniform style for all components. Furthermore, there is a well-defined community testing process, which in the case of GT4 included a six-month alpha and beta-testing program with close to 200 participants.

2.4 Future of Grid Computing

With grid computing coming to light in the 1990s as computing infrastructure distribution for complex science and engineering, there has been significant work done on the development and improvement of the infrastructure. Grid is known to be powerful for providing solutions to large-scale resource sharing, innovative applications, and this makes it an essential new field. It was argued that one of the main problems which grid technology has is in the resource sharing as well as the dynamic, multi-institutional VOs, where the problem is the access to computers, software, data, and other resources that are available to users. Therefore, such sharing should be controlled, and access to the resources should be well defined in order to ensure that only authorised users can be allowed to share and access certain resources. Grid technologies have come very far and have addressed these issues to a great extent such that grid computing has become a reality, used worldwide within and across thousands of experiments, projects and VOs. The focus of grid technologies is in the dynamic, and cross-organisational sharing, hence, such technology overcomes the limitations of existing establishments. Therefore, other distributed computing systems can collaborate with grid in order to accomplish powerful resource sharing across institutional limitations [12], [51].

The Globus software and community development is owed to the confluence of several factors. These factors includes the completion of GT4 which means that Globus has obtained a solid Web services base that allows clients and users to build additional services and capabilities. Another factor is the sustained funding for eScience support which provides the ability to the users to accelerate efforts aimed at meeting demands for greater scalability, functionality, usability, and so forth. Furthermore, there is the creation of organisations dedicated to the support needs of industry; this enables the commercial adoption contributions to accelerate. Additionally, there is the rapidly growing user community factor which has resulted in the increasing of the quantity and quality of user feedback, code contributions, and components within the larger Globus ecosystem. Finally, the revisions to the Globus infrastructure and governance processes make it easier for the developers to engage additional contributors to the software and documentation. The latest Globus Toolkit release GT 5.2.1 was released in April this year, Figure 2.9 depicts major components of the Globus Toolkit Version 5 (GT5).

Grid computing has continued to attract support from academia, government and computing industry giants including University of Edinburgh, the NSF, the National Aeronautics and Space Administration (NASA), Sun Microsystems, Microsoft and IBM. In 2004, Univa Corporation was formed in order to provide commercial support for grid computing and Globus software. Univa has released its latest Grid Engine 8.1 recently and his customers include Tata Steel Automotive Engineering. A number of conferences and events centred on grid computing have been taking place over the years. These include, among others, UK e-Science All Hands Meetings (http://www.allhands.org.uk/) and Open Grid Forum Events (Formerly Global Grid Forum: http://www.ogf.org/).

Grid computing has been seen by many as the global computing infrastructure of the future [52]. This is due to the fact that many organisations and researchers desired large scale computational resources but relatively very few organisations were able to afford it. Grid computing allowed sharing of large scale data and computational resources as well as sharing of experiments leading to many new discoveries. Grid computing developments also motivated researchers to collaborate and hence it acted as a source of accelerated innovation in computing infrastructure development as well as broadly in all areas of science, engineering, humanities, society and economy. However, activities in grid computing over the years are declining. Is it because the demand for sharing, collaboration and large scale resources is declining or that grid computing technologies have failed to deliver its promises? We explore these questions and look at the future of grid computing in the following.

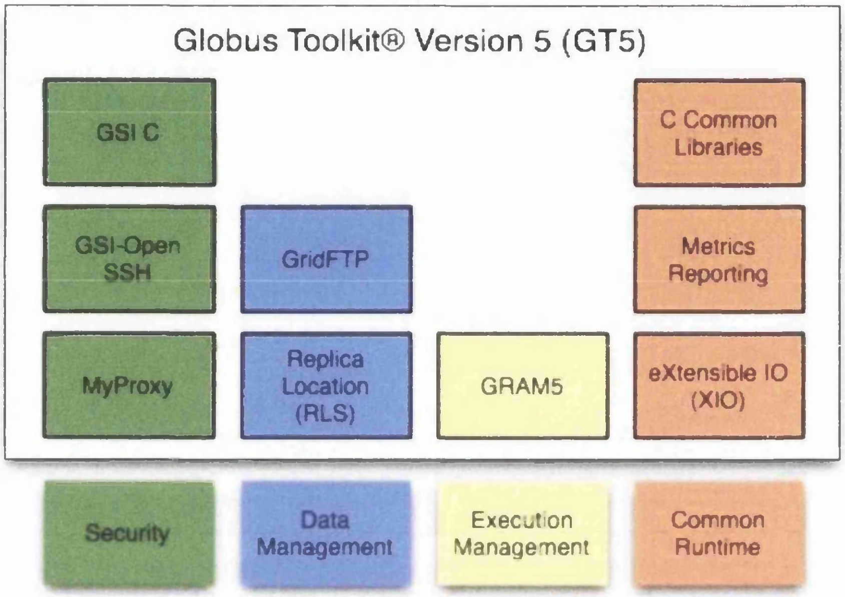


Figure 2.9 Globus Toolkit Version 5 (GT5) Components (www.globus .org)

Over the years, our ability to generate data has grown substantially. Development in sensor networks have benefitted many areas in science, engineering and Digital Economy. Transportation and healthcare are emerging sectors where data is being generated at a huge pace and the requirements to analyse this data has never been more important (see Chapter 56 where we discuss a disaster management system exploiting real time data). Hence, the major drivers for grid computing still exist rather the need for them has increased to a large extent. So why has grid computing been unable to maintain and increase its momentum? We believe that the grid computing ideas of resource sharing have been taken up by the industry and they have named it Cloud computing. For example, grid computing allows sharing of resources which is adopted by cloud computing vendors who own the resources which are shared by their customers. It is possible to create VOs by multiple individuals or organisations by renting computational resources from same cloud computing vendor. In this case, in contrast to grid computing, the administrations and interoperability issues would be dealt by the cloud computing vendor. Cloud computing however, is distinct from grid computing due to the relatively recent virtualisation technologies which were quickly adopted by the cloud computing industry while grid computing community failed to exploit it.

We believe that cloud computing holds the future. The need for grid computing (eScience, sharing, collaboration, and VOs) will remain and perhaps it will be satisfied by cloud computing through its virtual resources and highly interoperable infrastructure.

Having introduced grid computing in detail in this Chapter, the next Chapter will introduce and discuss the work on grid performance modelling of healthcare organisations.

Chapter 3: Grid Computing in Healthcare and its Workload Model

The rapid developments in information and communication technologies (IST) are shaping the world and all of the dimensions of our life. Healthcare is no different and is undergoing a complete ICT-driven transformation. Next generation health applications, services and organisations will require massive computational resources. Moreover, the next generation healthcare infrastructure will have to support a diverse range of applications and their ability to communicate with each other within and across the organisational boundaries. On the other hand increased demand and awareness of environmental and financial sustainability requires efficient and green use of resources. Grid computing provides a solution because, in addition to its other benefits, it is financially, computationally and environmentally efficient. Furthermore, it provides gradual and smooth technology deployment options with minimum disruption.

Grid computing is already playing a key role in the rapidly growing healthcare sector, providing storage and computing power for initiatives in several biomedical disciplines. The deployment of grid computing will enable researchers to examine diseases and rare conditions. Additionally, grid computing will provide doctors with new ways to analyse and treat patients. Moreover, by integrating the medical data available, specialists and doctors could start personalising treatments for patients. Collectively, grid computing and e-Health are shaping the future of healthcare as well as the future of collaborative research and business between healthcare organisations.

This chapter aims to quantitatively model and demonstrate the potential of computational grids for its use in healthcare organisations to deploy diverse medical applications. At the heart of this study is our understanding that sustainability, be it financial, environmental or computational, will be the key driver in the design, deployment and operation of the future services and systems. We consider multiple organisational and application scenarios for grid deployment in the healthcare area including four different classes of healthcare applications and 3 different types of healthcare organisations. We identify the computational requirements of key healthcare applications and build a Markov model of a Grid-based healthcare system. For each scenario, we compute steady state probability distributions of the respective Markov models and analyse the system performance using the results. Various performance measures of interest such as blocking probability and throughput could be computed from these state probability distributions. The work presented in this chapter has been published as conference paper published by IEEE Computer Society Press [1]. The paper was co-authored with John Williams who is a Professor of Health Services Research at the College of Medicine, Swansea University.

The rest of the Chapter is organised as follows. Section 3.1 reviews the changing healthcare landscape due to the rapid and continued developments in ICT. Section 3.2 discusses key issues in grid computing based healthcare. Section 3.3 describes our healthcare gird system Markov model. Section 3.4 evaluates the system performance using the Markov model. Finally, Section 3.5 concludes the Chapter.

3.1 Information and Communication Technologies in Healthcare

The increasing role and benefits of ICT in healthcare are already visible in the enhancement and emergence of technologies such as healthcare telematics, health informatics, epidemiology, bioengineering and Healthcare Information Systems (HIS). The major drivers for ICT based healthcare include demands for increased access to and quality of healthcare, rising healthcare costs, system inefficiencies, variations in quality of care, high prevalence of medical errors, greater public analysis of government spending, ageing population, and the fact that patients and the public want a greater say in decisions about their health and healthcare. However, the healthcare industry has not been able to reap the full potential of ICT mainly because of the social reasons (e.g. sensitivity, privacy and trust) and lack of business models. A HIS usually includes the following components; the hardware and software, and the databases of healthcare data such as patient records. A HIS could be seen as simple as a set of applications and data used to carry out eye examination in a clinic, or as complex as a set of applications and data that takes all the Computerized Tomography (CT) scans of patients and produces a set of symptoms through analysis. Medical research has emerged as one of the most demanding areas of computational resources. Most specialists and doctors are now demanding genomics and test results for their daily medical diagnoses. Genomics and test results are being produced using the latest available technologies to generate treatments for chronic diseases such as cancer. Production of images, genomics and bioinformatics, and storage, recovery, and processing of medical images require large computational resources [53], [54].

3.1.1 Impact of ICT on Healthcare and its Associated Risks

The use of ICT in healthcare no doubt has improved, and has the potential to further improve many aspects of healthcare systems, from healthcare business processes to the booking of appointments, as well as the quality and safety of healthcare. However, the increased introduction of ICT tools in healthcare could also inadvertently introduce and increase risks. It is important to be aware of these risk associated with new systems, as understanding the risks associated with the implementation of e-health systems will enable healthcare organisation to reduce the risks and increase its benefits. A report [55] has reviewed 46,349 documents to provide an excellent treatment of the quality and safety impact of ICT in healthcare.

3.1.2 ICT Penetration in the National Health Service UK

The National Health Service (NHS) is a fragmented service being operated in the four UK countries; Wales, Scotland, Northern Ireland, and England. Each of these countries has different approach to healthcare but they are all under the NHS, which provides free healthcare service. Healthcare is split into the following sectors. Primary: GPs and family doctors, Secondary: the hospital and Community. The NHS is a very busy organisation; it is believed that more than a million patients' visit their GPs everyday, 8000 are admitted to hospitals as emergency, 1.5million prescriptions are issued and over 100,000 nurse visits take place in the community. The problem is that there is no deeper insight into the data which is being collected, which means there are data but not information. The NHS covers over 60 million people in the country with massive amount of business processes and a large amount of data on activity, however, a relatively small amount of data on the clinical reasons, processes and outcome. The result is that a huge amount of data is available but it is not turned into intelligence. Recently, in July 2009, the NHS has been running the NHS Connecting for Health program which enables online patients booking. It is believed that using the program more than 16 million bookings have been made, and over 220 million prescriptions have been issued, 676,633 medical records have been transferred between GPs, and 983,152 messages on average are being sent/received daily. However, the process of creating patients' records is very slow; only 352,622 care records have been uploaded. The conservative party suggested the idea of going for a new technology for the patients' records such as Google health, the Cloud, and Microsoft health Vault. However, it is important to consider that there is a vital need to build up a knowledge base in healthcare. The knowledge base will be useful to identify how to plan new services, monitor the quality of work, distribute the resources, as well as to inform choice and research. There is a great opportunity to analyse the data to make effective decisions. For further details on this topic, please see [43], [44].

3.1.3 Electronic Data Management in Healthcare

In this subsection we discuss electronic data management in healthcare using Sweden as an example. InterSystems Corporation reports on the Swedish National Patient Summary project in [58]. Their aims lay on the implementation of a national system which enables the medical and patients' data and information to be shared among hospitals in Sweden. The report has also illustrated the issues related to data rights as well as business and public consideration. The population of Sweden is 9 millions and there are more than 20 healthcare regions. InterSystems have been given a contract with Tieto to implement healthcare system based sharing and exchanging information, locally, regionally, nationally or even international levels. The main objective is to improve the quality and consistency of the patient care by providing information to doctors that are accurate and complete. The system will ensure the improvements of the security and accessibility of the patients' information by identifying who can and cannot access such information. The project has been planned for about 10 years and this has led to a successful implementation of such system. The data is owned by the county, however; individuals have the rights to see and access their information. The healthcare system enables users to have centralised data or totally spread or join together model. The Swedish system is federated where some of data is held centrally; other data is held in the hospital HIS and produced as a medical record. There are future enhancements to the system such as educating the public about health, as well as allowing patients' access and adding more information types such as test results, allergies and so on. The system allows doctors to view all of their health issue related information such as history of their health problems and test results for the past 9 years.

3.2 Grid Computing and Healthcare

We have discussed in Section 3.1 the increasing role of ICT in the healthcare area and the opportunities and challenges that such developments entail. The next generation healthcare aims to provide an intelligent infrastructure that allows accessibility to the resources for all stakeholders and the ability to dynamically share, integrate, and analyse diverse data and resources across organisational boundaries. We have also introduced grid computing in detail in Chapter 2, including its emerging role in developing virtual infrastructure and organisations with minimum disruption through efficient integration, extension, interaction, and coordination of diverse resources transcending multiple organisations. The marriage of grid and healthcare hence is an excellent match. We discuss the opportunities and challenges of this marriage next, followed by a literature review on grid computing based healthcare in Section 3.2.2.

3.2.1 Opportunities and Challenges of Grid based Healthcare

The marriage of grid computing and healthcare has given rise to a number of opportunities and challenges. The main opportunity is the possibility of developing an intelligent collaborative virtual environment that allows integration, exploitation and analysis of heterogeneous healthcare related data at different scales. The healthcare research and industry is segmented, and the healthcare data has different dimensions and scales. For example, the data about a human body may be available in different databases at different scales: molecular and cellular, tissue, organ, and body-level. The integration and analysis of such data can reveal new correlations between various data such as genetic attributes, diseases, symptoms and medicine. Furthermore, integration of population-scale data generated from epidemiological studies or other analyses (e.g. correlation with the environment) can add to our insight leading to novel discoveries and improved, in time, intervention strategies. Grids enable easy integration, replication, sharing and management of computational and data resources within and across organisational boundaries. Grids harness the power of multiple resources to provide high performance and throughput at low cost. Grids are robust and fault-tolerant, allowing 24 hour reliable and efficient access to resources.

ICT technology is transforming the shape of businesses, governance, public services, and social networks. The domain of healthcare industry and its reliance on, and interaction with, other industries will be much broader than it has been in the past. A virtual infrastructure is needed that could support the next generation of businesses and micro industries, and allow the emergence and regulation of VOs in healthcare and otherwise. Grid based ICT (we refer to this Grid ICT from now on), in its broader sense, perhaps is the only technology that could realise the vision of the future of healthcare and other industries at relatively high efficiency and minimum disruption.

Perhaps the biggest challenge in realising a Global or national healthcare grid is security and privacy of data, and the trust in grid based ICT systems. Various healthcare related organisations – e.g. hospitals, NHS, regional councils – are reluctant to let the data flow outside the organisational boundaries. Security in grid ICT must be enhanced to make it suitable for the highly sensitive healthcare data and industries. Grid Healthcare aims integration and analysis of (private and sensitive) data, and hence new mechanisms must be developed to protect privacy of data. From industrial point of view, grid technologies need improved standardisation and stability before attractive business models can be developed. Grid ICT also requires significant developments in accounting and benchmarking areas to enforce QoS and ensure accurate billing mechanisms. In short, the socio-economics of Grid-based healthcare is an important area of consideration and a major hurdle in the uptake of grid technologies for healthcare.

3.2.2 Literature Review on Grid based Healthcare

We have defined and explained in detail in Chapter 2 that grid computing is the organisation of computer infrastructure which links various computers together and enables them to function as a larger unified system. Recently, its use has become widespread as an increasingly efficient means of providing large scale computing power on demand and inexpensively. In the 21st century, the healthcare industry has come to demand exponentially increasing numbers and sizes of high volume computer processing, especially to keep pace with advances in medical technology. Although research focused applications currently dominate medical computing needs, equally large growth is expected in the near future in applications centred around clinical diagnosis and customised treatment ordered by physicians, involving faces of genetics such as individual patient gene sequencing and genomics [59]. Applications of grid technology in the medically related fields include drug development in pharmaceutics, with the grid simulating human body systems in assessing effectiveness, drug interactions, and side effects. Imaging applications include ultrasound, CT scanning, and magnetic resonance imaging (MRI). Mass data storage and processing applications involve both uses such as genomics and bioinformatics, along with administrative purposes such as electronic medical records. The storage and retrieval of clinical data in the form of images, which have long been important for diagnostics, is one particularly promising use of grid computing given the trend toward the digitisation of images, with its attending problems of loss, storage, and access [60]. In these applications, the grid functions, among other things, to help insure increased availability and decreased loss or inaccessibility due to node failure or errors in archiving or retrieving images. Beyond this, grid computing can facilitate the sharing of data among healthcare institutions in the creation of a nationwide healthcare grid, supporting aspects as diverse as home-based and long-term care systems, research, as well as care and treatment decision modelling.

While industries as diverse as pharmaceuticals, publishing, graphic arts, and energy production and delivery have all benefited from grid computing [61], the field of healthcare has been slow to adopt this advance in technology, despite many obvious potential applications. Such areas in which the benefits of grid computing in the healthcare industry are currently significant include applications such as three dimensional modelling of either the body for surgical simulation and diagnostics or of epidemic transmission patterns, genomic research and genetically-specialised medicine, as well as large data processing operations, i.e., data mining for research [62]. Given the trend in healthcare toward curtailing government and private payer reimbursements, a linked grid of computers and servers offers a lower cost, alternative with significant savings, as opposed to high end systems while providing the same capacity to handle seemingly endless calculations and processing of data. Beyond keeping the cost per unit of applications low, grid computing can enhance the scalability and reliability of the operation. The healthcare industry in the unique situation of needing infrastructure that is ever more reliable, scalable, and secure, while at the same time accommodating rapidly advancing medical technology on a constrained budget. All of these characteristics are ones which grid computing is adept at providing.

A number of works on healthcare applications in grid computing have been reported in the literature. A great deal of research in the literature on grid based healthcare system development is focussed on collaborative and integrated middleware or workflow framework development; see [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]. For example, Koufi et. al. [63], [69], motivated by the increasingly collaborative and pervasive enterprise nature of modern healthcare systems, present a context aware access control mechanism that employs BPEL (Business Process Execution Language) to automate healthcare processes on a grid infrastructure. Their proposed mechanism acts as a mediator between mobile devices based clients and the underlying grid and allows authorised pervasive access to the integrated healthcare data. Phung et. al. [64] propose a collaborative task planning and workflow development framework for a healthcare grid with focus on elderly care enabling healthcare stakeholders to develop collaborative workflows such as for collaborative treatment. Savel et. al. [65] propose an architectural framework for a public health grid (PHGrid) enabling the community to exchange data, information and knowledge. Song et. al. [66] propose a workflow grid resource management system to support collaborative heart disease simulation applications. The focus is on workflow designs rather than on workload and capacity management. Kamal et. al. [67] propose an event based publish-subscribe grid middleware that addresses the current limitations of middleware for body sensors such as energy limitations and heterogeneity.

Calvillo et. al. [72] propose the development of a management infrastructure for virtual organisations looking into tasks such as user management (role assignments, privilege assignments etc) and definition of resource access policies. Boyd et. al. [73] in their paper entitled “The use of Public Health Grid Technology in the United States Centers for Disease Control and Prevention H1N1 Pandemic Response” report on the problems that isolated healthcare information systems have created and the work that they have developed on integrated information systems exploiting the grid technology. This is part of the research and development work on decentralised information architecture through the Public Health Grid (PHGrid) done by the US Centers for Disease Control and

Prevention's (CDC's) National Center for Public Health Informatics (NCPHI) and its partners.

Stell et. al [77] propose a system called VANGUARD, a Virtual Anonymisation Grid for Unified Access of Remote Data. Their motivation is to address problems associated with the disparate nature of data storage facilities of current healthcare systems, and the security ramifications of accessing, using, and potential misuse of that data. Their system VANGUARD supports adaptive security-oriented linkage of distributed clinical data-sets to support a variety of virtual EHRs. Hu et. al [78] propose a framework for bioprofile analysis over grid to support individualized healthcare addressing the challenges associated with distributed bioprofile analysis. The ASIDS (architecture for semantic integration of data sources) architecture is proposed by Naseer et. al [79]. They are motivated by the challenges related to sharing, integration, access and semantic interoperability of the healthcare data. Oh and Lee [80] express the need for integrating sensor networks and grid computing systems and present the design and implementation of a SensorGrid gateway to transparently connect sensor and grid networks. Further applications of grid computing in healthcare, for example, can be found in [81] (based on the analysis of healthcare data using genetic algorithms), [82] (Detection and Classification of Bacterial Contamination) and [83](data-driven Decision Support System for the acquisition and parallelised elaboration of Electroencephalography (EEG) signals).

Consequently, a number of grid based healthcare initiatives have emerged over the last 10+ years. These initiatives include, among others, the SHARE project, ACTION-Grid, ImmunoGrid, and Mammogrid. The SHARE [84][85] project was an EU Framework Programme project that aimed to identify the important milestones towards deployment and adoption of grid based healthcare systems in Europe. A number of communities have also emerged worldwide due to these initiatives, e.g. the HealthGrid community that defines HealthGrids as "Grid infrastructures comprising applications, services or middleware components that deal with the specific problems arising in the processing of biomedical data. Resources in HealthGrids are databases, computing power, medical expertise and even medical devices" [85].

It can be seen from the review of literature presented in this sections that the focus of most of the research have been on proposing grid infrastructure for healthcare applications in order to provide means for convenient, flexible collaborations across healthcare organisations; little is done on modelling the workload for healthcare grids.

3.3 The Grid Healthcare System Model

As explained in the earlier part of this thesis, medical applications are being used to enable healthcare organisations to obtain information that help them to provide the best care possible to the patients. In the proposed grid model we are examining four broad classes of healthcare applications. These are Medical Imaging and Image Processing (MIP); Electronic Health Records (EHR); Simulation Modelling, and Analysis (SMA); and Clinical Decision Support (CDS).

- ✓ MIP: Medical imaging is being used by specialists and doctors in healthcare organisations to diagnose and examine patients. Image processing provides doctors with consistent support when it comes to the analysis and treatment of the patients. Image processing of data within and across multiple healthcare organisations could enable processing and analysis of medical images, e.g. to produce new treatments and identify disease symptoms.
- ✓ EHR: is a record which holds a patient's entire health information, such as medical history, tests results, diagnoses and treatments. The use of EHR supports doctors in decision making and data management as well as generating complete reports about the patients.
- ✓ SMA: There are numerous applications of SMA within the healthcare domain, such as Genomic and epidemiological population studies.
- ✓ CDS: is an important part of the knowledge management technology used in Healthcare organisations. Therefore, doctors use CDS to support them in the medical process and on the use of knowledge available. CDS uses the data available of a specific patient to generate advice and recommendations.

Table 3-1 The Data for the Three Organisational Scenarios

MIP	EHR	SMA	CDS	CPU required	CPU available
(A_1)	(A_2)	(A_3)	(A_4)	(CPU hours)	(CPU hours)
Scenario 1 – A Hospital Grid					
3	64	2	31	145	200
Scenario 2 – A Grid connecting a Hospital and a University					
10	54	9	27	271	400
Scenario 3 – A Research Organisation Grid					
20	14	29	19	523	700

The applications discussed above are deployed in three different types of organisational scenarios. The first scenario considers a grid system to support a typical hospital; the second scenario considers an integrated grid encompassing the systems boundaries of a hospital grid together with a University Grid; and finally the third scenario relates to a grid installed to support a research organisation. Table 3-1 gives details for each of these three scenarios and a mix of applications. Columns 1-4 give the number of instances for each type of the four applications for a specific organisational scenario (i.e. Scenario 1, Scenario 2, and Scenario 3). For example, a hospital grid in Scenario 1 requires 3 instances of MIP application, 64 instances of EHR applications and so on. Column 5 gives the total CPU requirements for the mix of applications in hours of CPU power arriving each second in real-time. Column 6 lists the maximum capacity of each type of grid in CPU hours; i.e. the number of CPU hours each type of grid is able to provide per second. Note that a grid provides distributed power so an hour-CPU requirement can be fulfilled in a second or less if thousands of CPUs are available in a Grid. This is the case here too because, based on our assumptions, these grids can provide up to 200, 400, and 700 hour-CPU power per second for each specific organisation in Scenario 1, 2, and 3 respectively. The numbers of CPU hours required by each type of grid (Column 5) are calculated according to the following equation:

$$CPU\ required = A_1 * a\ +\ A_2 * b\ +\ A_3 * c\ + A_4 * d,$$

where $a = c = 10,$ and $b = d = 1.$ The variables $a, b, c,$ and d represent the number of CPU hours required for each type of applications.

We make now some notes on the data related to the medical applications and the three scenarios given in Table 3-1. The numbers of EHR and CDS applications in Scenario 1 organisation are lower than the ones for Scenario 2 and 3 organisations. Note also that the numbers of MIP and SMA jobs arriving each second are lower for Scenario 1 compared to Scenarios 2 and 3. These are very typical trends for hospital, University and research organisations, i.e., there are more EHR/CDS jobs in hospital because a hospital's primary role is to provide patient support. On the other hand, Universities and research organisations carry out SMA and MIP tasks more frequently than is the case for hospitals. Furthermore, the number of CPU hours required by the MIP and SMA applications are 10 times larger ($a = c = 10$,) than the EHR and the CDS applications ($b = d = 1$.). This is also justified by the fact that MIP and SMA applications are typically more computationally intensive than EHR and SMA. Unfortunately, we were unable to find a similar study in the literature and therefore the data formulated for this study is based on our experiences and broader knowledge of the area. However, on a positive note, our modelling study is independent of the data used herein; this study would provide useful insights into the subject using any reasonably appropriate sets of data. We intended to work with collaborators in the medical and other related professions to acquire real data for this study. However, due to the circumstances explained in Chapter 1, the focus of our research moved toward cloud computing; and we did manage to get real data from Amazon case studies; modelling of the Amazon data according to the modelling methods development in this chapter is our future work.

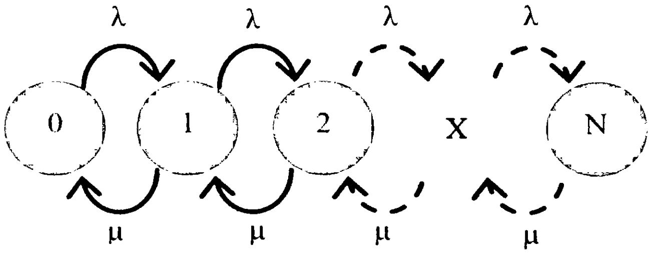


Figure 3.1 The transition diagram of the CTMC model

$$\begin{pmatrix} -145 & 145 & 0 & 0 & \cdots & \cdots \\ 200 & -345 & 145 & 0 & \cdots & \cdots \\ 0 & 200 & -345 & 145 & \cdots & \cdots \\ 0 & 0 & 200 & -345 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Figure 3.2: The generator matrix (Q) of the CTMC for Scenario 1

3.3.1 The Markov Model

We now explain our Markov model of grid systems for each of these organisational and application scenarios. Figure 3.1 gives the transition diagram of a Continuous Time Markov Chain (CTMC) model that could be applied to any of the scenarios listed in Table 3-1. Let us explain this by considering Scenario 1 (A Hospital Grid). The transition diagram shows that the model has 201 states (N+1 states; N = 200), where

$$2 < X < N$$

The initial state (leftmost) represents the idle system situation where there are zero jobs in the system to be processed, and the last state (rightmost) depicts that the system has reached its full capacity, i.e., the system is full with 200 jobs being processed in the system. The transition diagrams of Scenarios 2 and 3 will be similar except that the total number of states will be different; 201 for Scenario 1, 401 for Scenario 2 (i.e. N = 400), 701 for Scenario 3 (i.e. N=700). These numbers represent the capacity of the system; Scenario 1, 2 and 3 have a system to process 200, 400, and 700 jobs, respectively. Table 3-1 lists this data as well as gives the arrival rates (λ) and departure rates (μ) of jobs for each of the three scenarios; Scenario 1 has a λ of 145 (Column 5) and μ of 200 (Column 6), Scenario 2 has a λ of 271 (Column 5) and μ of 400 (Column 6), and Scenario 3 has a λ of 523 (Column 5) and μ of 700 (Column 6).

Figure 3.2 gives the transition rate matrix (Q) of the model diagram given in Figure 3.1 for Scenario 1 (i.e. $\lambda = 145$; $N = \mu = 200$). Similar transition rate matrices can be created for Scenario 2 (i.e. $\lambda = 271$; $N = \mu = 400$) and Scenario 3 (i.e. $\lambda = 523$; $N = \mu = 700$). Given the matrix Q, the steady state vector (π) of the system (containing probability distribution of the system to be in any of the states) can be calculated by solving the following system of linear equations:

$$\pi.Q=0.$$

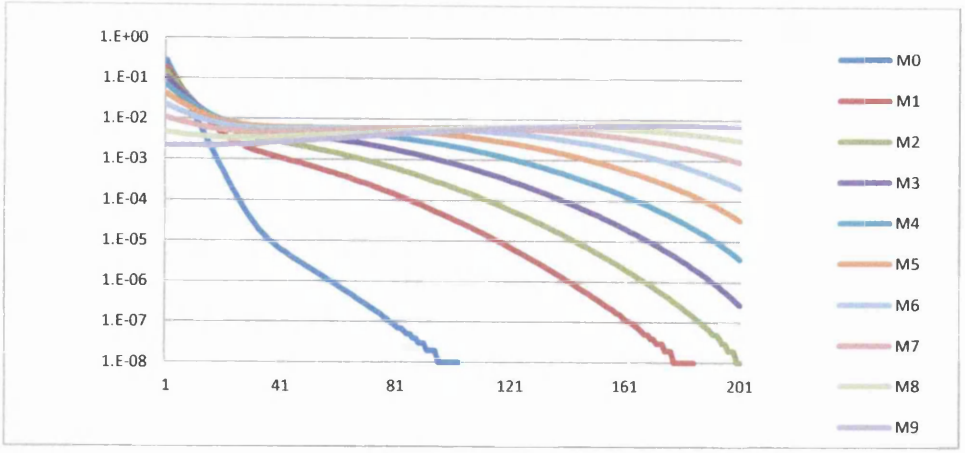


Figure 3.3 Steady state probability vector against the number of states for Scenario 1 Markov model (10 different intensity rates)

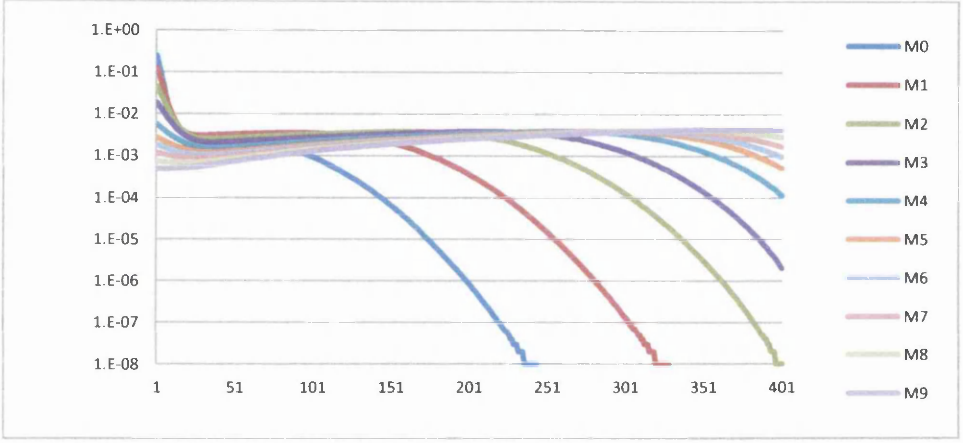


Figure 3.4 Steady state probability vector against the number of states for Scenario 2 Markov model (10 different intensity rates)

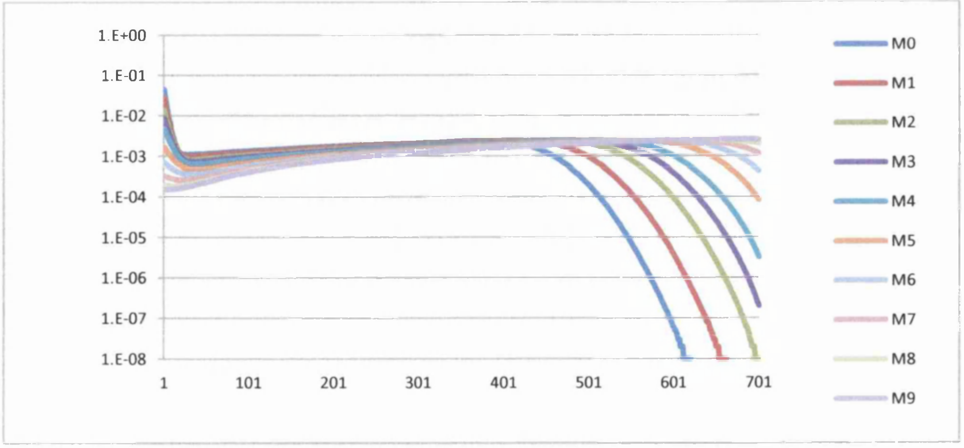


Figure 3.5 Steady state probability vector against the number of states for Scenario 3 Markov model (10 different intensity rates)

Any standard methods for the solution of systems of linear equations can be employed. Markov models arising from most real life systems are typically large and sparse, and hence iterative methods are usually employed for the system solution. We have used Gauss-Seidel method to solve the equations in our models. For further details of the solution methods for Markov Chains and some background on Markov modelling, see earlier work from the authors [86], [87] and the references therein (for example, [88], [89], [90], and [91]).

3.4 Performance Analysis

We now present results obtained by solving the linear equation systems associated with each of the three organisational scenarios presented in the previous section. Figure 3.3, Figure 3.4 and Figure 3.5 plot the respective steady state probability vectors, π, obtained for Scenarios 1-3. The steady state probability vector, π, provide the probability for the system to be in each state in the long run; the vector element $\pi[0]$ gives the probability for the system to be in state zero, $\pi[0]$ gives the probability for system to be in state one, and so on. Since most of the probability values lie near zero, we have used logarithmic scale for the y-axis. In order to explain the results, let us focus on Figure 3.3 which plots the probability vector against the number of states for Scenario 1. We have considered a total of ten arrival rates in Figure 3.3 ranging from 145 to 199, keeping the departure rate constant; each case is represented in the figure by Mx (M0 represents arrival rate λ=145, M9 represents arrival rate λ=199). The departure rate for all these plots is the same, i.e., μ=200, giving a range of 10 different intensity rates for Mx approaching almost 1 for M9. As we know from queuing theory that the intensity rate is defined by μ/λ and its value ranges between zero and one. The system becomes unstable for intensity rate exceeding 1. Basically, the relative workload for the health grids defined in the previous section (e.g. Scenario 1) will increase with the increase in the intensity rate, which in turn is directly proportional to the arrival rate λ, where, $145 \leq \lambda \leq \mu - 1$.

Note that in Figure 3.3 the probability of the system to be in one of the states varies from near 0.3 (30%) to10E-8. For M0, i.e. for the lowest intensity rate (or workload), the highest probability is for the states numbered below 10, while for the state number 100 (or so) the probability reaches almost zero (10E-8 to be precise). That is, the probability that the system will have 100 or more jobs in the system is almost zero. In this case, the system is likely to be in some of the states much below 100. The implication of this result is that the system will be stable and will operate at much below its capacity level. However, the system utilisation in this case will be quite low. For M1, the workload slightly higher than M0, the probability of the system to be in the states numbered below 10 is near 0.1 while the probability drops to the lowest for state number 180 or so. This trend is repeated as we move towards higher workloads (M2, M3, ... M9), such that the probability of the system to be in the rightmost states (nearer the state number 201) increases. For M9, the highest workload, the probability of the system to be in state 201 is the largest because the arrival and departure rate become almost equal. In short, higher arrival rates will give rise to higher probabilities of the system to be in the states nearer its total capacity which means that although the system's throughput is higher, the system is in a danger to become unstable or unreliable (because jobs may be lost as there is no queue to hold the jobs while the system is working on its full capacity, busy in processing jobs arrived earlier). Furthermore, such higher probabilities also cause higher blocking or queuing times for a system with input waiting queues (for a system where input waiting queues are not available, any jobs arriving when the system is running at its full capacity will be lost).

Figure 3.4 and Figure 3.5 present, in similar manner, the probability distributions of Markov models for Scenarios 2 and 3. As before, a total of 10 different intensity rates are used for each case. Further understanding of these figures could be obtained by considering the explanation given above for Figure 3.3.

3.5 Concluding Remarks

Next generation healthcare systems and organisations will require huge computational resources, and the ability for the various medical applications to interact and communicate with each other within and across organisational boundaries. In this Chapter, we presented a quantitative, Markovian, performance model to evaluate the suitability of computational grids for pervasive medical applications deployment in healthcare organisations. For a range and mix of medical applications, and three classes of healthcare organisations, we computed steady state probability distributions for the healthcare grid system. This study has quantitatively demonstrated the potential of computational grids for their use in the healthcare area by evaluating their performance for a range of diverse applications and organisations. The study has also shown innovative ways in which future healthcare organisations can use ICT to improve their business processes allowing them to prepare and survive in the digital economy era. However, the proposed scenarios and models in this chapter are by no means restricted to healthcare organisations and can be applied to organisations from other domains such as various government organisations requiring data/resource sharing and collaborations. The workload model and scenarios presented in this chapter gives an insight into the possibilities for resource sharing, collaborations and virtual organisations. This workload model can be used by researchers and practitioners for developing shared resources and collaborations, and for resource management of ICT systems. To further elaborate, suppose an organisation is looking to expand and improve its business processes by adding additional medical applications to its portfolio. The organisation can ask its employees and use other standard means for requirements gathering and identify a set of applications and the frequency with which these applications should be executed. These applications and range of workloads can then be modelled using the models described in this chapter. Similarly, a range of grid computing system capacities can also be modelled allowing the organisation to decide about the size of the grid to be acquired. This could be based on various business parameters including the level of efficiency required from the grid computing system and the level of risk tolerance of the organisation.

Another possibility is to model all the application workloads (for the three scenarios) discussed in this study combined together in order to model and analyse shared workload of multiple organisations. This could be used for example to design and manage a shared grid for multiple organisations under some service level agreements or for a virtual organisation (as discussed in Chapter 2). Furthermore, the modelling study described in this chapter can be used to model the workload of cloud computing providers for the design, capacity planning and resource management of cloud computing farms. Our Future work in this area will focus on improving the models, its analysis and validation by including in the model more detailed parameters, realistic applications and system related data, and by its application to cloud systems.

Chapter 4: Cloud Computing

Cloud computing is swiftly becoming a very attractive and foundational element in global enterprise computing. There are several companies across a wide range of industries which implement, develop and offer cloud technologies.

This Chapter aims to provide a detailed introduction to cloud computing. Cloud computing due to its complex nature has been a subject of much debate; Section 4.1 provides a discussion of its definitions and scope. Section 4.2 presents a history of cloud computing along with a discussion of its relationship with grid computing. Section 4.3 is an account of the various drivers of cloud computing and its current and future prospects. Section 4.4 describes its architecture while Sections 4.5 and 4.6 discuss its Service and Deployment models. Section 4.7 and 4.8 give an account of issues related to regulations and data protection for Cloud Computing. Section 4.9 gives a detailed account of migration process from traditional IT to Cloud Computing and finally Section 4.10 summarises the Chapter.

4.1 What is Cloud Computing – a Discussion

Cloud computing commonly is taken to be a phrase that basically renames common technologies and techniques that are known in IT. Cloud computing can be defined as web applications and server services that users pay for in order to access rather than software or hardware that the users buy and install themselves. Additionally, cloud computing can comprise everything from Software as a Service (SaaS) providers throughout development environment services such as structuring and building applications on the service provider's infrastructure, applications will then be delivered to the users over the Internet [92].

There have been several efforts to define the meaning of "Cloud Computing", and even though there is no comprehensive definition for the term, Buyya et al. P.44 [26] proposed the following definition: “Cloud Computing is a techno-business disruptive model of using distributed large-scale data centres either private or public or hybrid offering customers a scalable virtualised infrastructure or an abstracted set of services qualified by SLAs and charged only by the abstracted IT resources consumed"

Every organisation from IT companies to analyst firms to academics and industry practitioners has weighed in on the attempt to define the Cloud and Cloud computing. As an example of greater consensus in the defining process, Table 4-1(taken from [35]) shows how selected analyst firms have described Cloud computing.

Table 4-1: Cloud Computing Defined

Source	Definition
Gartner	*"a style of computing in which massively scalable IT-related capabilities are provided "as a service" using Internet technologies to multiple external customers"* [93]
IDC	*"an emerging IT development, deployment and delivery model, enabling real time delivery of products, services and solutions over the Internet (i.e., enabling cloud services)"* [94]
The 451 Group	*"a service model that combines a general organizing principle for IT delivery, infrastructure components, an architectural approach and an economic model – basically, a confluence of grid computing, virtualization, utility computing, hosting and software as a service (SaaS)"* [95]
Merrill Lynch	*"The idea of delivering personal (e.g., email, word processing, presentations.) and business productivity applications (e.g., sales force automation, customer service, accounting) from centralized servers"* [96]

Given the similarity of sources for the above definitions in Table 4-1, their common end-user perspective is to be anticipated. In describing cloud computing as the client experiences it, these analyst firms all stress the scalability of both applications and infrastructure as a service (IaaS). By contrast, in the scientific community there has been much less consensus over how to define cloud computing, in particular what it is versus isn't and which features are essential for a system to legitimately be called a cloud. One notable example comes from the Berkeley RAD Lab by Armbrust et al. p.1 [97], defining cloud computing in the following terms: "Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as SaaS. The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing. We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public. Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not include Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing".

One of the most significant aspects of this definition of Cloud computing is that it brings together the perspectives of both service providers and end users. For the provider, the data centre is the key component, encompassing everything from the software offered on a pay-per-use basis to the physical hardware for processing and storage. From the user's perspective and classified according to purpose, Clouds are either public or private. Meanwhile from the perspective of the IT professional, what is important is managing the integration of applications, system software, and hardware, in other words incorporating SaaS with utility computing.

In [98], the point is emphasised that while the Cloud itself is both software and infrastructure, the user sees none of the latter and needs no upfront capital outlay, but only needs to be ready to pay for what services are used as accessed through either a web service such as API, or more often by the simple, ubiquitous web browser. Foster et al., p.1 described cloud computing in [99] as:"a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet".

The above definition highlights scalability and virtualisation as central to what the Cloud is and how it achieves the tremendous advantages claimed for it. Virtualisation in what allows the cloud system to abstract both system software and the hardware it connects to, making them accessible through a predefined abstracting interface, also known as an API or a similarly functioning service. The interface serves to call up and dismiss the raw physical computing resources, i.e, the real hardware and software, applications, etc, with no delay or disruption in the client's use of the service. The interface with the user remains constant and the operation of applications continues smoothly, hidden from the user's awareness.

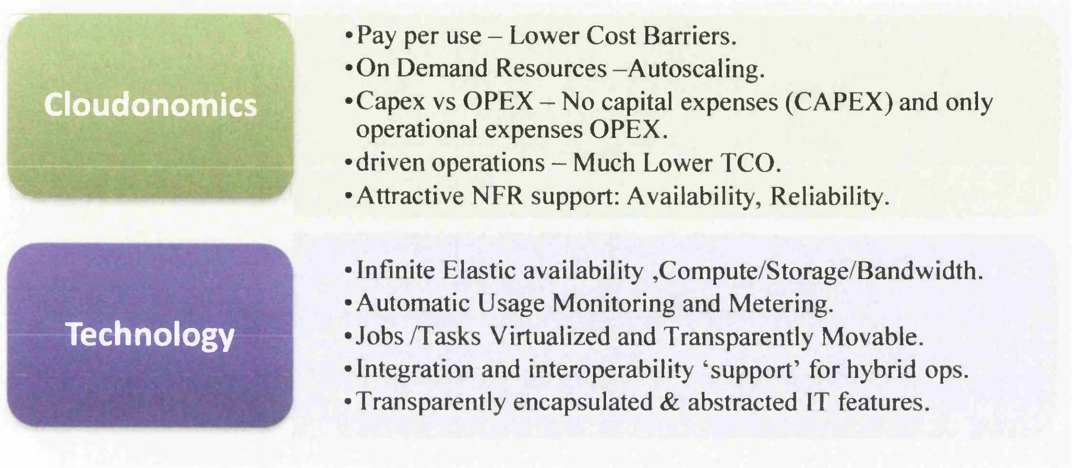
What underlies all the many and varied definitions, along with explanations, of Cloud computing is the sense that it represents a major paradigm shift in the way users access, as well as the way IT professionals deploy and provide every aspect of computing services from the fundamental infrastructure all the way through layers to the applications as the end user operates and manipulates them. For the client, what is new and fundamentally different is access over the Internet on a pay-just-for-what-you-use basis, with no thought to supporting infrastructure because an external service provider takes care of all that. Two key features of the cloud computing paradigm are scalability, which the end user may or may not notice, and virtualisation, which is normally invisible to the client. The former refers to the ability of the cloud service provider to respond almost instantly to any change in number of users, as well as changes in aggregate demand for computing capacity and data storage. Virtualisation refers to IT technology at the foundation of Cloud computing, which enables procedures such as encapsulation and abstraction in order to make the almost limitless scalability and responsiveness possible, as detailed in [100], and [99].

Cloud is defined in [101] very inclusively in order to cover all types of applications of cloud computing and services. The key characteristic of both aspects of the cloud is that it happens over the Internet or via a private network. The term cloud computing includes all components which support the computer applications, from the hardware and software to the networking functions and all services that are required for or go along with these applications. The term cloud services denote all possible services which the host, and in some respects the tenant, in a cloud system provides and all those which end users or consumers engage in. Beyond every conceivable existing or basic service, this inclusiveness extends even to any new service that the user creates by manipulating or reconfiguring that which is already provided, and to reiterate, all provided via the Internet or private network

What makes cloud computing so valuable, as well as revolutionary, to each of these users is its flexibility or elasticity in the system that enables "infrastructure shape-shifting." However, in anticipation of the major theme of [101], it is stressed that the same elasticity which gives cloud computing these advantages, also creates great challenges for security.

A very high level definition of Cloud computing is given by Menken and Blokdijk, p.8 in [102] as: " the use of computer technology that harnesses the processing power of many inter-networked computers while concealing the structure that is behind it".

It is the uniformity and simplicity of cloud computing that attracts most users, particularly those of smaller and mid-sized organisations, who also realise sizable economic benefits. In other words, the cloud is simple and easy to use. On the other hand, the economics of cloud computer services as a utility has led to the coining of the term "cloudonomics." Figure 4.1 detailed the variety of benefits which cloud computing promises to deliver to the business user through cloudonomics, as well as to the IT professional [103].



Cloudonomics

- Pay per use – Lower Cost Barriers.
- On Demand Resources –Autoscaling.
- Capex vs OPEX – No capital expenses (CAPEX) and only operational expenses OPEX.
- driven operations – Much Lower TCO.
- Attractive NFR support: Availability, Reliability.

Technology

- Infinite Elastic availability ,Compute/Storage/Bandwidth.
- Automatic Usage Monitoring and Metering.
- Jobs /Tasks Virtualized and Transparently Movable.
- Integration and interoperability 'support' for hybrid ops.
- Transparently encapsulated & abstracted IT features.

Figure 4.1: The promise of the Cloud Computing Services, adapted from [26]

There have been few trends that have developed in the growth of cloud computing. Furthermore, among these trends, early users moved to the cloud for web oriented operations and seasonally driven or fluctuating needs. Once start-up companies began opening with their IT service needs being met directly on the cloud, the pace of migration increased.

It has been noted in [104] that the advent of the cloud does not really represent any innovative technological component, but rather a new way on the part of IT professionals of thinking and organising the provision of computer services for what businesses call "the back office." The term "cloud" stems from the part of the definition referring to concealed structure. The user sees nothing of the massive scale, complexity, and movement of data which goes into making the application which the user accesses on-demand from any internet connection. A number of other circumstances of which the general public is unaware include: (a) the concept of cloud computing is not that new to the (professional) IT community; (b) already a majority of the IT industry's structural activity goes on in the cloud; and (c) there has been a slow but steady movement in the direction of businesses and organisations using the cloud propelled by the enormous savings of using (and paying for) computing services as a utility versus setting up and maintaining one's own hardware, software, and network.

What actually goes on in cloud computing on an application by application basis will be almost unnoticeable to the end user; the application is accessed from an internet browser on a remote server, to which the user pays no attention. An example of everyday computer functions would be Microsoft's Thesaurus which operates as cloud computing without users being aware of it. The implications, however, are revolutionary. Everyone receives major savings in terms of energy efficiency and pays for only what they use. The savings are most important to the smaller organisation, which can least afford excess expenditures; in fact, through the cloud they can compete with the largest counterparts on a level playing field.

For the average user at any machine that can link to the Internet, the cloud is an unseen extension, yet one with virtually limitless capabilities. Neither processing power nor data storage limits appear to exist, the only constraints being a dependable network connection and its bandwidth. Free email service with unlimited message storage is an example of what the user sees; the reduced chance of data loss due to the many networked computers is an example of the more hidden benefits.

The concept of virtualisation can be explained using electrical power as an analogy. Just as the user plugs an appliance into an outlet without regard for the source or delivery mechanisms of the current, cloud computing means running application without regard for where the hardware is, what servers power the OS and the software, or where the data is stored. In fact, in order to be virtualised, the various parts of the system must be distributed. Cloud computing actually continues the aims of cluster and grid technologies to deliver large-scale computing capacity by linking a conglomerate of resources virtualised and thereby made available to any of a number of end point users at any given time. An outgrowth of this goal is that computing comes to be seen as a utility, its service to be provided for a fee, much as electrical power is made available. This is what cloud computing actually is, with the like of Microsoft, Google, and Amazon operating like the power companies and businesses, organisations, and individuals the customers, who in this case, connect to the internet from anywhere and access whatever applications, software, and data they need, whenever they want it. Another interesting definition of cloud was proposed by Buyya et al., p.6 in [105] as: "A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on SLAs established through negotiation between the service provider and consumers".

Additionally, a number of definitions can be found in [26] from numerous sources, all of which may be summed up variations on the above definition, in combination with a report from the University of California Berkeley giving three essential characteristics of cloud computing. These essential three are: 1) the appearance to the user that computing resources are limitless, 2) the lack of pre-start up commitment, such as server and software investment, network design and set-up, etc, and 3) a payment for services model that permits the cost to vary directly with the amount of resources consumed [97].

Bypassing the hype which has accompanied the development of cloud computing and the attendant confusion as to just what the term means, the central achievement of the cloud is that it has transformed computing into a service delivering utility. By the present date, however, various IT components essential to realising the potential of cloud computing have had time to develop and mature. These will be discussed in the following sections along with integral concepts, architecture, and technical characteristics, which have influenced the development of public, private, and hybrid clouds, tools for managing IT in the cloud, and various frameworks for providing cloud services.

4.2 History of Cloud Computing – Evolution of IT Systems

The concept of cloud computing goes back to the 1960s, however, back then, it was not possible to use cloud computing due to the fact that high-speed internet was not available at that time. The cloud computing concept saw the light in the late 1990s after the huge infrastructure investments in broadband as this has made it possible for the could computing to be utilised and deployed in today's technology.

Winkle Vic (J. R.) [101], believes that cloud computing represents something of a full-circle movement back to the early days of computers, data storage, and information processing. In the 1960s, computer capability meant large rooms full of hardware. Winkle Vic (J. R.), p.10 [101] put it as "Mainframes were the epitome of control and centralisation in contrast to what followed in computing". In spite of its name, the cloud does have a physical presence, the data centre, which is described in terms of its huge scale, repeated patterns of identical computer hardware operating noisily, characteristics to which it attributes the great cost reductions of cloud computing.

The key difference between then and the cloud today, according to Majeed et al., [53], and Roth et al., [54], is that in those days mainframes represented a great inequality of opportunity and a source of power and control in the hands of a few. As cumbersome and limited as the computer capability of that era seems by today's standards, it was a tremendous advantage over the manual tasks it replaced. Still, it was quite expensive to acquire and maintain a mainframe; moreover, it required the services of trained elite, made up of those who could design the system and keep it running smoothly. Thus, the functions of the computer were available only to those with significant resources.

This very situation, which Health Informatics Unit1 (HIU1) [55], and HIU2 [56] have dubbed the "tyranny of the mainframe" is the problem that motivated the development of the microcomputer or PC, in the 1970s and '80s. The democratisation of access to computer technology and all of its benefits, along with the proliferation of small-scale systems throughout the work environment, was the result of this movement away from the large-scale mainframe driven, centralised system. Predictably, the decentralisation meant that the average employee, without computer engineering or technical expertise, could use a computer workstation analogous to the way he or she operated an automobile with only the most limited knowledge of automotive engineering and maintenance. Just as inevitable, however, was the state of affairs in which companies and organisations found themselves, having multiple department, branches, etc, using incompatible software or networking configurations, duplicating the storage of data and documents, ill-equipped to realise any saving from connection and cooperation.

One of the most significant results of all this proliferation was the accelerated development of the software industry. On one hand, the reduction in cost was making it possible for anyone or group who could afford the basic minimal level of hardware, and just a bit more for the software, to set up and starting computing. On the other hand, the variety of new possible applications of computer technology, in addition to the explosive growth of small-scale users with specific needs, created a flood of software, a significant portion of which was not well designed, and was especially vulnerable in terms of security [101].

As both the number of users and applications available to them increased, it was inevitable that desire to communicate and collaborate, in other words to network, would grow as well. Among the early perceived needs was that large numbers of employees or locations be able to access the same data base efficiently, a need which led to transaction processing systems. Additionally, Krutz & Vines [107] believe that the airline industry is a pioneer in making this possible on a world-wide scale. It was not long before the first model, in which the local point of computer access did not store or manipulate data but merely accessed it from the central database, proved inadequate. Users soon needed to not only access data but also to interact with and manipulate, in order to perform computations at the local level. This accompanied in both local-area and wide-area networks.

The new local-area-networks and their wide-area counterparts were more general and multipurpose in their orientation to user and their needs, bringing with this new orientation more connectivity, reductions in redundancy, and more user-friendly interfaces. At the same time on a less positive note, this networking brought with it more issues and points of vulnerability in relation to security concerns.

At the same time as this evolution in user access and networking was occurring, an equally fundamental change was taking place in the connection between points in the network. The strict leash-like wire cords that bound terminals to the mainframe, and later the server, were giving way first to modems, then to the Internet, and more recently to high bandwidth and even wireless transmission. However, just as remote access and data transference, as well as the resulting computing capabilities and applications were making quantum leaps forward, so too were the challenges of keeping data restricted to those with legitimate reasons to have access, as well as of preventing both mischievous and malicious attacks on the network or system. Further complicating the issue, although networking was leading to some reduction in duplication of data storage and processing, it was at the same time creating more, as well as more complex and less systematically organised infrastructure, siphoning off any cost savings. The overall effect at this period in the history of computers and IT was that, the operating costs were greater than before and the insecurity was very pervasive too [108].

From the perspective of computer applications and technology, it was not so much the Internet as it was the World Wide Web, which created explosive changes in the field.

This is what Winkle Vic (J. R.) [101] credits with fundamentally changing the human interface with information, and all that was necessary was a set of relatively uncomplicated browsers and a set of Internet Service Providers (ISPs), along with straightforward server software to make them work. Applications that had never been conceived of as anything but stand-alone were now transformed into Web applications.

Antonopoulos & Gillam [106] describe the evolution of computing, in one respect, as coming full circle from the days of mainframes to the cloud computing of today, it otherwise likens it to a spiral or to a tree branching out and growing in multiple directions. While on the surface, physical level there is similarity, as well as with the dynamic of centralisation, the power, flexibility, affordability, and democratisation in terms of user control have all made huge leaps forward with the advent of cloud computing.

The conceptual foundation of computing via the cloud is well established; the revolution surrounding it is way these concepts are being applied to change the nature of computing in businesses and organisations everywhere. One of the changes in society that has fostered this revolution is the change in the way information is disseminated and used, leading to explosive growth in the demand for ever increasing processing power. Web 2.0 is one of the outgrowths of this increase in the quest to organise and make accessible all this information in both new and traditional ways [109].

Menken & Blokdijk [102] asserts that the origins of cloud computing as a concept which grew out of the MIT computer scientist John McCarthy's notion of utility computing which conceived of computer processing as a metered service that companies could sign up for the way they do for electricity or telephone service. As far back as May 1964, in the Atlantic Monthly, Martin Greenberger author of its article “The Computers of Tomorrow” envisioned a world in which computers were ubiquitous. Around this time, IBM recognised the possibility of a huge income stream through utility computing, and got into the business of providing access to IBM's massive mainframes in something analogous to server rental for high end clients, including banks for hefty fees. The spread of the PC to business desktops, as well as homes, made utility computing an unnecessary excess for most firms, despite the significant restrictions posed by limited bandwidth and disk space. In the face of this trend, mainframe computing survived in some situations and organisations.

The concept of the computer as the network was proposed in the 1990s by Sun Microsystems and Larry Ellison, founder of Oracle, but according to , Krutz & Vines [107] failed to catch on because consumers were not interested in such a less-than-complete system.

While the personal computer was seen then and still today as having a major stand-alone dimension, the Internet was beginning to move the field of computing and IT in the direction of the cloud, by revolutionising the flow of information. Amazon was part of this nearly from the outset, amassing large server farms to handle much more than the mere retail business that the public conceives of. In fact, in the post dot-com bubble world, Amazon has retained the illusion of traffic-as-cash-flow that was the bubble, but gave it a reality in terms of the company's current operations, spurring the development of such technology as fiber optics. With their index-like search engines, Yahoo and later Google opened the path to accessing and interacting with virtually unlimited amounts of information. During this transformation, Google also began investing in enormous server farms, then a multitude of applications, including Web 2.0 while other companies trailed behind [102].

The analogy of computing becoming a utility with the advent of the cloud extends to parallel the way that the manufacturing sector shifted from creating and supplying its own electrical needs factory by factory to a model of connecting to a utility's existing power grid [110]. The IT world is experiencing a switch at present, from computing power that is generated in-house into utility-supplied computing resources which are delivered over the Internet as Web services. The following definition of computing in the cloud environment highlights the analogy and it constitutes, "On demand delivery of infrastructure, applications, and business processes in a security-rich, shared, scalable, and based computer environment over the Internet for a fee" Rappa, p.38 [20].

Grid computing has also played a critical role in the development of the modern day cloud computing paradigm (we will discuss more about it later in Section 4.2.1). We note that while hardware and system software providers as Sun and IBM have encouraged advances in grid technology, at the same time there has been an evolution in application availability. Software vendors such as Microsoft and SAP have embraced the SaaS model of computing. Both Utility Computing and SaaS are integral parts of the movement toward computing and operation of applications increasingly as a service which accesses all the resources it needs externally. It may be believed by many that SaaS evolved out of utility computing, the two however must be seen as concurrent developments, which also go hand-in-hand, but which must achieve a critical usage mass for market success to become possible. For SaaS to function so as to fulfil its promise it must have a secure foundation in terms of a readily accessible, flexible, and scalable infrastructure in which to operate. Therefore, in order to realise the potential of both, SaaS and utility computing need to be integrated seamlessly, as part of a holistic approach to computing providing a physical infrastructure that is flexible and scalable, as well as robust and reliable; platform services which access this physical infrastructure for the purpose of programming by means of abstract interfaces; and SaaS offerings that have been developed and deployed to run on this scalable, flexible infrastructure [111].

Possibly the most essential IT innovation in the development of the cloud is the process of Virtualisation. Simply explained, software 'creates' a virtual computer, which functions throughout the network in all respects to the point that one cannot tell it apart from a physical computer made of real hardware. An operating system or specifically designed program sets up an environment within an environment and in that space operates just as the traditional computer would do. It has been explained, for example, in [101] as to how the enormous quantity of physical infrastructure is virtualised, so that it may be presented over the Internet and made available on-demand to large numbers of end user, clients who will access the cloud's services at will, at fluctuating and often unpredictable times and levels of usage.

Developments in IT in the areas of hardware, Internet technology, parallel and distributed computing, and systems management, form the medium as it might be described out of which the cloud has grown. Such innovations as multi-core chips and virtualisation, Web services and SOA, grids and clusters, along with data centre automation form the necessary foundation. As these advances merged, their combined force made the birth of cloud computing inevitable. From the maturation of these developments arose the cloud [26]. In the end, what has been achieved through all these related developments in various aspects of the industry is the reliability and speed of the in-house IT network together with the cost and resource saving of merely calling up a service provided cheaply to all users on a massive scale.

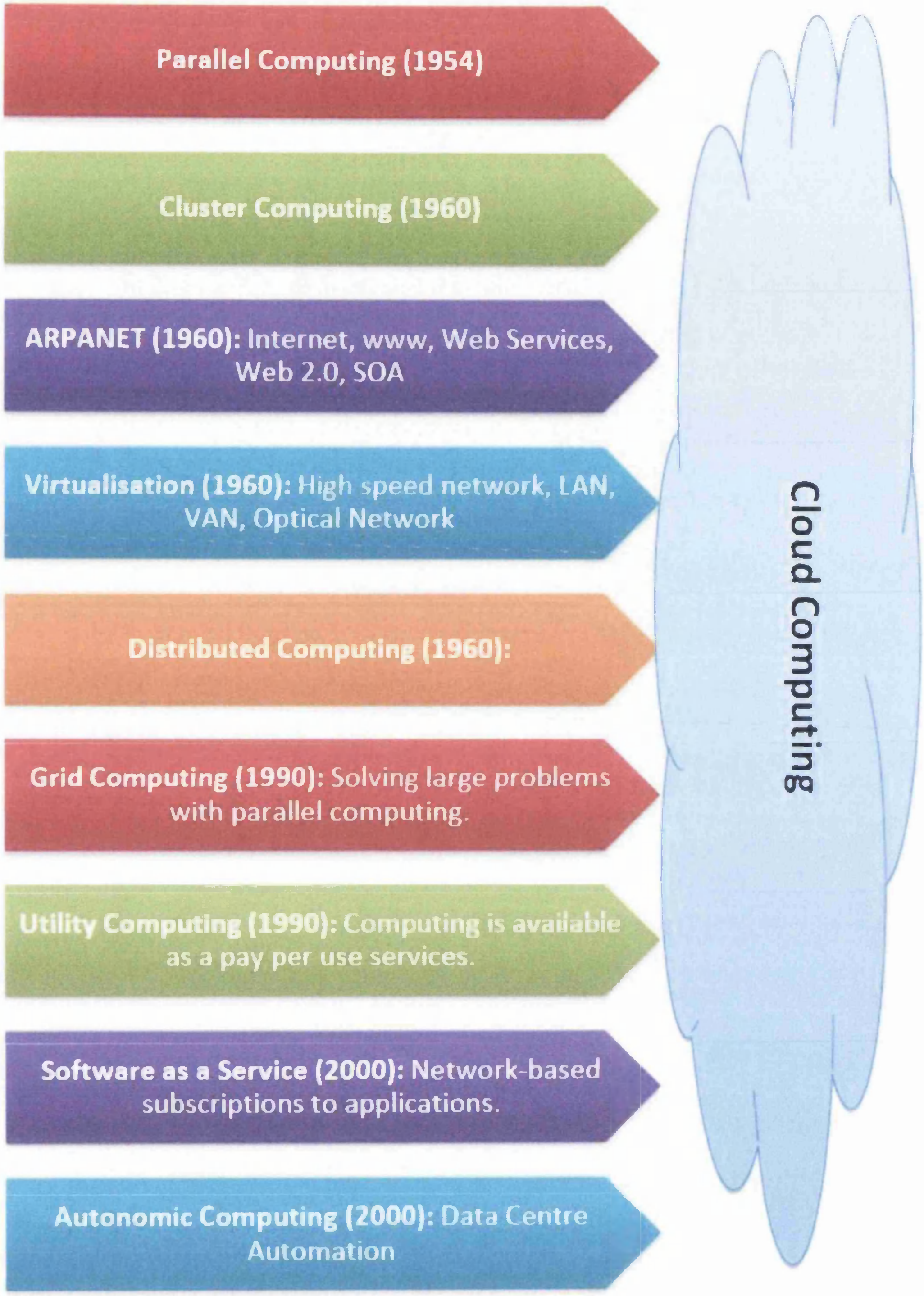

Parallel Computing (1954)
Cluster Computing (1960)
ARPANET (1960): Internet, www, Web Services, Web 2.0, SOA
Virtualisation (1960): High speed network, LAN, VAN, Optical Network
Distributed Computing (1960):
Grid Computing (1990): Solving large problems with parallel computing.
Utility Computing (1990): Computing is available as a pay per use services.
Software as a Service (2000): Network-based subscriptions to applications.
Autonomic Computing (2000): Data Centre Automation
Cloud Computing

Figure 4.2: Evolution of IT technologies and systems to Cloud Computing.

Our discussion presented up until now in this section on the evolution of IT systems and technologies to cloud computing is captured in Figure 4.2. The various horizontal blocks show the major technologies (e.g. 'Parallel Computing') and their decades of origins; these technologies, together, form the platform which enables existence and provision of cloud computing as a ubiquitous utility.

4.2.1 Grid Computing vs Cloud Computing

In the 1980s, the solution to the challenge of getting the maximum processing power was to cluster computers, hardware and all, together with special communication protocols, in effect creating one very large computer. The processing load or work was divided up between individual CPUs, a feature of no consequence to the user; data residency, the determining of where data would be held and transferred to, was crucial to making the system work efficiently. We have known in Chapter 2 that in 1990s, Foster & Kesselman [30] came up with the concept of the grid, analogous to the power grid for electricity, permitting users to gain access to computing time and resources as a metered utility to be provided by a third party in the role of utility provider and paid for as the service was consumed. Again, the obstacle was data residency. Under the grid concept, computational nodes could be anywhere on the globe; receiving data requests and making data delivery could easily lead to delays in execution and bottlenecks in the system. When the data is needed from disparate locations, the problem becomes exponentially more complex and open to problems and inefficiencies. The open source GT was designed and is being maintained by the Globus Alliance to deal with just such difficulties [29]. The distinction between grid computing and cloud computing has been discussed, for example, in [112]. As we know, the concept behind the former is that one task is subdivided into parts for each of a very large number of processors to contribute to working on a single application too large for any one computer or server to handle. The latter, cloud computing, by contrast involves many independent operations working to achieve their separate goals on the same system, the cloud. The cloud is the natural outgrowth of the grid computing concept, providing packaged services[1] which are attractive to businesses and organisations for the impressive cost saving and the ability to dispense with heavy investment in infrastructure and IT expertise.

In today's business and organisational environments, the majority of them are looking for different ways that provide them with the opportunities to reduce costs, increase storage capacity and optimise mobility. As a result, this has essentially attracted the

[1] See, for example, Amazon's Simple Storage Service (Amazon S3) [113], a simple web services

move toward cloud computing and it is continuing to be a leading infrastructure deployment theme for many business and organisations. There are many common features which are shared in grid computing and cloud computing when it comes to their goals, architecture and technology. However; both grid computing and cloud computing have their differences in many other aspects, such as the differences in architecture, security, virtualisation, applications and abstractions and concepts [18].

As we know from Chapter 2 that, based on their functionality, grid computers have two major types, one is called Computational Grids which deals with the way that computing resources from geographically distributed multiple administrative domains are shared and utilised, and the second type is called the Data Grids which deals with how the sharing and management of large amounts of distributed data is controlled. Cloud computing models also follow these two major categories in some ways. However, most of scientific computing applications nowadays require both of these features together due to the fact that they need to have extensive computational power as well as operate on large size data sets [10].

Grid computing and cloud computing share the same vision, and it is to be able to reduce the cost of computing on one hand and to be able to increase reliability and flexibility on the other hand. This can be achieved by renovating computers and the way businesses and organisations deal with them. So instead of buying and operating the computers, businesses and organisations will be able to have the computers run, managed and controlled by a third party [99], [114].

Grid computing has grown and evolved into the synthesis of diverse physical resources through the technology of virtualisation into a single integrated computing unit operating as a whole. In the context of its convergence with Service-Oriented Computing (SOC), grid computing and its attending infrastructure have come to meet the needs of application developers for a platform on which to create and deploy their work. A key part of this evolution has been the applying of the pay-per-use business model to computing, which is provided as a service [115].

Cloud computing and the cloud is the phenomenon in technology, which brings together both grid and utility computing, along with SaaS, integrating them into a unified whole to be accessed via the Web. Cloud Computing is where computational power, storage, and business applications come together, accessed externally as a service.

4.3 Cloud Computing is here to stay

There has been considerable attention and focus on Cloud computing among business analysts and in the media in recent years, with a fundamentally positive spin which has led to much public interest. An example of this is reported in [93], which predicted Cloud computing to become "no less influential than e-business."

One of the results of all this favourable publicity has been optimistic market forecasts such as recorded in [94], anticipating that expenditures on Cloud services would triple by 2012, and topping 42 billion dollars. Even the negative direction of the world economy during this period has been used as a reason for growth in Cloud computing through the increasing attractiveness of its cost savings. Other significant predictions include that Cloud computing will become the pathway to Green Information Technology.

From the user's perspective, the tremendous benefit to cloud computing is that even with a very modest terminal or "point of contact," the user can instantly access and operate the most complex of tasks, such as database indexing. Several drivers of cloud computing have been discussed by Winans & Brown in [116] in terms of information technology as organised around the following technological components: its infrastructure, its IP based networks, its virtualisation, its software, and its service interfaces. In terms of infrastructure, the cloud is revolutionary because the organisation of the servers, their storage capability, and network components give the system the ability to adjust incrementally to constantly varying needs of scale and required by the changing needs of an individual consumer or different consumers. Furthermore, IP-based networks constitute the component of cloud systems in which two potentially conflicting system goals intersect, namely the need to maximise efficiency by creating an optimised, unified network and the need to keep the multiple classes of traffic generated by different users separate, particularly to preserve security. It is believed that there are two fundamentally distinct needs in IT systems; the first is cost effectiveness coming from efficiency. This would tend to value things like multiple users sharing server resources.

On the other hand, there is the need to keep the individual computing, data storage, and processing of multiple users, separate from each other. This is the place where virtualisation plays an important role. One of the core advantages of virtualisation is that it allows the system to rapidly and efficiently put additional servers into operation or take them out again as needed. This is a key contributing factor to the immense cost savings that the cloud offers [109]. The software is what enables the cloud to function as a system. It must monitor and enforce all the features of the cloud system which guarantee the separation and isolation of data and processing needed to ensure security. At the same time, the software must direct all the operations and reallocation of resources which make the massive infrastructure elastic and efficient. The service interfaces of the cloud open up a new relationship between the service provider and the consumer which makes the cloud innovation that it is. It brings competition among providers of cloud computing services, which augments the already tremendous cost saving in terms of scale. A key component in the revolutionary nature of the Cloud is that it empowers consumers to set up automated interactions with the cloud; and it even allows users to define and manipulate their own interface [39], [110].

There are two types of users of cloud computing and services can be distinguish; tenants and end users. The tenants take advantage of the cloud capabilities by leasing part of cloud's infrastructure on an ever-changing basis as client needs change. End users, on the other hand, normally have specific, on-going applications for which they directly use cloud services; often these users are less dynamic in some of their aspects although they still benefit from elasticity of the cloud model [101].

From the point of view of the cloud service provider, the data centres require space, proximity to an electricity generation plant, the ability to conserve energy usage, and full internet accessibility. Additionally, Bohm et al. [117], explained that the Internet has traditionally been shown in IT schematics as a cloud, and as such lends its name to this innovation in computing. Since what goes on in place of the traditional in-house network (which took up significant space and resources to set-up and maintain), is amorphous and largely invisible to the user whether an individual, business, or organisation, cloud is considered to be an appropriate symbol.

Users will be aware of the fact that the applications which they activate and operate run on hosted servers somewhere, but what and where are concerns of the cloud service provider. This is in keeping with the overriding advantage of cloud computing to the customer. What's more, is that Cloud computing guarantees that organisations will be able to reduce the operational and capital costs, and in addition to that, it will importantly, allows the IT departments in the organisations focus more on planning strategic projects instead of maintaining the data centres and keeping them running [112].

According to Menken & Blokdijk [102], one of the main drivers to the technology of cloud computing is the fact that users will not be requiring a powerful computer with high specifications in order to handle complex database indexing tasks that server clusters can. The reason for that is because with the use of cloud computing and through a broadband connection, the users can without difficulty connect to the cloud, which would normally be referred to as the point of contact with the larger network. By the user of this point of contact, the users of the cloud computing from all over the world can obtain the benefits of huge processing power without the need of major capital or technical knowledge. Another major driver to cloud computing that has greatly transformed the way forward to the use of technology is the inspiration that large numbers of cheap computer hardware could be put together with the aim of creating an enormous network data centre which is as good as a smaller number of more expensive, higher quality server hardware. Therefore, the fact that to have such a huge amount of power in terms of data capacity has created a great flexibility in information, and this has never been available before.

Nowadays, there are many largest technology companies that are presently deploying this concept and providing information that can be used to help make our lives easier and more convenient. On the other hand, companies are using such concept to be more efficient and profitable. However, there are some known drawbacks to cloud computing, such as the legal and business risks of using cloud computing, especially problems that are related to the downtime and data security. In addition to that there is the complexity of managing cloud computing [92].

According to a study that was conducted by Version One [118] in 2009 41% of senior IT professionals in fact do not know what cloud computing is about and 66% of senior finance professionals are confused by the concept of cloud computing, the study was highlighting the young nature of the cloud computing technology. However, another study that was conducted in 2009 by the Aberdeen Group stated that a large number of well-organised companies achieved averagely about 18% reduction in their IT budget as a result from cloud computing and they have also achieved a reduction of 16% in data centre power costs [119].

4.4 Cloud Computing Reference Architectures

In this section, we will describe cloud computing reference architectures (CCRAs) proposed by two major cloud computing players; the National Institute of Standards and Technology (NIST) and IBM. These are discussed in Section 4.4.1 and Section 4.4.2, respectively. The non-functional properties of cloud computing will be discussed in Section 4.4.3.

A Reference Architecture (RA) provides a blueprint of a to-be-model with a well-defined scope, requirements it satisfies, and architectural decisions it realizes. By delivering best practices in a standardized, methodical way, an RA ensures consistency and quality across development and delivery projects [120, p. 0]. Other notable RAs proposed by the industry and relevant bodies include the HP CCRA [121], the Microsoft Hyper-V CCRA [122], and the CCRA proposed by the Distributed Management Task Force (DMTF) [123], [124]. It is promising to see such detailed reference architectures coming forward from major players in industry and standard bodies. We will highlight the core concepts and components in these reference architectures making comparisons between them as we move from one reference architecture to the next. We deliberately keep these discussions brief; for further details, the reader is referred to the cited reference architecture documents.

4.4.1 The NIST Cloud Computing Reference Architecture

The National Institute of Standards and Technology (NIST) have released a Cloud Computing Reference Architecture (CCRA[2]) in September 2011 [125]. The motivation to create this CCRA was to develop an independent, vendor-neutral architecture that is consistent with the NIST definition of Cloud Computing (the definition as in [126]). Its objectives are to relate different cloud services in the context of an overall cloud model, provide a reference to compare cloud infrastructures, and to facilitate the development of standards for cloud implementations. It focuses on "what" of cloud computing, and

[2] The term "CCRA" is not specific to the NIST Cloud Computing Reference Architecture. We will use it to refer to all the Reference Architectures which we have discussed this Chapter.

excludes "how to" of cloud design and implementation. The NIST CCRA is depicted in Figure 4.3. Note in the figure that NIST in its CCRA has defined five major actors; these are Cloud Provider, Cloud Carrier, Cloud Consumer, Cloud Auditor and Cloud Broker. Each of these actors plays a distinct role and performs a set of functions and activities. We explore these five actors in the following.

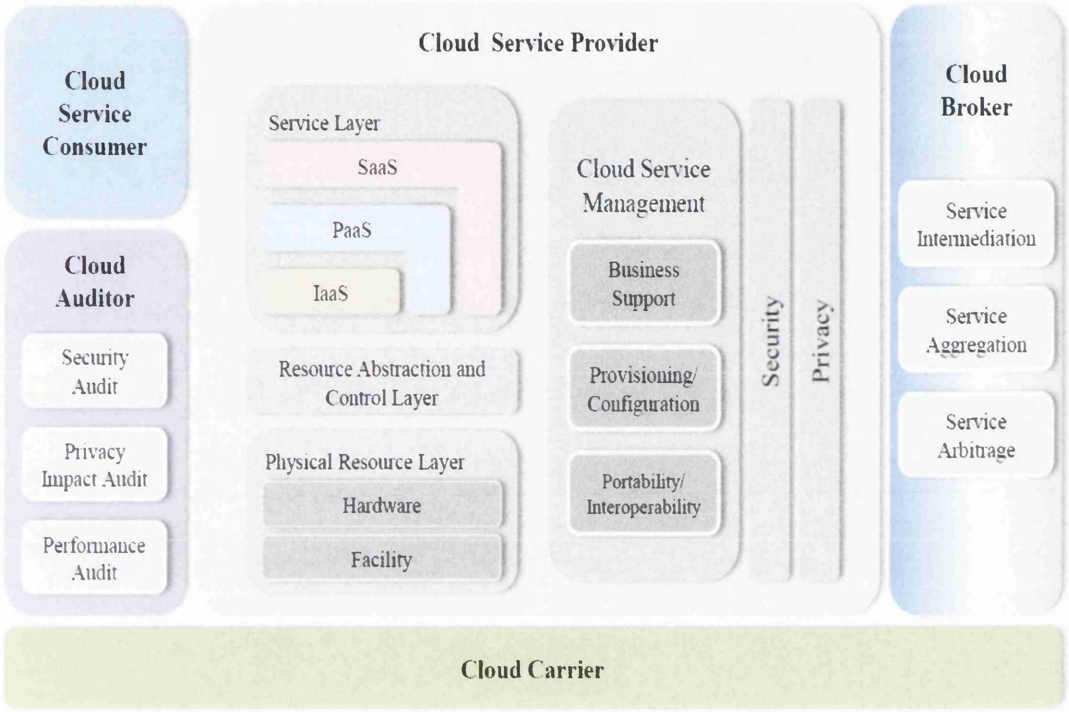


Figure 4.3: The NIST Cloud Computing Reference Architecture, [125]

4.4.1.1 Cloud Provider

Cloud Provider is defined as a "person, organisation, or entity responsible for making a service available to interested parties". Its responsibilities differ according to one of the three service models: SaaS, Platform as a Service (PaaS), and IaaS. See Figure 4.3, where these are depicted within the "Service Layer" block of the Cloud Service Provider. We will present a detailed general discussion of Cloud Computing Service models later in Section 4.5. We come back to the responsibilities of a Cloud Provider; so, for example, for IaaS, a Cloud Provider makes available to the IaaS Consumer the computing resources including the servers, storage, and networks through a set of service interfaces such as virtual machines and virtual networks. The various example services, including IaaS, PaaS, and SaaS, provided to Cloud Consumer by a Cloud Provider are depicted in Figure 4.4. Figure 4.3 depicts the five major areas in which a

Cloud Provider conducts its activities; these are service deployment, service orchestration, cloud service management, security, and privacy.

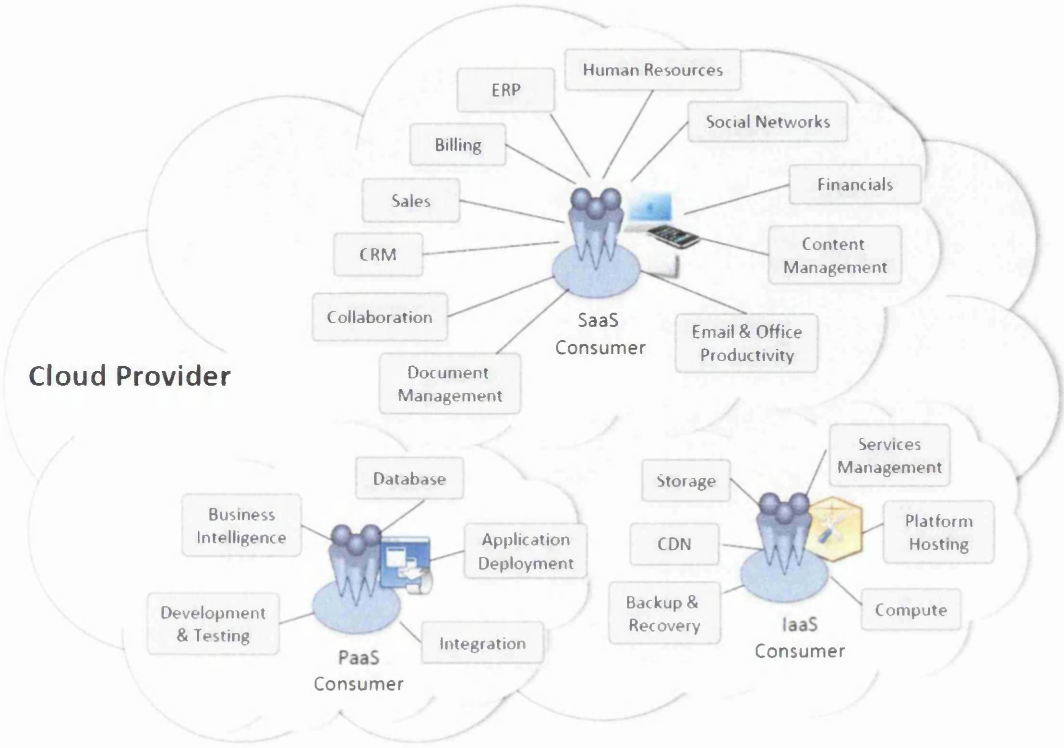


Figure 4.4: NIST Example Services Available to a Cloud Consumer (IAAS, PAAS, SAAS), [125]

4.4.1.2 Architectural Component: Service Deployment

In terms of service deployment, a Cloud can deploy its services in one of the four models; Private cloud, Community cloud, Public cloud, and Hybrid cloud. A general detailed discussion of various cloud deployment models can be found in Section 4.6.

4.4.1.3 Architectural Component: Service Orchestration

Service Orchestration is defined by Liu et al., p.15 [125] as "the composition of system components to support the Cloud Providers activities in arrangement, coordination and management of computing resources in order to provide cloud services to Cloud Consumers". The NIST CCRA uses a three-layered model for Service Orchestration (i.e. the *composition* that underlies the provisioning of cloud services). The three-layer model represents the grouping of three types of system components a Cloud Provider needs to compose to deliver its services. These three conceptual components or layers are the Service Layer, the Resource Abstraction & Control Layer, and the Physical

Resource Layer. These are depicted in Figure 4.3, in the left hand stack of the Cloud Provider block. We have already discussed the Service Layer which contains three service models (IaaS, PaaS and SaaS).

The Resource Abstraction & Control Layer controls (resource allocation, access control, and usage monitoring etc.) access to physical resources through resource abstraction (virtual machines, hypervisors, virtual data storage, virtual network etc.). The Physical Layer includes hardware resources (CPU, networks, storage etc.) as well as facility resources (heating, ventilation, air conditioning, power etc.).

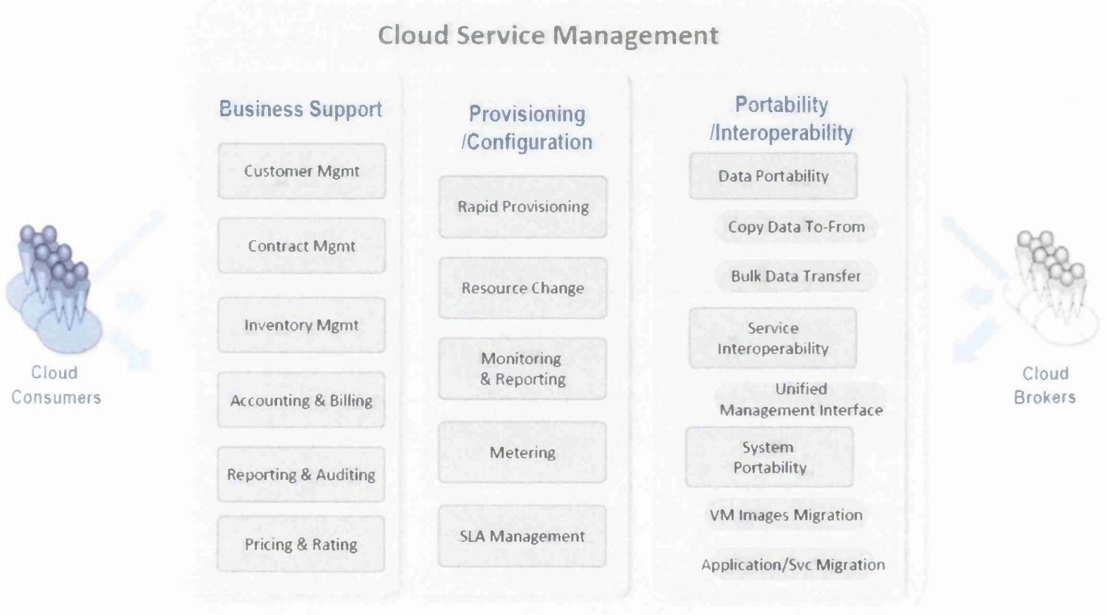


Figure 4.5: Cloud Provider - Cloud Service Management, [125]

4.4.1.4 Architectural Component: Service Management

Finally it is promising to see that the major players in Cloud computing are giving due importance to cloud service management. The Cloud Service Management block of the NIST CCRA (see Figure 4.3) scope includes all of the service-related functions that are necessary for the operation and management of services provided to cloud consumers. As depicted in Figure 4.3, a cloud provider performs service-related functions in three major areas; Business Support, Provisioning & Configuration, and Portability & Interoperability. Each of these three areas is elaborated in Figure 4.5 with more details and examples. Business Support, as depicted in the figure, includes client-facing business operations such as customer management, contract management: Inventory Management etc. Provisioning & Configuration includes rapid provisioning, monitoring and reporting, metering, SLA management etc. These terms are defined in the CCRA document [125], for example, metering involves "providing a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts)".

It is of interest for cloud consumer to know whether they can communicate as well as their data and applications can be moved across multiple clouds at low cost and minimal disruption. Therefore Cloud Providers should provide mechanisms to support data portability, service interoperability, and system portability.

4.4.1.5 Architectural Component: Security

The NIST CCRA emphasise the fact that security is not solely under the preview of the Cloud Provider but also other actors in the CCRA including Cloud Consumer. Security is stressed as a cross-cutting aspect of the architecture spanning all layers of the cloud computing reference model from physical security to application security. It is stated in the CCRA document Liu et al., p.16 [125] that "Cloud systems still need to address security requirements including authentication and authorization, availability, confidentiality, identity management, integrity, audit, security monitoring, incident response, and security policy management". The CCRA continues to discuss the implications of various architectural components on security including the three service models and the four deployment models. It further stresses on the fact that security in cloud environments is a shared responsibility for all actors including Cloud Providers and Cloud Consumers.

4.4.1.6 Architectural Component: Privacy

Cloud computing provides a flexible solution for shared resources, software and information; however, it poses additional privacy challenges to consumers using the clouds [125]. Privacy is stressed upon by NIST in its CCRA by citing an excerpt from the Privacy Management Reference Model (PMRM) Technical Committee of Organization for the Advancement of Structured Information Standards (OASIS), "Cloud providers should protect the assured, proper, and consistent collection, processing, communication, use and disposition of Personal Information (PI) and Personally Identifiable Information (PII) in the cloud" [127].

4.4.1.7 Cloud Carrier

Cloud Carrier is defined in the NIST CCRA [125] as "an intermediary that provides connectivity and transport of cloud services from Cloud Providers to Cloud Consumers". It provides access to Cloud Consumers through networked devices such as PCs, laptops, mobile devices etc). The distribution of cloud services could be through various networks, telecommunication carriers or transport agents, where a transport agent is a business organization that provides physical transport of storage media such as high-capacity hard drives. A Cloud Provider may setup SLAs with a Cloud Carrier requiring the Cloud Carrier to provide dedicated and/or secure/encrypted connection between a Cloud Consumer and Cloud Provider.

4.4.1.8 Cloud Consumer

Cloud Consumer is defined by NIST in its CCRA [125] as a "person or organization that maintains a business relationship with, and uses service from, Cloud Providers". Cloud Consumers are categorised into three groups based on their requirements, these are indeed according to the three service models; IaaS, PaaS, and SaaS Cloud Consumer. We have already depicted in Figure 4.4 the various example services available to a Cloud Consumer based on the three service models.

4.4.1.9 Cloud Auditor

A Cloud Auditor is defined in the CCRA as "a party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation". A Cloud Auditor "can evaluate the services provided by a cloud provider in terms of security controls, privacy impact, performance, etc". As stated in the document, "a privacy impact audit can help Federal agencies comply with applicable privacy laws and regulations governing an individual's privacy, and to ensure confidentiality, integrity, and availability of an individual's personal information at every stage of development and operation".

4.4.1.10 Cloud Broker

The NIST CCRA defines Cloud Broker as "an entity that manages the use, performance and delivery of cloud services and negotiates relationships between cloud providers and cloud consumers". It is envisaged that integration of cloud services in the future can be too complex for Cloud Consumers to manage. A Cloud Broker can address these problems and can bring innovation through Service Intermediation (improvements in some specific capability of a service and provision of value added services), Service Aggregation (integration of multiple services into one or more new services), and Service Arbitrage (integration of services from multiple agencies). Cloud Broker is the only optional actor in the NIST CCRA.

4.4.1.11 Cloud Computing Scenarios

The NIST CCRA document provides a number of interesting usage scenarios to illustrate the concepts of various actors in the reference architecture. A usage scenario depicts a Cloud Consumer to request a service directly from a Cloud Broker who provides new services by combining multiple services from a Cloud Provider. In another scenario, a Cloud Provider enters into two unique SLAs to arrange dedicated and encrypted services for a Cloud Consumer through a Cloud Carrier. A third usage scenario depicts a Cloud Auditor conducting independent assessment of the operation and security of the cloud service implementation possibly involving both Cloud Provider and Cloud Consumer.

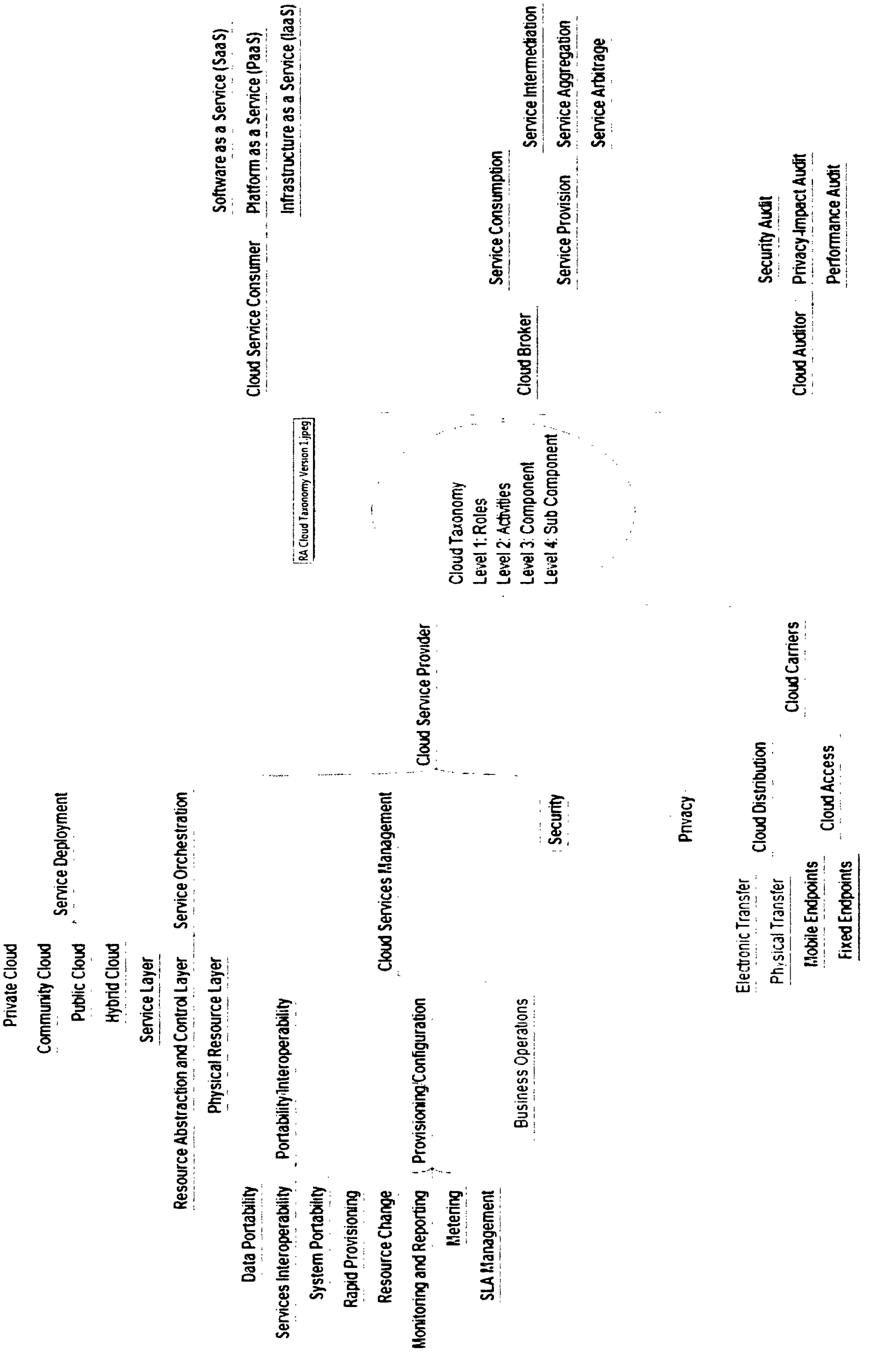


Figure 4.6: NIST Cloud Reference Architecture Taxonomy, [125]

4.4.1.12 Cloud Taxonomy

Finally, the NIST CCRA document provides a four-level Cloud Taxonomy associated with the CCRA (see Figure 4.6 which is taken from [125]). The four levels are:

1. Role: set of obligations and behaviours as conceptualized by the associated five Actors
   - ✓ e.g. Cloud Provider
2. Activity: the general behaviours or tasks associated to a specific role
   - ✓ e.g. Cloud Service Management
3. Component: the specific processes, actions, or tasks that must be performed to meet the objective of a specific activity
   - ✓ e.g. Provisioning/Configuration
4. Sub-component: a modular part of a component
   - ✓ e.g. SLA Management

4.4.2 The IBM Reference Architecture

IBM has released v2.0 of its CCRA in February 2011 [120]; the CCRA was submitted to the Cloud Architecture Project of the Open Group[3]. The mission for IBM CCRA is to provide a definition of a single Reference Architecture for Cloud Computing so to enable cloud scale economies in delivering services with resource optimisation and delivery of a design blueprint for workload optimised cloud services and projects managed by a common management platform. The IBM CCRA is based on aggregate experience gained from the real-world cloud computing projects as well as general knowledge of IBM systems, software and services. It consists of 21 documents that provide information ranging from the overview of the fundamental architectural building blocks of Cloud architecture to the definition of the architectural principles providing guidance for creating a range of cloud environments.

[3] http://www.opengroup.org/

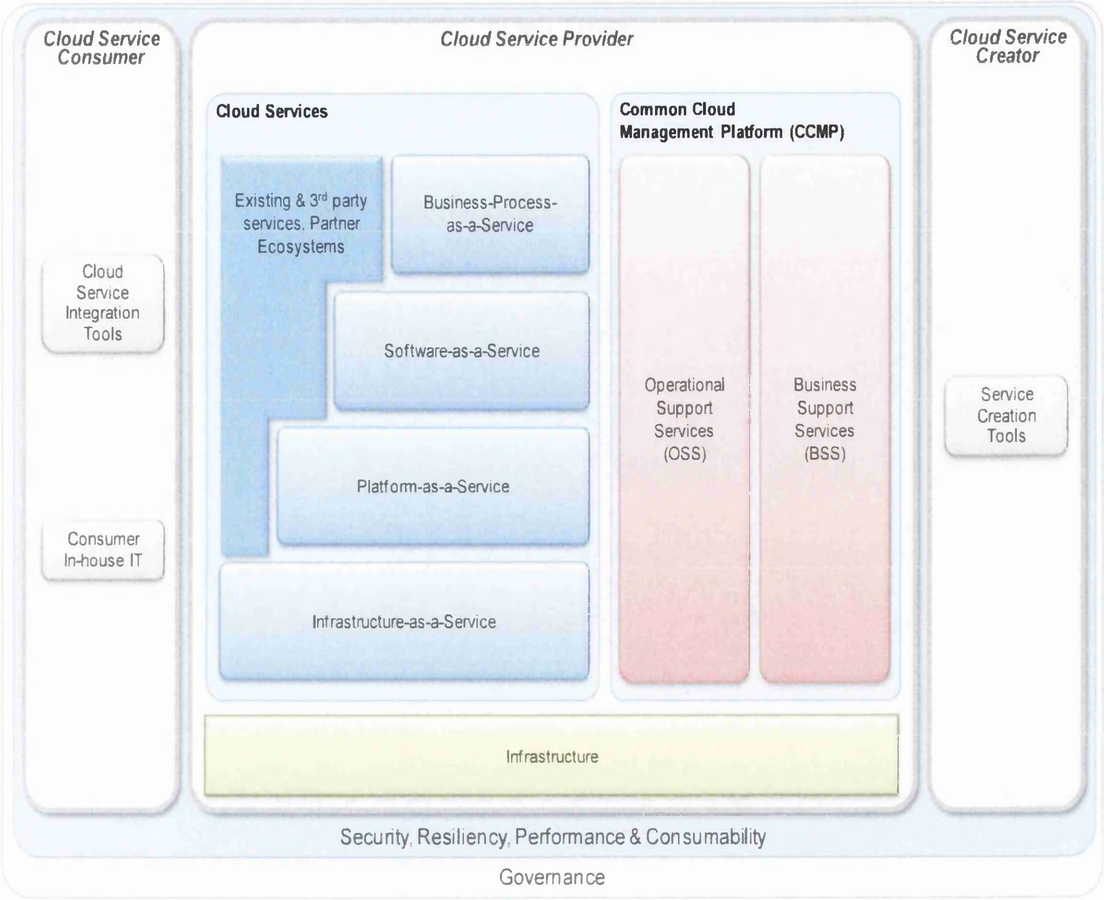

Cloud Service Consumer
Cloud Service Provider
Cloud Service Creator
Cloud Services
Common Cloud Management Platform (CCMP)
Existing & 3rd party services, Partner Ecosystems
Business-Process-as-a-Service
Software-as-a-Service
Platform-as-a-Service
Infrastructure-as-a-Service
Operational Support Services (OSS)
Business Support Services (BSS)
Cloud Service Integration Tools
Consumer In-house IT
Service Creation Tools
Infrastructure
Security, Resiliency, Performance & Consumability
Governance

Figure 4.7: The IBM Cloud Computing Reference Architecture (CCRA) – the overview, [120]

The IBM CCRA is not a detailed deployment specification of a single specific cloud implementation; rather the intention is to use it as a blueprint for designing and developing cloud implementations. The CCRA document [120] also explains the relationship between SOA (Service Oriented Architectures) and Cloud Computing making notes of the areas where SOA knowledge and standards can be applied to Cloud. The IBM CCRA is shown in Figure 4.7. It depicts a higher abstraction view of the fundamental architectural components that constitute a Cloud.

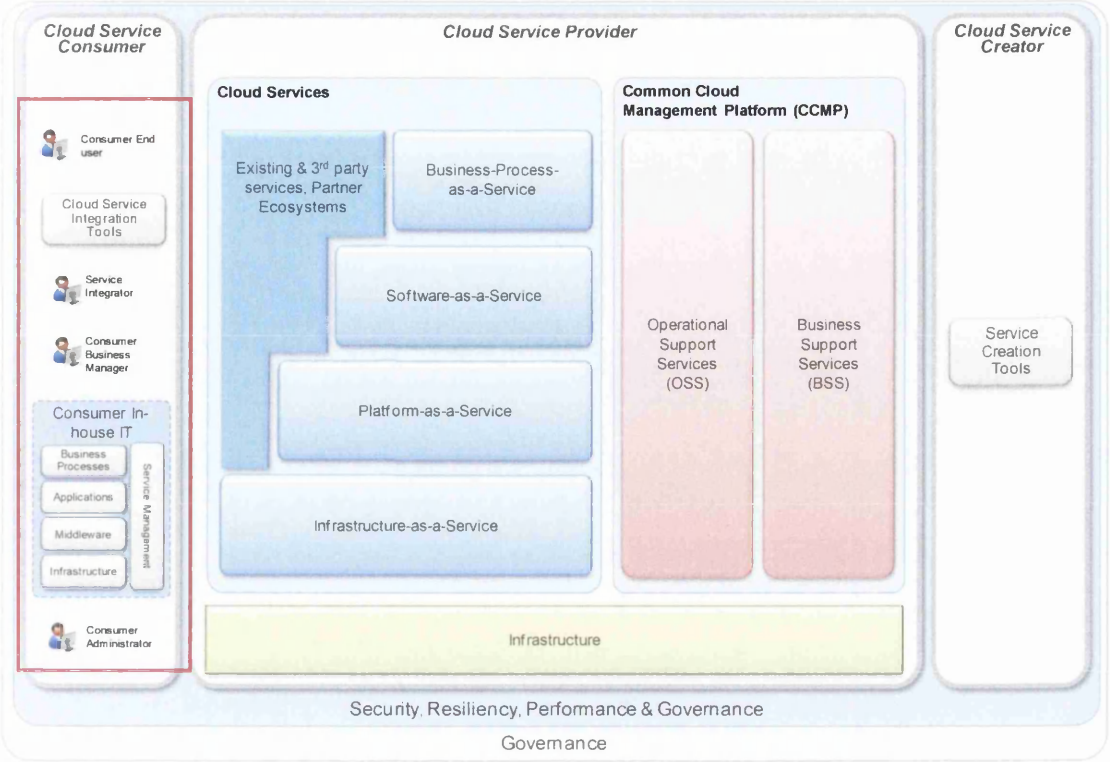


Figure 4.8: The IBM CCRA – Cloud Service Consumer, [120]

The CCRA is modular and the figure does not include details of the architectural components. The details of a few major components are included in Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11, and Figure 4.12.

As in the NIST CCRA, the IBM architecture also includes Cloud Computing Actors or Roles. However, in contrast to the NIST approach, the IBM architecture includes three Actors. These are:

1. Cloud Service Consumer (the block on the left in Figure 4.7)
2. Cloud Service Provider (middle), and
3. Cloud Service Creator (right).

Note that the IBM CCRA only defines 3 roles as compared to the NIST architecture; Cloud Auditor, Cloud Broker, and Cloud Carrier are absent from IBM CCRA. Note also that Cloud Service Creator role was not used explicitly by the NIST CCRA discussed in Section 4.4.1, rather, in NIST CCRA, the Service Creator role was fulfilled by Cloud Provider and Cloud Broker.

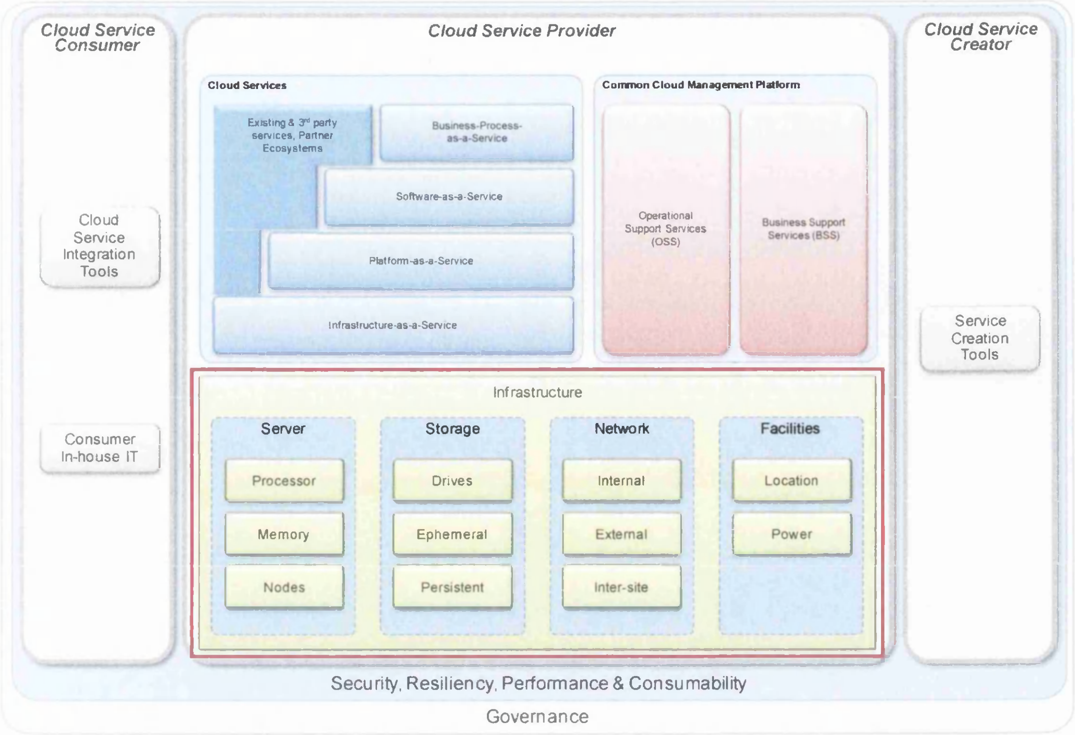


Figure 4.9: The IBM CCRA – the Infrastructure Components, [120]

A Cloud Service Consumer is defined in the CCRA as "an organisation, a human being or an IT system that consumes (i.e., requests, uses and manages, e.g. changes quotas for users, changes CPU capacity assigned to a VM, increases maximum number of seats for a web conferencing cloud service) service instances delivered by a particular cloud service.". It is expected that a Cloud Service Consumer may also have in-house IT to meet some specific requirements. In such cases, the consumer would require Cloud Service Integration Tools, allowing the consumer to integrate cloud services with their in-house IT. These service integration tools will provide support for integration on all layers of the technology stack (infrastructure, middleware, applications, business processes, and service management). This is due to the fact that the consumer in-house IT could take place on any of these layers. This sounds very helpful, in particular in hybrid cloud environments where seamless integrated management, usage and interoperability of cloud services in integration with in-house IT is important. The details of the Cloud Service Consumer role are highlighted in Figure 4.8.

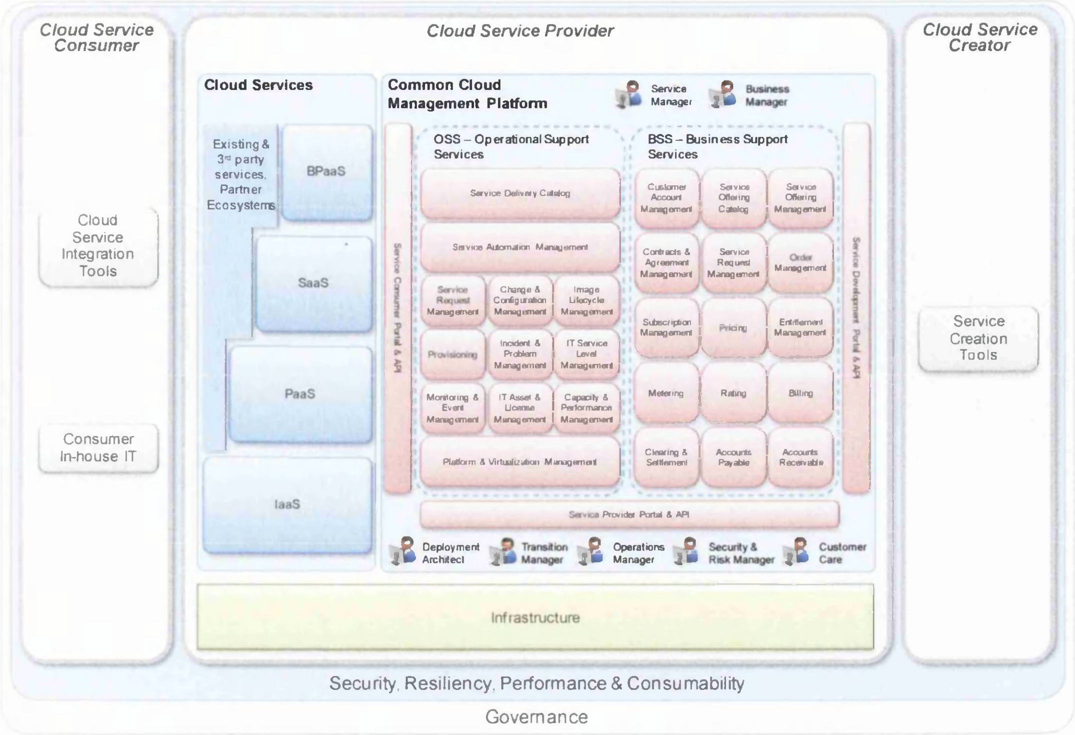

Cloud Service Consumer
Cloud Service Provider
Cloud Service Creator
Cloud Service Integration Tools
Consumer In-house IT
Cloud Services
Existing & 3rd party services, Partner Ecosystems
BPaaS
SaaS
PaaS
IaaS
Common Cloud Management Platform
Service Manager
Business Manager
OSS – Operational Support Services
BSS – Business Support Services
Service Consumer Portal & API
Service Delivery Catalog
Service Automation Management
Service Request Management
Change & Configuration Management
Image Lifecycle Management
Provisioning
Incident & Problem Management
IT Service Level Management
Monitoring & Event Management
IT Asset & License Management
Capacity & Performance Management
Platform & Virtualization Management
Customer Account Management
Service Offering Catalog
Service Offering Management
Contracts & Agreement Management
Service Request Management
Order Management
Subscription Management
Pricing
Entitlement Management
Metering
Rating
Billing
Clearing & Settlement
Accounts Payable
Accounts Receivable
Service Development Portal & API
Service Provider Portal & API
Deployment Architect
Transition Manager
Operations Manager
Security & Risk Manager
Customer Care
Service Creation Tools
Infrastructure
Security, Resiliency, Performance & Consumability
Governance

Figure 4.10: The IBM CCRA – the Common Cloud Management Platform (CCMP), [120]

The Cloud Service Provider in the IBM CCRA is composed of three major architectural sub-components. These are Cloud Services, Infrastructure (see Figure 4.9) and Common Cloud Management Platform (CCMP) (see Figure 4.10). Note in Figure 4.8 (other IBM CCRA figures also contain the same information) that the IBM CCRA defines four types of service models (in contrast to NIST which defines three). Business Process as a Service (BPaaS) is the additional forth Cloud Service model, obviously IBM has many services in this domain. We will discuss Cloud Service models in general in Section 4.5.

The Infrastructure architectural component of the IBM CCRA is detailed in Figure 4.9. It includes servers (CPU, memory, notes etc), storage, network (internal, external and inter-site), and facilities (location, power etc). The infrastructure is managed by the Operational Support Services (OSS) as part of the CCMP which we discuss next. Note that the CCMP itself runs on the infrastructure.

The Common Cloud Management Platform is a general purpose cloud management platform to provide management of cloud services across all four cloud service models; see Figure 4.10. The CCMP also supports virtualisation of cloud services on any level; hypervisor, operating system, platform or application level. It comprises three portals;

Service Consumer Portal, Service Provider Portal, and Service Development Portal. A portal provides an interface and although these are detailed separately, all three can be realised in a single implementation with different access rights.

The CCMP is composed of two main sub-components the OSS, and the Business Support Services (BSS). BSS "represents the set of business-related services exposed by the CCMP, which are needed by Cloud Service Creators to implement a cloud service". OSS "represents the set of operational management / technical-related services exposed by the CCMP, which are needed by Cloud Service Creators to implement a cloud service".

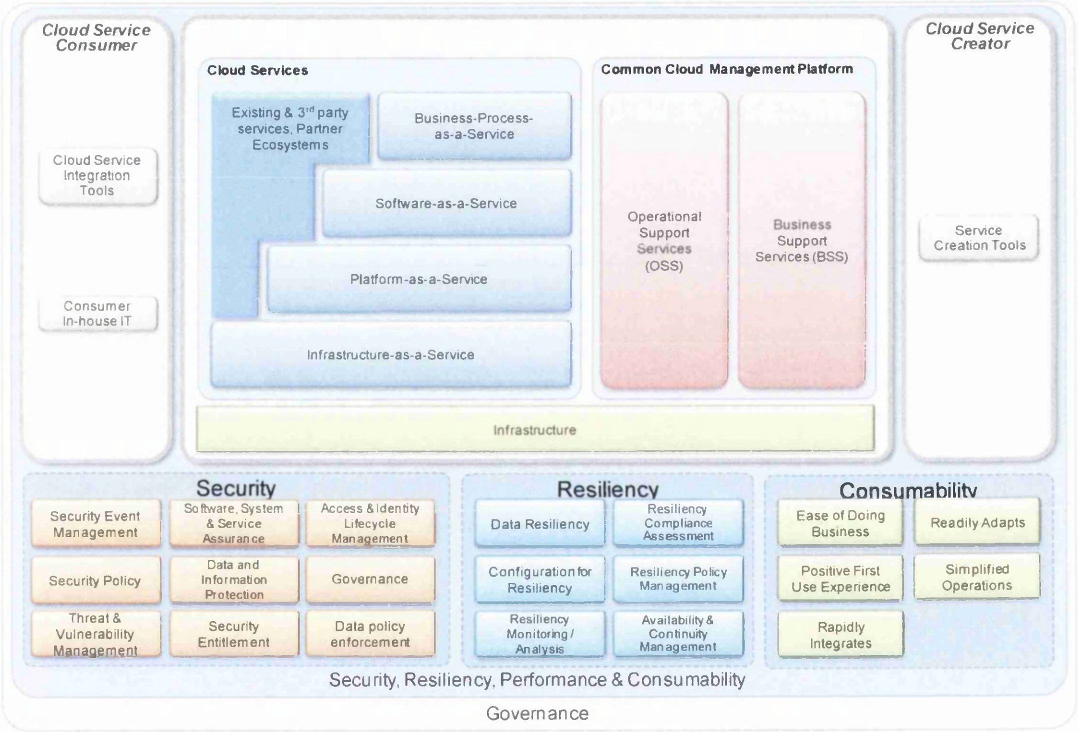

Cloud Service Consumer
Cloud Service Integration Tools
Consumer In-house IT
Cloud Services
Existing & 3rd party services, Partner Ecosystems
Business-Process-as-a-Service
Software-as-a-Service
Platform-as-a-Service
Infrastructure-as-a-Service
Common Cloud Management Platform
Operational Support Services (OSS)
Business Support Services (BSS)
Infrastructure
Cloud Service Creator
Service Creation Tools
Security
Security Event Management
Software, System & Service Assurance
Access & Identity Lifecycle Management
Security Policy
Data and Information Protection
Governance
Threat & Vulnerability Management
Security Entitlement
Data policy enforcement
Resiliency
Data Resiliency
Resiliency Compliance Assessment
Configuration for Resiliency
Resiliency Policy Management
Resiliency Monitoring / Analysis
Availability & Continuity Management
Consumability
Ease of Doing Business
Readily Adapts
Positive First Use Experience
Simplified Operations
Rapidly Integrates
Security, Resiliency, Performance & Consumability
Governance

Figure 4.11: The IBM CCRA - Security, Resiliency, and Consumability Components, [120]

Security, Resiliency, Performance, and Consumability (see Figure 4.11) are cross-cutting, end-to-end, non-functional attributes of a cloud environment and therefore these span across infrastructure, cloud services, CCMP, as well as Cloud Service Consumer and Cloud Service Creator.

Cloud Service Creator, finally, is detailed in Figure 4.12. It includes Service Creations Tools (Service Management Development Tools, Service Runtime Development Tools etc), Service Component Developer, Service Composer and Offering Manager. These tools allow design and development of new cloud services.

IBM devoted a whole chapter in the CCRA document [120] to the rrchitectural Principles and related guidance. The architectural principles include the Efficiency Principle, Lightweightness Principle, Economies-of-scale principle, and the genericity principle.

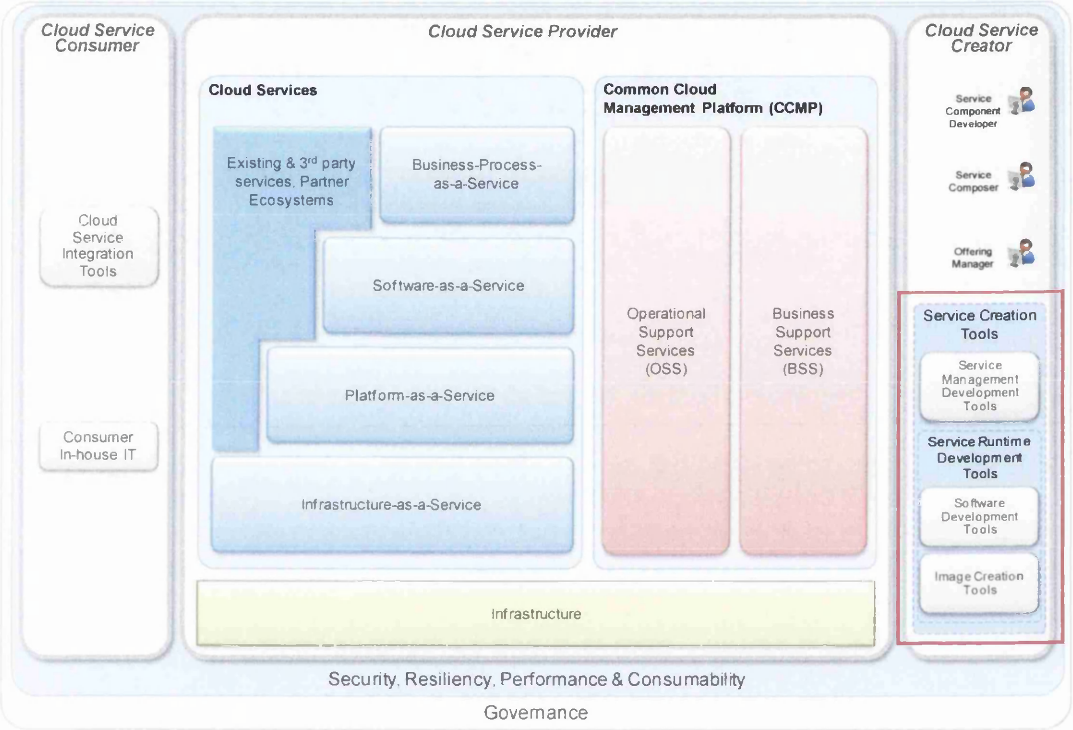

Cloud Service Consumer
Cloud Service Integration Tools
Consumer In-house IT
Cloud Service Provider
Cloud Services
Existing & 3rd party services, Partner Ecosystems
Business-Process-as-a-Service
Software-as-a-Service
Platform-as-a-Service
Infrastructure-as-a-Service
Common Cloud Management Platform (CCMP)
Operational Support Services (OSS)
Business Support Services (BSS)
Infrastructure
Cloud Service Creator
Service Component Developer
Service Composer
Offering Manager
Service Creation Tools
Service Management Development Tools
Service Runtime Development Tools
Software Development Tools
Image Creation Tools
Security, Resiliency, Performance & Consumability
Governance

Figure 4.12: The IBM CCRA – the Service Creation Components, [120]

4.4.3 Non-functional Architectural Properties

We discuss below the five core non-functional attributes as defined by the National Institute of Science and Technology (NIST) [126]. These are on-demand self-service, resource pooling, rapid elasticity, measured service, and broad network access. The idea of being on-demand is the first critical element of the cloud's overall revolutionary impact. By accessing the cloud, the end user can access whatever applications and perform whatever operations he or she wants at any time. Moreover, users can do this without making any prior arrangements to have the required IT components ready and waiting for the users at a desired time. The self-service aspect of this characteristic means that, just as server time, storage space, network connection, etc are instantly available, they are moreover available without the need for human interaction. The second essential prerequisite property of the cloud is Resource Pooling, i.e. the pooling of hardware, software, server time, network connection, data storage space, and all the other IT resources. They are shared among the many clients of the provider who in the process achieves both the highly enhanced capability described above and phenomenal cost savings for all at the same time. One fundamental consequence of this resource pooling is that the client in some respects loses strict control over data in particular, and even knowledge of the exact location, virtual and physical, of operations and storage. However, the constraints imposed by the customer on broad location of data are enforced by the cloud computing vendor.

Cloud architecture must also enable Broad Network Access. In accessing the cloud, the client or end user is able to put into operation any of the capabilities which are provided in that particular cloud, whether they are SaaS, PaaS, or IaaS. Furthermore, and most critically for this characteristic of the cloud, the user is able to do so any standard means of connection in common use, examples of which include smart phones, laptops, and personal digital assistants. Furthermore, a Cloud should be able to demonstrate Rapid Elasticity. The cloud only achieves its advantages by virtue of its ability to respond automatically and virtually instantaneously to the changing needs of any one client, as well as a rapid onset or drop off in the aggregate needs of its multi-tenant/end user base of consumers, with resources scaling up or down, coming in and out as needed. Resources and capabilities of the cloud should appear limitlessly available from the consumer perspective. And finally, a Cloud should be able to provide Measured Services. The level of abstraction needed to achieve measured service will be determined by the type of service involved; however, all forms of client usage in the cloud must be monitored, metered, and controlled. All this observation and measuring is needed so that transparency may be guaranteed. This guarantee, in turn, insures that both the end user and provider will know exactly the quantity and types of service consumed.

4.5 Service Models of Cloud Computing

It is believed that the foundation of cloud computing was a result of a set of many pre-existing and researched concepts. Such concepts are known to be distributed and grid computing, virtualisation or SaaS. Even though, it is very obvious that several concepts do not come into view to be new, however; the true innovation and improvement of cloud computing is seen in the way its computing services are being provided to the customers. There is a wide range of business models which have developed in modern times with the aim to provide services on different levels of concepts and abstraction. Additionally, such services comprise the ability to offer software applications, programming platforms, data-storage or computing infrastructure.

Within the range of services that fall under the heading of Cloud computing, three distinct layers or levels (also referred to in various sources as types, shapes, styles, or segments) can be differentiated based on definitions given in [97], and while terminology may vary, classification of service into these three layers has become standard, as noted in [100], and [128]. The term layers is preferred here, especially over a discussion framed in terms of capabilities, as most logically capturing the relationships among of the three in terms of the architecture of the Cloud. Either way, fundamentally Cloud Computing is about providing IT capabilities and resources in a manner illustrated in Figure 4.13 according to integration and with real world examples. The three layers are uniformly known as SaaS, PaaS, and IaaS.

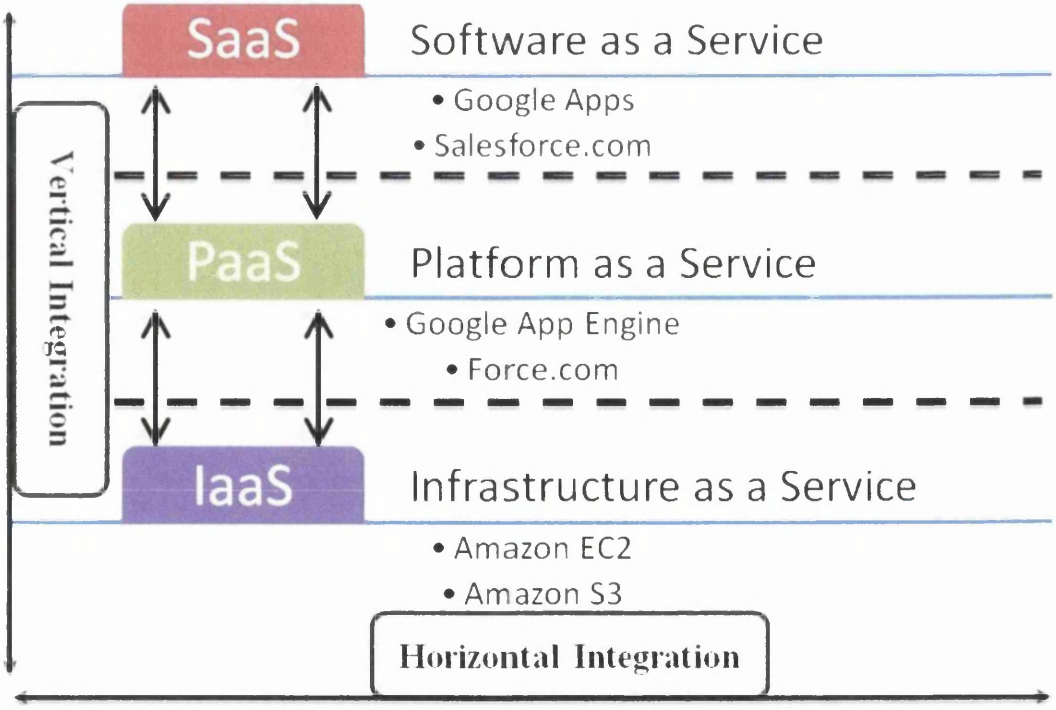


Figure 4.13: Illustration of Cloud Computing Layers, adapted from [18]

Three commonly recognised levels of service provided in the Cloud are SaaS, PaaS, and IaaS, which may be distinguished from each other by either or both of two criteria. The first is the level of abstraction that the services provided represent; the second is as layer of integrated architecture [126]. An illustration which blends these two is portrayed in Figure 4.14

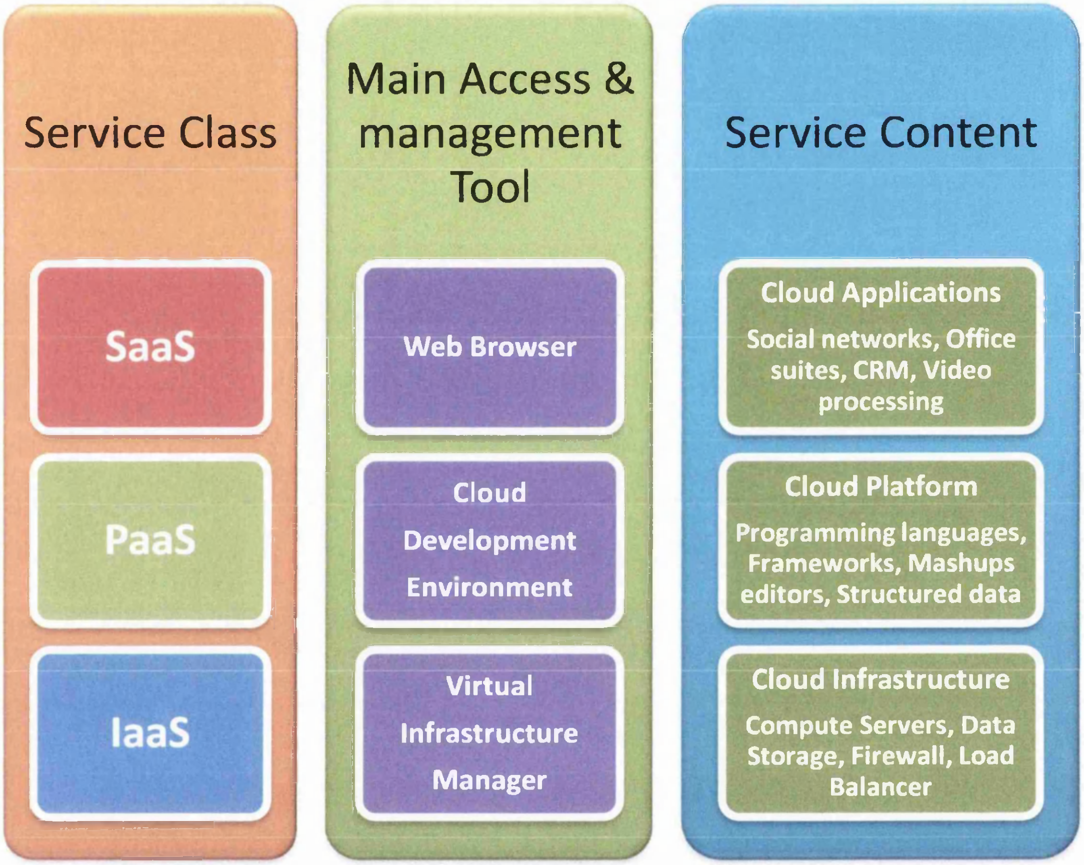

Service Class
SaaS
PaaS
IaaS
Main Access & management Tool
Web Browser
Cloud Development Environment
Virtual Infrastructure Manager
Service Content
Cloud Applications
Social networks, Office suites, CRM, Video processing
Cloud Platform
Programming languages, Frameworks, Mashups editors, Structured data
Cloud Infrastructure
Compute Servers, Data Storage, Firewall, Load Balancer

Figure 4.14: Illustration of Cloud Computing Stack, adapted from [26]

Of the three, IaaS is the least abstracted, meaning that users are most connected to the detailed aspects of creating and deploying applications. When described in terms of layered architecture, IaaS represented the foundation level of cloud service, with the other two resting on it, as in Figure 4.14.

According to Menken & Blokdijk [101] the three components of cloud service which are IaaS, PaaS, and SaaS collectively referred to as the SPI model of cloud service delivery. One of the significant advantages of cloud computing is the flexibility in meeting the client's needs in terms of service structure.

The attractiveness of cloud computing for those running both the financial and technology aspects of any business is due to its cost reductions coupled with its ease of use. The foundation of these advantages is the large scale provision of service to many clients simultaneously by major cloud service vendors, such as Google, Amazon, Microsoft, and a few others. Moreover, from a technology perspective, IaaS cloud offerings have been to date the most widely used and have had the most positive results. At the same time, PaaS has great potential, with most if not all efforts directed at new services using this model. Additionally, the model of the cloud application deployment and consumption has three levels, meaning types of clouds, based on the composition of clients or users that the cloud has. The first level is the public clouds, normally provided by large vendors, the second level is the private clouds within large corporations and organisations, and the third level is the hybrid clouds, which utilise the service of both public and private clouds [26].

The services available through IaaS are generally abstracted, virtualised, and scalable with respect to hardware, things such as data storage, computing power, and bandwidth. The PaaS offerings are focused on support in terms of a platform for programming is the forte of PaaS. On PaaS, applications perform better because the programming platform has built-in levels of support in the cloud. The use of large software packages is part of the circumstances under which SaaS is the most effective service within the cloud. Most users, however, pay little if any thought to the underlying cloud support. The example of effectively limitless storage of messages in Gmail is presented as illustrating the lack of user knowledge, interest, or concern [102], [112]. Figure 4.15, brings the service models, IaaS, PaaS, and SaaS together with the deployment models, public, private, and hybrid cloud models, illustrating their connectivity.

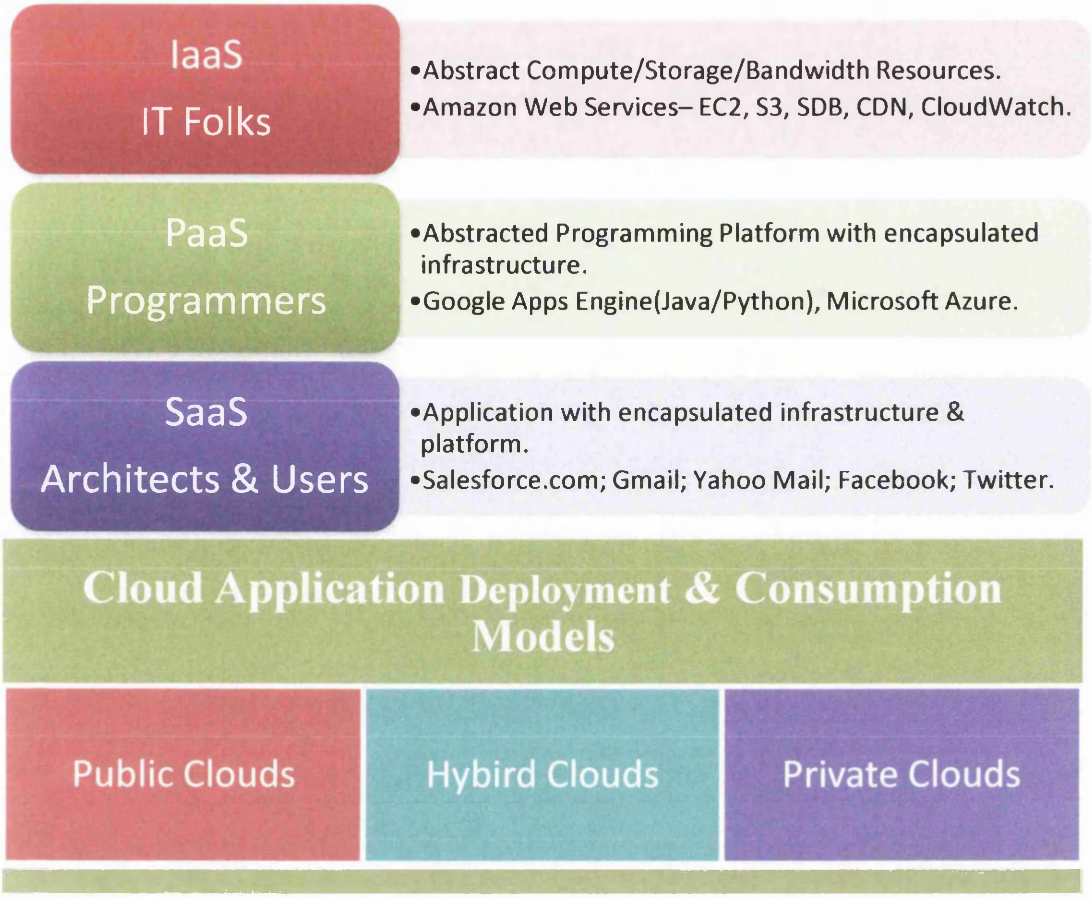

IaaS
IT Folks
•Abstract Compute/Storage/Bandwidth Resources.
•Amazon Web Services– EC2, S3, SDB, CDN, CloudWatch.
PaaS
Programmers
•Abstracted Programming Platform with encapsulated infrastructure.
•Google Apps Engine(Java/Python), Microsoft Azure.
SaaS
Architects & Users
•Application with encapsulated infrastructure & platform.
•Salesforce.com; Gmail; Yahoo Mail; Facebook; Twitter.
Cloud Application Deployment & Consumption Models
Public Clouds
Hybird Clouds
Private Clouds

Figure 4.15: The Cloud Computing Service and Deployment Models, adapted from [26]

4.5.1 Software as a Service

Fundamentally, Software as a Service (SaaS) occurs as the host or tenant provides applications for a variety of end users, who access and utilise these applications though any number of devices, such as browsers which constitute a thin interface. The client or end user has no direct-controlling interaction with the foundation cloud infrastructure, whether it is the storage, servers, network, operating system, or even the applications, with the exception of those that have limited, and proscribed, consumer-chosen control settings. These are basically transparent to the user [101].

SaaS can be explained as the software being hosted by the service provider while the user connects to the Internet and uses the needed software, as it is, basically, without reconfiguring it or integrating it with other applications or program. The trade-off is that while the organisation is relieved of the responsibilities of software upkeep and maintenance, the organisation also gives up control over its software and often its data, in addition [112].

SaaS means that instead of accessing word processing, spreadsheet, and similar programs from where they reside on the local computer or its LAN connected server, the user access to same software directly from the World Wide Web. This transfers all the responsibility of software licensing, installation, upgrading, trouble shooting, and other maintenance away from the user to the cloud service provider. Moreover, in spite of giving up stewardship, the user still has ample ability to reconfigure and customise according to individual needs [129], [130].

There are seven factors that make the SaaS model decidedly and increasingly attractive to businesses and organisations [112]. First, as employees have become increasingly familiar with the World Wide Web. Today, most have access to a computer access and experience so training for cloud computing can be quite minimal. Second, contracting for service from a vendor as provider mean needing a smaller IT staff with all the payroll savings that it entails. Third, customisation of SaaS applications on the cloud is much easier and to varying degrees already provided for as part of the service. Fourth, better opportunities for marketing niche applications gives more incentives for providers who had developed limited-demand applications, making more of these available to consumers. With SaaS, the entire world is a potential market for providers. Fifth, while discussions of the Web here and elsewhere have called attention to its weaknesses, the Web's overall reliability is quite strong. Sixth, Secure Sockets Layer (SSL) is a straightforward, trustworthy tool for insuring customers secure, as well as simple access to their applications. Seventh, recent increased bandwidth helps to ensure prompt access and efficient speeds [112].

As described by Mertz et al., in [96], SaaS entails a provider (or several vendors), who owns and manages the software, delivering or making it available to an end user, who pays according to usage. As the previous statement suggests, this is the layer most salient and comprehensible to the user. Software applications, the component of a system with which the end user is most familiar, are available on-demand and paid for as used, avoiding initial and subsequent infrastructure investment, creating significant and readily apparent cost savings. To more explicitly describe SaaS from the end user's perspective, when the user signs up for or subscribes to Google Mail, Docs, or

Spreadsheets, or something such as at Salesforce.com, both the platform and the infrastructure are out of sight, mind, and control of the user. The provider is responsible for maintaining these however, either directly or by outsourcing; the software application which the end users operates may have been developed by the vendor on the vendor's platform, all run on a third party's infrastructure.

The SaaS provider can achieve cost savings in terms of avoiding substantial infrastructure investment and licensing expenses, in roughly the same way that end users bypass these necessities and concentrate on what their area of expertise. Over the past years, the strongly felt desire to lower the cost of IT, along with the increased acceptance in the business arena to contract for SaaS, has fuelled the tremendous growth of not only this layer of service but also of Cloud computing in general. As early as August 0f 2007, according to Mertz et al., [96], Gartner's predictions of growth in SaaS cloud computing internationally were for a yearly average increase of 22.1%, arriving at $11.5 billion in business by 2011, which the analyst firm revised upward in late 2008, projecting a doubling to $14.5 billion by 2012 [131]. During the same period, the analyst company International Data Corporation (IDC) was predicting a revenue growth rate in SaaS of 31% in 2009, more than quadrupling the equivalent rate for the total software market, as reported in [94].

It is believed that there are three difficulties to the growth and more widespread use of SaaS. The first one is the lack of availability of software for those who have specialised needs or for those who need unique adapted version of more widely available applications in more general form. The second one for companies and organisations results in part from surrendering control; it is the potential for getting locked into continuing with the services of one vendor who may, in the future, cease to satisfy the user's needs. The third one is that between open source applications and inexpensive hardware, the needs of some users can be met with greater control for the organisation and at lower cost than through using the cloud [112].

4.5.2 Platform as a Service

Platform as a service (PaaS) goes beyond SaaS in that it gets rid of the limitations of dealing with predesigned and configured applications, services, and software. Beyond allowing all phases of creating and implementing applications, PaaS enables activities, such as integration of databases and web services, group collaboration, creation of various versions of an application, to name a few representative examples. The most significant disadvantage to the PaaS model stems from its linkage to one specific cloud service provider. Having created an application on one provider's platform, a company often cannot easily or inexpensively transfer the application to a competitor's service. Moreover, if the platform goes down for reasons such as the provider suddenly ceasing operations, applications and data on the platform disappear with it [112].

For users who need more abstract level of service, PaaS enables the creation and deployment of applications, using a platform which offers practically unlimited resources, such as processors, memory, and data storage. Through the use of specialised services in combination with multiple models for programming, users are empowered to develop entire hierarchies of new applications [132].

In terms of cloud PaaS, the client is now enabled to manipulate the system to the extent of using it as a basis or platform for obtaining, creating, or adapting applications. To these ends, the service provider supports or supplies programming languages and other tools for the client's work. The end users neither manage nor control the hardware, the network, operating system, or storage; however, they are typically involved in application deployment, and may even get into reconfiguring the application hosting environment [101].

According to Velte et al., [112] PaaS is typically provided in one of three forms, namely add-on development facilities, stand-alone environments, and application delivery-only environments. Furthermore, features favouring the increased use of PaaS include collaboration between individuals that are geographically isolated, merged web services from multiple sources, cost savings from utilising built-in infrastructure services built and tested by a provider, as well as from the use of higher-level programming concepts. However, Marks & Lozano in [108] pointed out two main drawbacks to widespread business of PaaS relate to vendor exclusivity. Either the services are proprietary and the customer can get locked into using that vendor and prevented from changing providers, or if change is permitted, vendor fees for migrating applications may be much more than they would have been with traditional hosting.

As illustrated in Figure 4.13 intermediates between the layer of software services (SaaS) and the virtualised layer which represents the infrastructure (IaaS) is a layer more abstract than the former, yet more concrete than the latter, known as the platform and provided under the designation PaaS. At this level, the user may develop, adapt, and reconfigure software applications simply by uploading code, at which point the platform takes over managing, including up scaling with growth in usage, all without the developer having to worry about maintenance or supporting infrastructure. Among the best known of these commercially available PaaS systems are the Force.com from Salesforce and App Engine from Google. There is connectivity between the PaaS layer and the SaaS layer above it, the platform functioning to allow development of the application which will make up the SaaS offering. Meanwhile, the platform rests on the foundation of virtualised infrastructure, which gives access to the resources of computing power and storage, which the platform needs to function and which it links to through the Cloud by means of standardised interface [109].

4.5.3 Infrastructure as a Service

Concisely, Infrastructure as a Service (IaaS) means providing virtualised resources on demand, including servers, operating systems, and software stacks. The key distinguisher of this service model is the ability which it gives the user to reconfigure the server, customising it in a more or less permanent way to fit their own needs [133]. Examples might include installing custom software packages, attaching virtual disk storage, or customising security features such as firewalls and access codes [134]. Furthermore, according to Reese, [98] the IaaS part of the SPI model goes the furthest in terms of client control over elements in the cloud. While the underlying infrastructure is still in the control of the host or service provider, the end users has freedom to arbitrarily deploy and run applications and software, including operating systems. Even some network components, such as host firewalls, may be accessed and manipulated by clients in certain contexts. Each of these service models, SaaS, PaaS, and IaaS, represents differing levels of control and freedom of operation for the user, as shown in Figure 4.16. From the figure we can see that customers have greater control when using PaaS and IaaS as a service and less control when using SaaS as a service. On the other hand, providers have less control on PaaS and IaaS as a service and more control on SaaS as a service.

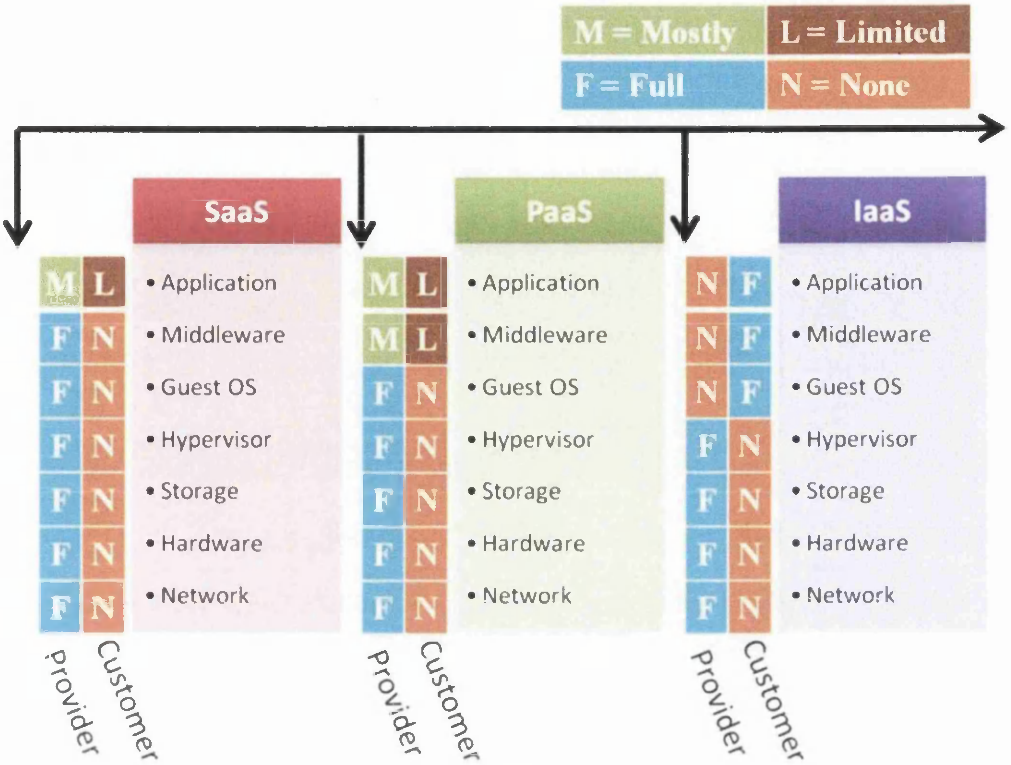


Figure 4.16: Illustration of the level of control over security in SaaS, PaaS, and IaaS, adapted from [101]

Velte et al., [112] explained that IaaS is distinct from both SaaS and PaaS in that it provides the hardware, which the company or organisation rents and is thereby given the freedom to make whatever use of that it needs. Such rentals may include server space, network equipment, memory, CPU cycles, and storage space. This service model typically involved dramatic up-and-down scaling according to changing needs, especially given the multiple users at any one time, each of whom typically pays based on resources consumed.

Most often provided in the form of virtualised infrastructure as opposed to physical hardware, Information as a Service (IaaS) entails the vendor providing the fundamental resources for computer applications and operations, e.g., data storage and processing. Standardised interfaces allow software-as-service offerings, as well as platform-as-service offerings to utilise an IaaS layer from any provider to in turn provide their services. This layer, which normally encompasses network resources, in addition to data storage, computing capacity, and the actual physical as well as virtualised hardware, has been described as the fabric layer in [99].

Infrastructure had been available, provided as a service for some time before the advent of Cloud computing. At the time, the current IaaS was referred to as utility computing, a term still in use by [97], [128], [109] to denote the Cloud computing infrastructure layer. A key characteristic change from utility computing as it was originally used and the infrastructure layer of Cloud computing today is that it has come to be integrated in a supporting function to platform and software as services. The realisation and implementation which made this integration possible was the development of a readily comprehensible, accessible, programmable, and operable interface, so that most if not all users would be able work with it, whether they are end-user clients or middle-level developers. Obviously overall, the IaaS provider must achieve a 'critical mass' of clients; therefore, access to the Cloud's physical underpinning must be practically effortless, and for this reason platforms for programming development, as well as the virtualisation of infrastructure, have become defining characteristics of Cloud computing.

4.5.4 Business Process as a Service

The Business Process as a Service (BPaaS) is an additional service type in Cloud Computing Service Models defined by IBM; these are defined by IBM as: "any business process (horizontal or vertical) delivered through the Cloud service model (Multi-tenant, self-service provisioning, elastic scaling and usage metering or pricing) via the Internet with access via Web-centric interfaces and exploiting Web-oriented cloud architecture. The BPaaS provider is responsible for the related business function(s)". Examples of BPaaS include employee benefit management, business travel, procurement, and even IT processes such as software testing [120].

4.6 Deployment Models of Cloud Computing

While it is common to equate cloud computing with multiple diverse end users accessing vendor provided services via the Web as a utility in essentially a publicly connected manner, any cloud may operate according to either public, private, community, or hybrid deployment model. Figure 4.17 illustrates three of the four classifications of clouds.

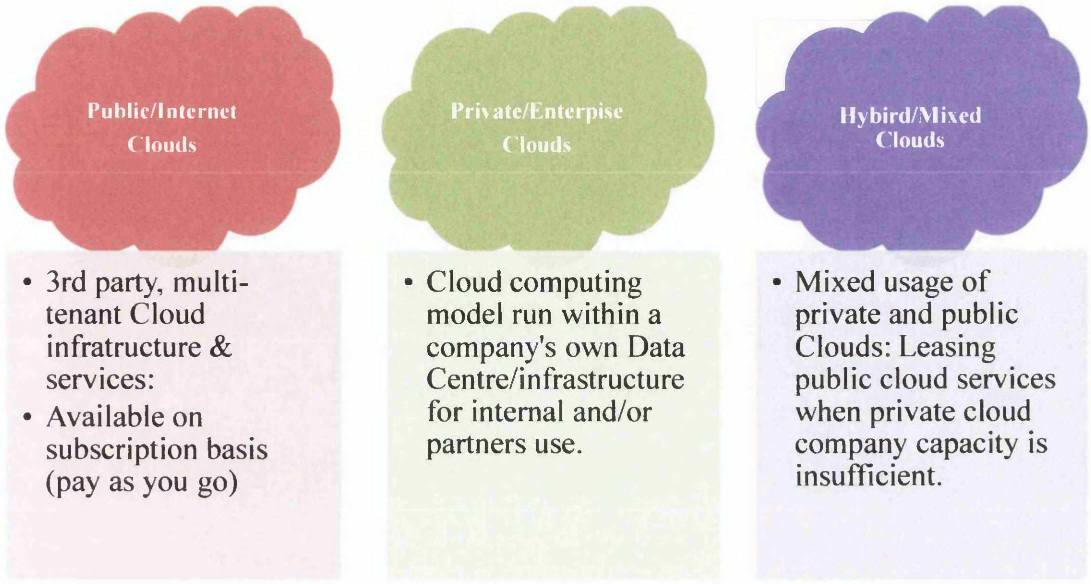

Public/Internet Clouds
3rd party, multi-tenant Cloud infratructure & services:
Available on subscription basis (pay as you go)
Private/Enterpise Clouds
Cloud computing model run within a company's own Data Centre/infrastructure for internal and/or partners use.
Hybird/Mixed Clouds
Mixed usage of private and public Clouds: Leasing public cloud services when private cloud company capacity is insufficient.

Figure 4.17: Illustration of Clouds Deployment Models, adapted from [26]

Public versus private cloud being in fundamental contrast to each other, and according to Armbrust et al., [97], who has attempted to define the distinction by propose definitions for public clouds as being made available in a pay-as-you-go method to the general public, while on the other hand, private clouds consist of the internal data centre of a business or other organisation, and it is not being made available to the general public. Furthermore, private clouds frequently develop as a business or company's network is transformed through virtualisation into a cloud environment. Somewhere between these two is the community cloud, which is defined by Mell & Grance, cited in Buyya et al., p.98 [26] as "Shared by several organisations and support a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations)".

Hybrid clouds occur most commonly when an initially private cloud that requires additional computing capacity, which is obtained by linking into a public cloud [133]. According to Winkler [101] there are four models for the deployment of cloud systems and they are: Public, Private, Community, and Hybrid.

4.6.1 Public Clouds

As the name implies, the public cloud is available to basically anyone who is paying to use it, with no stipulations on location, access point, or intended purpose. To function both effectively and securely, public cloud environments depend on a number of factors.

Perhaps most fundamentally, the end users (i.e, the subscribers) must be able to have faith in the integrity and competence of the cloud service provider. This trust is important since, the open nature of this model, along with features of the cloud in general, leads to the very real potential for the storage of data belonging to unknown-other subscribers mingled together will one's own data, leading to the need for highly accurate and dependable mechanisms for encryption, identification, ease of access by the legitimate user, and prevention of access by others [101]. Furthermore, many larger organisations which have and continue to use their own infrastructure can still benefit from access to the cloud by borrowing resources o an as-needed basis to cover periods of sporadically additional or unexpectedly high demand. This would fall under the hybrid cloud model, requires management for the life cycle of the virtualised resource usage, and needs to be maximally transparent [135].

An organisation, buying a virtual server on the platform run by a cloud services provider such as Amazon Elastic Compute Cloud (Amazon EC2) is a particularly common example of virtualisation in the context of a public cloud. While this is a public cloud, because of the way Amazon's business has developed, the platform actually began as a private cloud, which Amazon decided to take out of its own back office by leasing out its excess capacity. Thus, Amazon Web Services (AWS) was born as a PaaS offering, out of the excess capacity created for Amazon.com retail sales, taking advantage of its scalability and resilience, with the built out platform offering computing capacity to others as a subscription service, turning that portion of their business into a public cloud in the process [136].

The advantages of this kind of public cloud service to the customers are: 1) That they only have to spend money for what they end up using after they have used it; in other words, for relatively few pennies per hour of running the virtual server or consuming the disk space; 2) That they can dynamically scale up and back down, accessing more virtual servers as needed; 3) That they never have to pay for anything they have not used, and therefore do not have to carefully calculate or end up paying for anticipated needs which never materialised. Additionally, [13], [107], [108] give number of illustrations, which highlight what the firm, can gain from computing via a public cloud.

Testing and Quality Assurance: Short period needs for a server or for the need to accommodate atypically high levels of capacity are ideal situations for utilising cloud computing because when a firm rents the time it needs, it no longer has to worry about load testing of applications or software upgrade cycles.

Web-based Application Hosting: Applications on Web sites are a natural place for cloud computing in that they already are accessed from there and are susceptible to problems stemming from unexpected peaks of demand. Since Web applications are already accessed from the cloud, it can be a natural fit to hosting them in the cloud. Additionally, Web applications suffer from peak demand issues. Any event or development in the news could trigger thousand times jump in visitors to a Web-page on a given day, only for the level to fall back to the normal level within 3 or 4 days. Keeping capacity in reserve for such changes at the price of having it sit idle most of the time is cost prohibitive for all but the largest of firms.

Outsourcing Needs: Creating and maintaining data centres is something which many businesses simply do not wish to get involved with; therefore public cloud services let them outsource the IT infrastructure and support for their operations. Outsourcing these computing necessities frees the firm to concentrate on its primary mission.

High-performance Computing: Cloud computing services can provide 'horsepower' for massive applications and operations that far outdistance anything an individual organisation can put together unless they make a truly huge investment in infrastructure. Therefore, any firm that needs such computing power from time to time would find it advantageous to turn to cloud services.

Small Organisations: The enormous scale of cloud computing services brings the cost down to a matter of pennies per hour of use, making the cost well within reach of small businesses and organisations many of whom could simply not afford to equip themselves to accomplish what they need to have done.

4.6.2 Private Clouds

The Private cloud exists strictly within a particular organisation, and by definition, access by a user is restricted to that group's designated membership. One obvious contrast to public clouds is that, barring a breach of security, from the organisation's perspective, there are no anonymous users, and furthermore, all legitimate users are to some degree under the organisation's control or supervision. In spite of these advantages, individuals, departments, or sub-groups within the organisation may have legitimate and pressing needs to keep data isolated, a perfect example being the human resources department of a large corporation which maintained a private cloud. Partly for this reason, [107] cautions against thinking of the private cloud model as merely the same as the public model, except with the drawbacks and constraints gone.

According to Chee & Franklin Jr,[92] the more fundamental difference between public and private clouds is that, in the latter case, the provider of the cloud service is inherently more directly invested in making the service interface a perfect dovetail between host and end user. It would seem at first consideration that a private cloud should have such an inherent advantage in terms of security to make it the undisputed choice wherever feasible. However, as Takabi et al., [137] point out, in this case the greatest strength can also be the greatest weakness; typically the enterprise that sponsors the private cloud is the one that administers its IT security. Just as commonly, security needs and issues are not a high priority in terms of either funding or oversight.

Authors of [133], [138], [139] Stated that many organisations have already become involved in cloud computing through work with VMware and other such firms that deploy virtualisation products and thereby take advantage of the economies of scale. These days simple virtualisation is in regular use when server needs are clustered around certain times of the day or when applications need dedicated server operating system readiness. Furthermore, Velte et al., [112] predicts growth and evolution in the needs of virtualised infrastructure coincident to the growth in the number and variety of applications and demands on resources. Different areas such as Business Continuity and High Availability as well as scalability have been assessed by [101], [106] in which this growth can be anticipated. The need for even higher levels of availability is expected to grow along with increasing virtualisation of applications and resources. Additionally, it is anticipated that an IT strategy of reproducing identical virtualised infrastructure at multiple data centres and interconnecting them as a means of satisfying this demand. Furthermore, Information technology is driving many developments in the way businesses function, in turn placing new and complex demand on IT infrastructure. One of the changes is an anticipated greater need for scalability, especially and firms acquire new subsidiaries or ventures. This is an area in which private clouds fit the needs very well.

4.6.3 Community Clouds

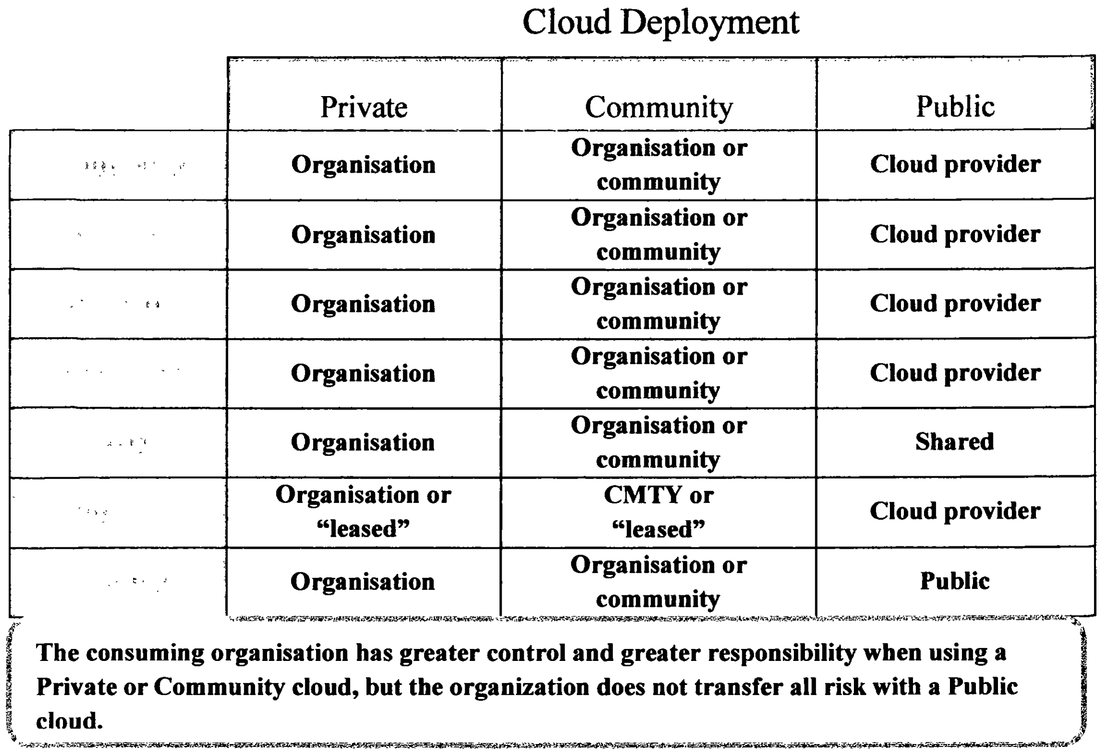
The Community Cloud model is particularly attractive for those groups who want to combine the best of the private model with the public model and its advantages. From the private model, these entities gain the cost savings of a cloud on the scale larger than what their individual organisation could effectively maintain and utilise. They also are likely to have the comfort of knowing at least the identity of all the other users of the cloud, making it possible to anticipate any threats to security [98]. Moreover, for some groups, such as governments, in particular there is the security advantage of keeping sensitive data off the flow of the Internet altogether. One circumstance that greatly enhances the feasibility and desirability of a community cloud is the situation in which the various participants share common legal or regulatory mandates and constraints. Community clouds often function best when supported from multiple data centres [107].

4.6.4 Hybrid Clouds

As its name suggests the hybrid cloud model features a combination of two or even three of the other models; hence, hybrid clouds can differ markedly from one another in their configuration. The explanation of how a hybrid cloud is created suggests that [140] sees the private model aspect to be more fundamental, in that he describes an organisation starting with the launch of its own private cloud and then augmenting the system by leveraging its way, by connecting into either a public or a community cloud, or both.

The motivation for building into such a structure would be subtly, yet significantly different from that which would lead an entity to join a community cloud. One possibility would be a situation in which short-term paid use of a public cloud was the best way to make the incremental steps toward expanding or upgrading one's own private cloud [109]. Alternatively, while public clouds may give a cost-benefit ratio too great to pass by in the quest for a unified system, legal, other regulatory, or policy constraints might necessitate the operation of a private cloud for certain, sensitive applications. This would seem to suggest that the hybrid cloud is uniformly the superior model. However, sometimes the extreme sensitivity of client data, in business or government entities precludes its storage in even the most securely encrypted form or on the most securely administered cloud. Meanwhile, the entity has other applications that are perfectly fine to be carried out in a public cloud setting, and in fact, are most cost-effective in that environment [141].

Figure 4.18 provides a more complete illustration at one time for the public, private, community, and hybrid models. This Figure spells out organisational responsibility for various aspects of cloud services, as well as legal and regulatory responsibilities that go with them, contrasting them for each of the three deployment models, Private, Community, and Public. Obviously the same determinations for any given Hybrid cloud will depend on the nature of the hybridisation.



Cloud Deployment

	Private	Community	Public
[illegible]	**Organisation**	**Organisation or community**	**Cloud provider**
[illegible]	**Organisation**	**Organisation or community**	**Cloud provider**
[illegible]	**Organisation**	**Organisation or community**	**Cloud provider**
[illegible]	**Organisation**	**Organisation or community**	**Cloud provider**
[illegible]	**Organisation**	**Organisation or community**	**Shared**
[illegible]	**Organisation or "leased"**	**CMTY or "leased"**	**Cloud provider**
[illegible]	**Organisation**	**Organisation or community**	**Public**

The consuming organisation has greater control and greater responsibility when using a Private or Community cloud, but the organization does not transfer all risk with a Public cloud.

Figure 4.18: Organisational responsibility in the Cloud Deployment modes, adapted from [101]

Hybrid clouds generally come into existence as private clouds become interconnected with public clouds; and Menken & Blokdijk, [102] demonstrated two typical scenarios of how this develops are described below. The first scenario is security where the requirements for maintaining the security of sensitive data may necessitate that the organisation keep certain portions of its operations in-house, even though it has others which be advantageous to conduct in a cloud environment. An example of the latter would be applications which need the greater scale and resources of the public cloud. The planning out of how these two component mesh could result in the creation of a hybrid cloud, as for example in the case of a bank, which could not use Amazon EC2 to store its client account data, but needed the Amazon infrastructure number crunching and new system testing. The second scenario is scalability, since a private cloud typically has fewer tenants, the large-scale cost savings can be greatly reduced in this environment [141]. It is suggested that an organisation's mixing the operations it conducts on private cloud (perhaps 20-25 linked systems) with others of its operations on public or community clouds (perhaps thousands of systems), the overall saving could be significantly better than on a private cloud alone [142].

4.7 Regulations in Cloud Computing

At the most basic level the acquisition of a cloud service is like any other and organisations must assess the operational risk. Weaknesses that may be associated with a cloud service and which would warrant particular attention include security, restrictions on access to data connectivity and the organisation's ability to retrieve data and transfer to an alternative solution at the end of the service. This is not to say that regulations will undermine cloud computing but rather organisations will need to pay particular attention to the need for controls that will help to prevent system and process failures, or to implement measures that will enable prompt rectification of a problem and continuity of operations in the event of an outage [142].

4.8 Cloud Computing Data Protection

The lack of transparency associated with the cloud computing creates a significant issue in terms of compliance with privacy and data protection requirements. It is not possible under European data protection rules for an organisation that processes personal data in the cloud to give up all control over the processing by the service provider.

Although most of the processing will be carried out by the organisation i.e. where the applications within the cloud are used by the organisation's end users, tasks such as hosting, storage and back-up are likely to be performed by the service provider, who will be considered as a data processor for data protection purposes [143].

There are several areas which need to be considered. The first area is the features of the service which must enable the organisation to comply with data protection regulations. For example, there may need to be access controls, data may need to be encrypted, and data fields restricted in order to minimise the capture and retention of data. The second area is the engagement of the cloud service provider. This area which must include terms requiring appropriate organisational and technical measures to be taken against unauthorised or illegal processing of, or the loss of or damage to, the personal data. Putting this in practice, it means the organisation must be satisfied with the service provider's standard offering and reflect this in the contract or agree specific arrangements. Another area is that the organisation must know where the data is processed in order to determine whether the rules on the adequacy of the data's protection abroad will apply, and if so, the way in which the organisation will comply [144].

4.9 Migration from Traditional IT Environment to Cloud Computing

While moving to cloud services in order to fulfil computing needs and operations has tremendous advantages, Antonopoulos & Gillam, [106] acknowledge that it will be a more complicated and often more difficult and problematic task than what the organisation usually anticipates. One of the major issues in migrating data to the cloud is the potential for problems if migrating again at a later date becomes necessary. This can be a major problem because of the proprietary nature of some providers' software and platforms and can result in data lock-in, the loss of data caused by moving between different proprietary systems. The enormity of the potential problems this might lead to can make the organisation's officers very nervous and reluctant to go through with a migration to the cloud, already an unsettling proposition, given the question of where exactly the data will be in the first place, once it is in the cloud.

Given that the chances of achieving zero loss of data in any move are a long shot, there is a real need for open standards and platforms being able to communicate with each other. Otherwise, Buyya et al., [26] believe that the company is held hostage to phantom infrastructure instead of being liberated from the need to care about the infrastructure. Further worsening the risks is the generally agreed on point of view that data is of greater value to an organisation than cash, assets, or any other part of its IT operations. Therefore, as Menken, [104] points out, under the wrong circumstances, the firm could be at the mercy of the cloud vendor in a dispute; for that reason, careful research and caution are needed. These possible eventualities need to be factored into the cost-benefit calculations

Among the issues that Rittinghouse & Ransome [13] recommends the cloud migrating firms prepare for include, missing data, broken links, misrouted emails, corrupted databases, and malfunctioning programs. They might not be noticeable during the actual migration, but could well turn up later. Therefore, potential customers should be sure that they have the full support of the vendor before migration is implemented.

Other questions to be address well in advance of executing any move include: 1) how to minimise the loss of system integration inherent in any migrating of data or else reintegrate the system, 2) what are the details the support plan, and 3) is it in place and adequate to address the possibilities described above.

In this section we lay out important considerations for both the planning and execution phases of a company's move from relying on a tradition model of self-provided computer and IT services to relying on a cloud vendor and obtaining, as well as carrying out, its computer operations in the environment of the cloud.

4.9.1 Planning the Migration

An important starting point for planning is keeping certain systems separate from the move because of their needs for security, i.e, systems such as those involving complex financial data, network administration functions, systems that require the kind of integrity which is not easily compatible with distributed characteristics of the cloud, etc; these will in all likelihood remain in-house applications. Among the high security components, one that can successfully migrate to the cloud, assuming the necessary security can be worked into the SLA and adequate contingency plans are in place, is storage capacity. The potential savings are great, but to is the need for security.

Good candidates for migration include noncore functions, such as email and web servers, along with application development servers, all of which have the added advantage to the user that problem arising with the infrastructure are the burden of the contractually obligated cloud service vendor. Furthermore, it is noted that the decision to migrate to using cloud computer services is an individual organisational determination, for example, the choice to move being much better suited to a firm whose culture is at home with open source software versus one that is committed to proprietary solutions to computing needs [136].

4.9.2 Execution of the Migration

When it comes to migration from traditional environment to Cloud Computing it is very crucial to understand that preparations for the execution are likely to take a significant amount of time, even after the planning phase of the migration is finished. For example, legacy systems will need specialised conversion although automatic programming obviates the need for tedious human labour [145].

Additionally, Menken, [104] describes the following steps or phases in the execution of a move to the cloud; 1) data extraction, 2) data loading, and 3) data verification. In the first phase, data must be systematically extracted from the old storage and prepared for insertion into the new, in this case the cloud vendor's, system. The second phase, data loading is the actual installation of the data into the cloud, i.e., the cloud services provider's system. The vendor takes on the role of expert/consultant at this point and may charge a fee if the loading is sizeable. Regardless, with the IT professionals on its staff who know the new system better than anyone else, the vendor provided vital support. The third phase, data verification, is frequently conducted by a specially designed program, which reviews all the inserted data. This process represents a testing of the data's proper storage and configuration in its new home, so to speak.

The data verification program checks for the following: 1) that extraction and loading were translated correctly, done by comparing both the old and new as sources; 2) that all data in the new location is complete, given that awareness at this stage can be addresses, avoiding loss of productivity down the road; 3) that beyond the successful translation of the data, there is complete and correct reading of the newly installed data.

Emulation is a process that has been used by various companies to get around this obstacle to smooth migration of data; it essentially simulates the environment of the old system within in the new.

Emulation is not always the perfect solution in that among other things, many older systems were technologically inefficient. By recreating the old set-up in the new, i.e, cloud, system, the data set in question gains nothing from the move to the cloud.

Given, that in such cases the decision can be made to avoid all the issues and potential problem and not move the type of data in question. Such a decision is often made midway into the process. Data migration involves substantial costs and emulation often makes the process more efficient, but has its drawbacks. Sosinsky [146] reiterates that getting to know the potential cloud services provider is essential to the successful migration.

4.9.3 Model for Migration into Cloud Computing

Especially for an organisation of any size, to move from a network to the cloud is a process that is complicated, needs to occur in an orderly fashion, and most importantly, while the organisation continues its ongoing operations. Therefore, as the term migration indicates the process must occur in stages, as opposed to a sudden shift. Furthermore, Buyya et al., [26], presents a 7-step model for accomplishing such a migration. As presented in the complementary illustrations Figure 4.19 and Figure 4.20, the steps are in brief:

1. Assess the costs and return on investment.
2. Isolate the dependencies within the existing system.
3. Map out what stays local and what moves to the cloud.
4. Re-architect and re-implement that which moves to the cloud.
5. Augment these applications as features of the cloud enable.
6. Test the migration.
7. Iterate the migrated applications and optimise.

While Figure 4.19 presents the tasks as stacked, Figure 4.20 emphasises the cyclical nature of the process.

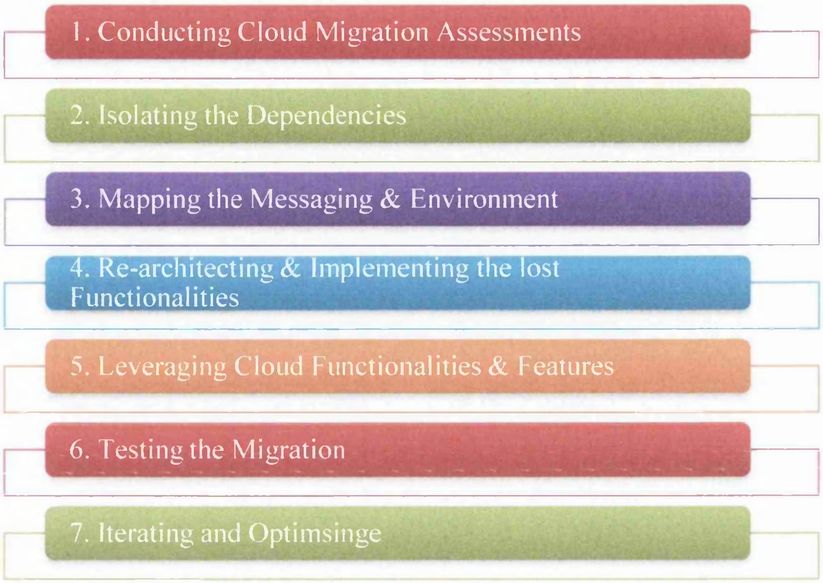

1. Conducting Cloud Migration Assessments
2. Isolating the Dependencies
3. Mapping the Messaging & Environment
4. Re-architecting & Implementing the lost Functionalities
5. Leveraging Cloud Functionalities & Features
6. Testing the Migration
7. Iterating and Optimsinge

Figure 4.19: Illustration of the Steps of Migration into the Cloud Model, adapted from [26]

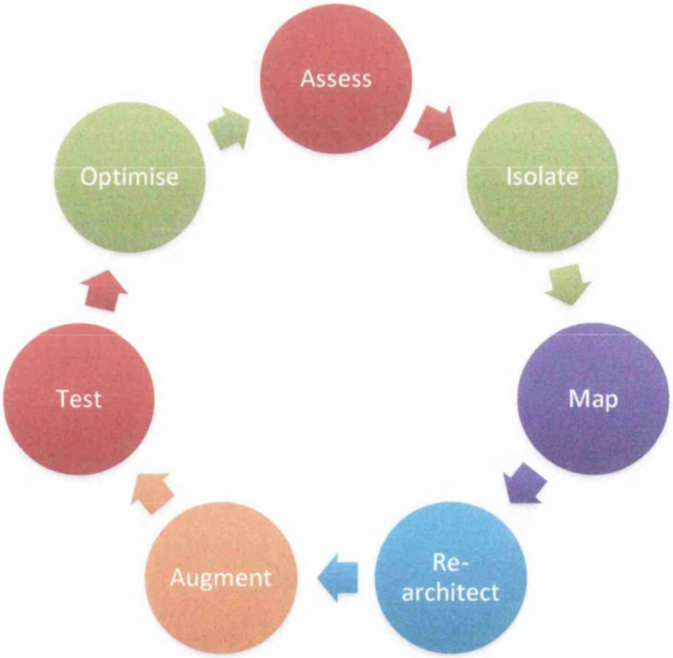


Figure 4.20: Illustration of the iterative Model of Migration into the Cloud, adapted from [26]

Certain tasks must be accomplished at each phase in the model in order for the migration to progress smoothly. Furthermore, Buyya et al., [26] explained how

Amazon's guide, sheds light on the important goals at each step or phase: Step 1 – Dependencies need to be isolated and strategies devised to handle their movement, Step 2 – Concepts need to be tried out in preparation for creating a reference architecture, Step 3 – Segmentation and cleansing of data needs to occur in preparation for the database's migration, Step 4 – Migration needs to happen in an orderly fashion, often either a key application with all its dependencies in one move, called the 'forklift strategy,' or by 'hybrid migration' in which non-critical segments moved first to be followed by the crucial elements, Step 5 – Various feature of the cloud system, such as auto-scaling, elasticity, and cloud storage need to be leveraged, and Step 6 – The migration, now complete needs to be optimised.

4.9.4 Risk of Migration into Cloud Computing

Behind both the creation of the 7-step model and Amazon's guide to the phased of migration, is an understanding of the significant risks involve in undergoing the process, whether for all or parts of an organisation's computing activities. According to Chee & Franklin Jr, [92] the first and most crucial element in migrating successfully is identifying risks and forestalling or mitigating any negative eventualities which the risks anticipate. Beyond simply envisioning what risks there will be, the testing and validating throughout the model, along with the optimisation step represent a thorough and integrated approach to achieving these goals. Risks that arise because of migration fall into two major categories, namely general risks and the security-related risks.

Furthermore, Buyya et al., [26] and Simmonds et al., [147] point out that general risks include: 1) problems with performance monitoring and tuning, 2) disruption of business continuity, 3) disaster recovery problems, 4) threats to portability and interoperability, 5) the potential for vendor lock-in, 6) threats to QoS parameters, 7) licensing issues, and 8) misunderstandings on the part of senior management of the importance and complexity of migration. Additionally, Howell & Kotz, [44] cites a significant number of security risks, which arise during cloud migration. These involve threats to or problems with: issues of trust and privacy, legal compliance, failings of cloud service providers, data leakage in the cloud, issues with vulnerability management and incident response, and issues of consistent identity management.

4.10 Chapter Summary

Cloud Computing is real; it is being adopted by governments and industries worldwide. Cloud computing has been a subject of much stir and debate, this chapter aimed to explain the cloud computing space, its definition, history, drivers, its relationship with other technologies, its architectural details including non-functional properties, various service and deployment models, issues surrounding data protection and regulations, as well the process to migrate to cloud computing from traditional IT space. Having explained cloud computing technologies, we are now ready to introduce, in the next Chapter, a few major players in the area of cloud and cloud services offerings.

Chapter 5: Cloud Computing Vendors and Services

Cloud computing is a disruptive innovation. Although dismissed by some in the industry as not a particularly substantial innovation, in fact cloud computing allows companies and their IT professionals to turn much of their attention away from maintaining their own data centres and focus it on the higher level tasks of delivering information technology solutions for the organisations' primary goals [104], [112]. Like the Internet business boom of the 1990s, there is currently a boom in terms of companies of all sizes getting into cloud vendor services. Of particular interest are the business models by which the major established firms in the field are seeking to expand into the cloud. While many of the largest firms are moving toward the more on-demand structure that comes with cloud technology and service, they are continuing to maintain aspects of their more traditional model, given that it has been so profitable to date [104]. The migration to cloud services has been a gradual step-by-step endeavour with an immense bulk of effort directed at developing the economies of scale needed to provide efficient cloud computing, through investment on infrastructure. Furthermore, offering customers the most non-essential features will be the hallmark of the providers who will come to be the most successful; however, they contend that what is far more crucial to overall success is the providers' ability to change potential and real clients' perceptions of disadvantaged and risks in cloud computing. Reliability and data security are the biggest areas of concern, and the already established names in the computer/Internet/IT industry have the advantage of a comparatively long and generally positive track-record [108].

Having introduced cloud computing in the previous chapter (Chapter 4), we are now ready to look at specific cloud services available in the market today. There are already literally dozens of providers of cloud services from among which a company or organisation may choose. The three widely known names offering cloud services are Amazon, Google, and Microsoft, [104], [112]. As we have evidenced in Chapter 4, IBM is also among the top players offering Cloud systems and services. We will review the services offered by these four cloud vendors in this chapter (IBM, Microsoft, Google and Amazon will be reviewed in Sections 5.1, 5.2, 5.3, and 5.4, respectively.

Subsequently, in Section 5.5, we will discuss the specific cloud services offered by Amazon in some detail. We have selected Amazon cloud services for a detailed discussion in this thesis because Amazon is currently among the top cloud services vendors who have made in good detail the services related information available on its website in the public domain. The explanation provided for Amazon cloud services in this chapter will also be useful in Chapter 7 where we model and analyse Amazon market sectors, applications, and workload. Finally, the Chapter is summarised in Section 5.6.

5.1 IBM

As a cloud services vendor, IBM offers computing services in public, private, and hybrid models to organisations of all sizes, incorporating its background in providing consulting for specific industries. Some of these features include design, implementation, and security services for cloud computing as well as industry-specific business consulting services for cloud computing and technology consulting. According to Schmotzer and Donovan [148], IBM's Blue Cloud is a return for this company to the cloud computing business and is aimed primarily at the financial services market and all of the processing work that went on in the firms' back offices. IBM's goal has been to add cloud services to the portfolio of very profitable consultative services they provide for top flight businesses, such as those on the Fortune 500 lists. IBM has a powerful Cloud platform which is known as LotusLive. The platform provides clients with online services that are delivered using the SaaS model. Users of LotusLive can enjoy the flexibility of it as it allows them to easily work together and do businesses with anyone, anywhere, at any time at a low monthly rate. LotusLive provides users with solutions for Web conferencing, Collaboration, and email [149].

Due to the nature of its consulting services IBM is in a good position to assist and advise clients, many of whom will be incorporating the services of both public and private cloud models. Beyond providing the cloud services, platforms, and infrastructure, there is a need for guidance in transitioning businesses and organisations in their movement to cloud computing, a service in itself for which IBM is well suited. Furthermore, there are several efforts by IBM to assist new clients in moving their computer operations to the cloud. Examples of such efforts include working with organisations such as Neighbourhood Centres of Houston, TX, developing industry specific cloud based services such as Lender Business Process Services and Healthcare Process Services, and even developing a cloud-based service platform for businesses in China [112].

With decades of experience in creating, as well as managing, corporate data centres, IBM envisions such centres as being open and accessible, just as is the Internet (obviously to those inside the firm with legitimate reason to access.) To this end, IBM has invested considerable resources into conducting research on cloud computing technology at research centres all over the world, including the Almaden centre, famous for its pioneering work with DB2 databases and Linux servers, and the Shanghai centre currently studying efficient resource utilisation, application workload and power consumption, in order to enhance the smart scheduling of workload migration, as a result preventing the unnecessary consumption of energy [102], [150] .

Recently, IBM has been teaming up with Google to work with academic and scientific institutions, doing research into cloud computing in anticipation of a massive up scaling of operations within the decade. In this collaboration, IBM has provided expertise with large scale networks while Google has supplied the enormous volumes of data for processing by the server networks [104].

5.2 Microsoft

Until recently, Microsoft leased most of its data centre operations from others; now it is working on creating its own server farms, such as the one in the vicinity of its Washington state complex. Microsoft labels its strategy software plus services, which denotes a setup with software remaining on client's terminals for the foreseeable future, in other words, a less-than-total migration of the support for computer services to the cloud. Microsoft is apparently anticipating that its clients will have concerns with security or dependability [104]. Microsoft is in the enviable position of having built a business model on supplying the software to computers, becoming a fixture in the industry all the while working on large operating margins, which now enable their company to move into opening the data centres needed to provide cloud services, creating a hybrid business model consisting of software and cloud services. Microsoft's

SharePoint is the quintessential example of the company's model in that it firstly is aimed at the biggest of corporations as clients, secondly has capabilities such as enabling single database, browser accessible enterprise sharing, and thirdly has the greatest base of large-scale customers, as exemplified by the reputedly 30,000 users which Microsoft brought on board through the efforts of the Coca Cola company. While Microsoft is set to provide cloud services to even the largest scale enterprises, one of their advantages is that they can provide for the family business and for individuals equally well. Since many of these people in small operations already use Microsoft products, the similar cloud versions would be familiar [112].

Windows Azure works on and with Windows applications and Microsoft's data centres. Its components include Windows Azure, Microsoft SQL Services, Microsoft .NET Services, Live Services, Microsoft SharePoint Services, and Microsoft Dynamics CRM Services. The Windows Azure Platform allows users to have the ability with the use of Microsoft data centres to build, host and scale web applications. The Windows Azure is considered to be a PaaS. It is a range of on-demand services that are hosted Microsoft data centres [151]. The Windows Azure Platform is available to users in two ways, subscription where users can subscribe on a monthly basis or they have the choice of going on pay as you go where they are only billed for what they use. In spite of its name and offerings, Microsoft has been evaluated as not being at the forefront of a rapidly growing and changing field [112], [152], [153]. Furthermore, Microsoft has announced plans for Live Mesh, a service designed to synchronize Windows and Windows Live services. Among younger members of the video game population, Xbox Live, Microsoft's Xbox 360 platform is quite popular as a per month subscription service based in the cloud [104].

5.3 Google

The recent pace of Google's development of its cloud services has been remarkably rapid, with two billion US dollars worth of infrastructure investment per year for their data centre along. Moreover, four new data centres are under construction, at a projected cost of over $600 million dollars each. Expert estimates suggest that Google's cloud may encompass as many as a million inexpensive hardware servers. Beyond this, Google is preparing to work with IBM in the area of cloud research in a manner similar to that of IBM's current efforts involving its China centre [102], [110] .

With users in the millions and free to all who have Google accounts, Google Apps is central to the company's cloud services, in particular because it provides essentially the same productivity software as what Microsoft Office provides at $500 per user. What is more, the premium software addition enables collaboration and storage for quite large files [104], [112].

Although General Electric and other sizable corporations have settled on Google Apps for their cloud services and Google claims over ten million users, Menken and Blokdijk [102] believe that Microsoft Office will continue to dominate the field in terms of major clients, given that while most of the ten million are students or small business proprietors attracted by the cost saving on software. Google's App Engine, on its Python platform and Google servers, provides 500 megabytes free storage, enabling developers to create their own software. Additionally, one of the features that contribute to integration is the ability to use App Engine to build web-based applications on the very infrastructure Google uses for its own applications. The enterprise version of this service runs $0.10-0.12 an hour (CPU time) and $0.15-0.18 a gigabyte [104]. According to Velte et al [112] Google's cloud service provision is considered to be among the largest business ventures that Google has to date undertaken. Furthermore, with Google App Engine, developers can accomplish the following tasks: they can write code once and deploy it, configure the system to absorb spikes in traffic due to sudden increases in usage, integrate applications with other Google services.

5.4 Amazon

As a cloud service provider, Amazon has one of the best known packages in its portfolio, which offers scalability, speed, and reliability, all at a very inexpensive price tag because it makes use of the data centre infrastructure of Amazon's global web-based retail business. Amazon Simple Storage Service (Amazon S3) is storage service for the Internet. It provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites [113].

Amazon was one of the first companies to offer cloud services to the public, and they are very sophisticated. Amazon's cloud service offerings include the Amazon EC2, Amazon S3, Amazon Simple Queue Service (Amazon SQS), and Amazon SimpleDB (Amazon SDB), a web service for running queries on structured data in real time. In addition, Velte et al [112] briefly explain each, going on to note the disadvantage of needing to use the command line for access. Amazon is one of the main players in Cloud Computing as it offers several platforms. AWS is considered to be a very powerful and complete cloud services platform. Such platforms, enable businesses to have a range of services such as compute power, storage, content delivery, as well as functionalities with the aim of allowing businesses to organise and deploy applications and services that can be achieved at reduced cost with better flexibility, scalability, and reliability [153], [154]. Furthermore, Menken [104] explained that Amazon has grown from an Internet based retailer and sales facilitator of books to an enormous firm whose business is focused much more broadly on technology and one of the largest vendors of cloud services to date. Having built a large computer processing infrastructure in anticipation of internal needs, they were well positioned to move into the cloud service vending business at the very start, simply by selling Amazon's spare processing capacity to other companies. To date, the bulk of Amazon's customers are small businesses, many of them start-ups with little or no capital to set up their own systems.

At mere pennies per hour, Amazon's computing services such as Amazon S3, Simple Queuing Service and the Elastic Compute Cloud (collectively known as AWS) are incredibly attractive to small businesses lacking capital funds to purchase and set-up hardware, especially since the services are charged for on a per usage basis. To illustrate the point, Amazon S3 charges £0.10 per gigabyte per month, all predefined to store photos, audio, and video files. The pay-as-you-go structure is familiar to most users (in many cases comparing quite favourably with other similar systems used for self-storage of property, mobile phone service, etc), instilling confidence in the system and comfort in the absence of charges for unused, but contracted for service. Beyond the confidence boosting effects of flexibility and scalability, the ability to specify data storage in either Europe or North America encourages trust in the Amazon system [102], [154]. Amazon has been declared to have most extensive cloud service on the market.

5.5 Amazon Web Services

This Section discusses Amazon Web Services (AWS) in two parts. The first part entitled "Non-Functional and General Characteristics", comprising Sections 5.5.1 to 5.5.5, discusses Amazon cloud services' characteristics that make it among the top cloud computing vendors. The characteristics that are discussed include Flexibility, Cost-Effectiveness, Scalability and Elasticity, Security, and Experience. The specific sets of cloud computing services offered by Amazon (as on May 2012) are reviewed in the second part entitled "Specific Cloud Services offered by Amazon", comprising Sections 5.5.6 to 5.5.14. The specific sets of Amazon services discussed include Compute Services, Content Delivery, Database, Deployment & Management, Application Services, Networking, Payments & Billing, Storage, and Workforce.

NON-FUNCTIONAL AND GENERAL CHARACTERISTICS

The internet retailer Amazon branched out into the business of Cloud computing as a service vendor in 2006, by providing IT infrastructure through its AWS subdivision. What it was offering individuals companies and other organisations was the opportunity to forego the traditional initial capital investment in creating infrastructure in favour of paying for services commensurate with usage adaptable to the changing needs of the business and at a much lower cost per amount of use. Taking advantage of the inherent features of Cloud computing, the company could dispense with months' worth of planning and installing IT infrastructure and simply call up servers numbering in the thousands if need be, all made available by AWS. Unsurprisingly, the service has gained a client base which currently numbers in the hundreds of thousands spanning over 190 nations across the globe, all of whom receive access to an infrastructure platform that scales up or down nearly instantly, all with excellent reliability for a miniscule cost [155]. The physical infrastructure behind AWS includes data centres in Europe and the United States, Japan, Singapore, and Brazil, but regardless of either hardware or user's location [155], [154].

As an Internet-based cloud platform, AWS consists of a variety of computing services provided by Amazon.com as a vendor, which as in with all cloud computing systems are made available to the customer through Internet access to a highly reliable, scalable infrastructure that is distributed and remotely located [155].

The variety and organisational relationships among the services which constitute the AWS cloud platform are illustrated in Figure 5.1, together with the AWS-specific terminology, as a means of explaining the possible patterns of interaction between a given user's applications and the various related AWS offerings, as well as among the various services themselves. The goal of the platform architecture as a whole is to minimise the costs of administration and support, while maximising the flexibility for users with all sorts of needs and service usage. Central in the diagram are Amazon EC2 and Amazon S3, the best known of the company's offerings [156].

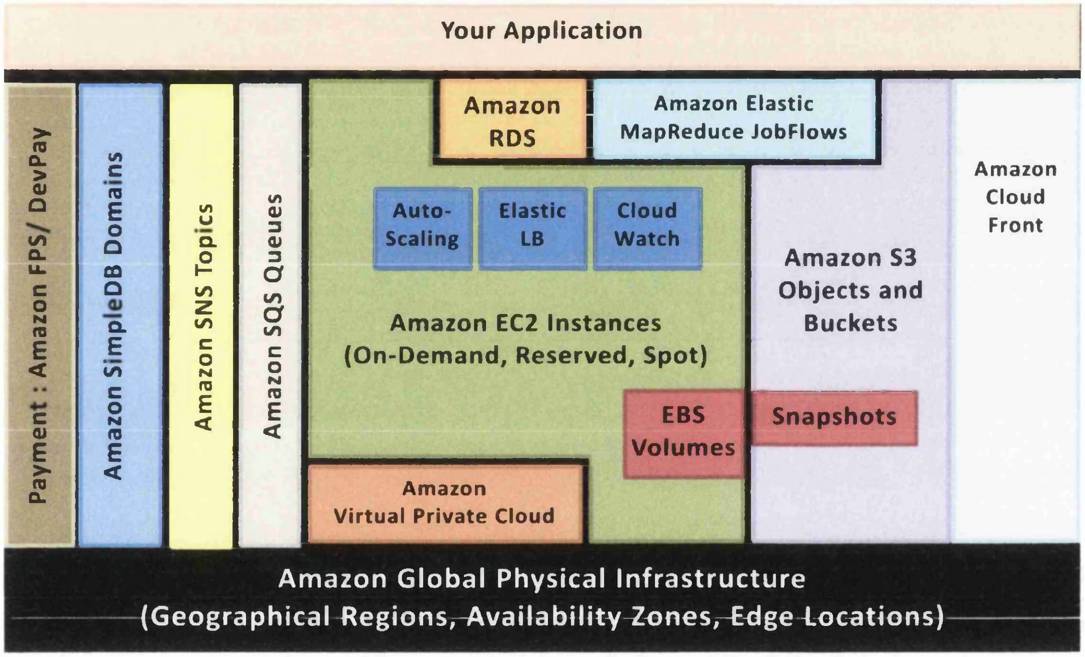

Your Application
Payment : Amazon FPS/ DevPay
Amazon SimpleDB Domains
Amazon SNS Topics
Amazon SQS Queues
Amazon RDS
Amazon Elastic MapReduce JobFlows
Auto-Scaling
Elastic LB
Cloud Watch
Amazon EC2 Instances (On-Demand, Reserved, Spot)
EBS Volumes
Snapshots
Amazon Virtual Private Cloud
Amazon S3 Objects and Buckets
Amazon Cloud Front
Amazon Global Physical Infrastructure (Geographical Regions, Availability Zones, Edge Locations)

Figure 5.1: Amazon Web Services, [156]

As technology and the field of IT have experienced such phenomenal growth and change over the last few decades, keeping abreast with and managing such innovation and its consequences has presented enormous challenges for those executive responsible themselves as the IT experts, let alone the management of other aspects of business organisations. For example, the architecture for an average business application during the past decade has seen transformation from being a desktop-cantered work station, to being a part of a system of client/server solutions, and most recently to being loosely coupled point of access for web services and their service-oriented architectures [154].

The key to achieving the recent, dramatic reductions in operating costs coupled with significant advances in reliability has been the introduction and growing use of the technology known as virtualisation. Working in tandem with developments such as grid computing, virtualisation enables organisations to engage in data crunching, analytics, and business intelligence previously impossible due to time and money constraints. Not only has the world of IT developed rapidly in recent years, but so also has the world of business, from the sheer speed of product innovation to the fundamentals market functioning. The provision of SaaS and its increasing utilisation by the business community has interacted with these other trends to spark the inception of Cloud computing, with AWS among its earliest and most developed platforms. The platform which constitutes AWS is scalable and flexible, while at the same time extremely cost-effective, yet above all easy for the user to access, navigate, and manipulate making it well suited to all sizes of companies and organisations [154].

In terms of availability and cost, the principle advantages of AWS to users of cloud computing services are threefold: 1) No matter the size, from the individual user to the largest organisation, users gain access to a great array of cloud-based computing services whether software or infrastructure; 2) All this access is on-demand, and the charges are on a pay-per-use basis; 3) Users can do away with procuring, installing, and maintaining their own hardware and related infrastructure, leading to significant cost savings. Beyond these benefits, operational advantages provided through AWS cloud-based services include: 1) Users are able to start new applications and operations in minutes, as opposed to the months it would take to prepare the necessary infrastructure; 2) Users can carry out computational projects requiring huge quantities of resources rapidly, calling up all the required computational capacity and storage at a moment's notice, and then dismissing it all as soon as it is no longer needed [157].

Amazon's history of building and improving upon massive-scale, distributed IT infrastructure, making it increasingly reliable and efficient, goes back as far as 1995, with its investment in the infrastructure exceeding several hundred million dollars in the first decade alone. All of the experience and infrastructure that have grown out of the building and administering of Amazon's global internet-based retail business, which has come to be one largest in the world. All of AWS services are designed to be elastic, allowing an organisation to access additional computing power or data storage virtually instantly whenever needed. On top of this, a company or individual may select from any of the programming models or even development platforms whichever appears most appropriate for the task at hand. Moreover, according to AWS [154], all customers do so paying only for each service used and only for the length of time it is in use.

The following features of AWS discussed in sections 5.5.1 to 5.5.5 set it apart from other providers of similar computing services [154], [155], and [158].

5.5.1 Flexibility

AWS can claim tremendous advantages in terms of flexibility over any traditional IT system by virtue of AWS being cloud-based computing. Before cloud computing, extensive new architecture, operating system, programming language, and software investments programming languages, and operating systems were frequently a prerequisite to any organisation's commencing computer operations. Once these investments are in place, they often tend to act as a weight, or worse an anchor, making it difficult for a business to rapidly and efficiently incorporate new technologies, or to respond to market dynamics and opportunities by moving in new directions or launching 'next-generation' products and services. The cloud computing model in AWS obviates the need to adapt to pre-existing infrastructure and applications or go through an extended procurement process in order to expand, reorganise, or innovate. This flexibility extends to the user choosing the optimal from among programming models, languages, and operating systems on a project by project basis, a feature which also means that IT staff and developers do not need to engage in learning, and that legacy applications may be transferred to the Cloud inexpensively and without significant difficulty, taking advantage of increased computational capability without the need of rewriting or massive reconfiguring.

The differences between building new applications on AWS, with its virtual infrastructure and developing them in traditional IT environments is insignificant; moreover, services may be integrated into their own platform as a group or launched independently for separate functions. Furthermore, AWS can handle the entire spectrum of complexity, whether the operation is as limited as backing up off-site data or batch processing, or as complex as complete integrated web-based applications.

Beyond enabling businesses to respond to time-sensitive market driven opportunities with new applications developed directly on the AWS platform, organisations have the opportunity to move selected SOA parts of their existing computer operations to

Amazon's cloud platform as best fits the individual organisation's needs. Normally, the first to be moved would be limited internal dependency applications that profit most dramatically from the scalability and accessibility offered by the cloud environment. To date, many of the most sizable companies have settled into a hybrid mode, with part of their computer operations in the cloud while others remain on the company's own network. While these organisations may with more experience move a greater portion of their activity to the cloud, it is clear that platforms such as AWS are becoming a permanent component in meeting their business's IT needs.

In a tradition IT model, launching a new product or service and its attendant applications would necessitate a process of planning, budgeting, procuring resources, setting-up infrastructure, deploying applications, and hiring new personnel, taking up weeks or more often months before actual operation or implementation. By contrast, whether its goal is to create the prototype of an application or to host a production solution on a long term basis, a company merely needs to sign up for AWS services and immediately commence deploying cloud-based operations, which could call for the computational resources of anywhere from one to one thousand servers.

5.5.2 Cost-Effectiveness

Throughout the history of the field of information technology, there have existed major trade-offs between the cost saving benefits promised by each innovation to be introduced and the investment that would be needed to achieve those savings. One notable instance of this phenomenon is the inception of e-commerce applications. Their development and deployment proved to be a significantly inexpensive endeavour, and they promised to yield substantial increases in business generated; however, the additional hardware and bandwidth required for successful deployment offset those projected cost savings and revenue increases to a significant degree, especially when these required resources were not continuously in use. The Cloud breaks this paradigm in two ways: first, the infrastructure is supplied by the provider, so the user avoids all the associated costs of set-up and maintenance; second, the user need only pay for the resources actually used. Furthermore, with cloud-based computing, computational resources, data and application related storage, as well as bandwidth are each available in unlimited quantities. Given that user's exact future needs are difficult to predict, even on an on-average basis, and that these needs often fluctuate markedly over time, cloud-based computing offers unprecedented returns on decidedly lower investment (ROI).

Beyond the cost saving that cloud services such as AWS enable by allowing organisations to dispense with infrastructure set-up and administration, the agility that they offer by providing virtually instantaneous unlimited scalability enables companies to respond quickly and effectively to emerging opportunities and problems in the marketplace, at times gaining a competitive advantage, but always in ways the tend to reduce costs and drive new business.

In addition to dispensing with the initial investments in procuring and installing hardware and software to create infrastructure, organisations using AWS do not have to deal with either long-term commitments or to required minimums in terms of usage or spending. Aside from any IT consultation support that a client may need and approach AWS staff for, the entire cloud service relationship is carried out without the interaction of personnel, such as sales representatives. As an on-going process, cloud computing through AWS also saves the client the cost of maintaining its own system, such as the expenses of electrical power for running the system and cooling the facilities and hardware, expenses involving real estate whether leasing or owning, personnel cost for a sizable IT administration staff, etc.

5.5.3 Scalability and Elasticity

Before cloud computing, IT systems required substantial investment in and development of infrastructure in order to achieve any significant degree of scalability or elasticity; thus, cloud services represent a major step forward in terms of reducing costs and increasing Return on Investment (ROI). As used in describing AWS offerings, the term elasticity refers to the system's ability to scale available resources up or down effortlessly, on-demand and without planning or friction to the system. The need becomes readily apparent in cases such as that of a company's open benefits enrolment period when the traffic heading to an application could be expected to double or even triple, during which the routine operations and activities of the business must continue unaffected by any spike in IT traffic anticipated or not. The Elastic Load Balancing and Auto-Scaling features of AWS are perfect example of cloud computing ability to scale resources in both directions in accord with changing demand.

Operations which are one-time, and short in duration but at the same time, vital to the company's on-going operations are well suited to the core features of cloud computing. For instance, through AWS, a drug manufacturer can call up the computing capacity it needs to run trials of a new product and then switch those resources off as soon as finished. Alternatively, storage and computing resources might be suddenly needed for an indefinite length of time in order for relief organisations to deal with a natural disaster. However, these characteristics of cloud computing extend to more regular and predictable tasks such as invoice processing and billing or payroll, enabling cloud users to lower their costs and save their own computing resources for other operations. The elasticity of cloud services such as AWS with its simple API calling procedure for additional resources makes it possible to avoid upfront investment in infrastructure for operations that are short-term or fluctuating in their demand for computing power.

According to the way Cloud computing services are organised on the model of utilities providing service through a grid, all responsibility for infrastructure set-up, administration and maintenance rest with the cloud computing services provider. As the clients, companies, organisations, and individuals simply sign up, use the services whenever and to whatever extent they need or desire, and pay for what they have consumed from among the seemingly limitless computing capacity, storage space, and other resources.

5.5.4 Security

Having been approved by or through all of the following industry accepted review mechanisms: a) Level 1 of the Payment Card Industry Data Security Standard (PCI DSS), b) International Organization for Standardization 27001 (ISO), c) Federal Information Security Management Act (FISMA) Moderate, d) Health Insurance Portability and Accountability Act (HIPAA), and e) Statement on Auditing Standards 70 Type II (SAS), AWS's platform can claim to be both secure, through multiple physical and operational layers and durable, guaranteeing the safety and integrity of the customer's data. Given the wide range of services available through AWS, along with the scalability and flexibility of its platform, as well as the great degree of availability and reliability that customers expect, it become a matter of necessity that Amazon guarantee both security and privacy seamlessly from one end of the platform to the other. While customers themselves have to make use of the security features and follow best practices in order to ensure the security of individual application environments, Amazon is prepared to work with clients to ensure that their data, applications, and operations in general are always maintained at the highest levels of availability, integrity, and confidentiality, since in the end not only is the actual, practical security important, but so is the customers' perception and their resulting confidence and trust. Amazon has demonstrated a commitment to maintaining the infrastructure of AWS in a completely secure and reliable manner, through attention to the following four areas:

- ✓ Certifications and Accreditations, Having completed a SAS70 Type II Audit demonstrating satisfactory performance, Amazon continues to proactively seek appropriate security certifications and accreditations so that both existing and potential customers may have every confidence in AWS's infrastructure and service security.
- ✓ Physical Security. AWS and the physical infrastructure supporting them are housed in data centres across the globe under sole control of Amazon, with knowledge of the locations limited even within the company on a need-to-know basis and with access restricted by security clearance procedures, as well as physical barrier to guarantee authorised access only.
- ✓ Secure Services. The AWS cloud has been configured to simultaneously maintain cloud-level flexibility while preventing all unauthorised access to any of it systems, as well as any unauthorised use or appropriation of applications or data, whether part of the AWS infrastructure or the client's applications and data.
- ✓ Data Privacy. With AWS, a user can encrypt any or all personal or business data within the AWS cloud; furthermore, redundancy and back-up procedures for services are published enlightening the customers, as well as any of the end-user clients they may have just how their data flows throughout Amazon's cloud.

5.5.5 Experience

AWS has been set-up to facilitate the organisation's smooth transition into cloud computing. However, migrating to the cloud, as with as any other new business venture involving IT, requires consideration and planning, including detailed communication with vendors and service providers. This applies equally to cloud computing service vendors, just as it would to hardware and software vendors in setting up a traditional IT network system. Furthermore, the relationship with a cloud services vendor is more ongoing, that is to say involving more direct interaction over the long term, even in the absence of problems. Therefore, the vendor's trustworthiness, which is a function of both experience and track record, becomes paramount in a business or organisation's choice of service provider.

It would be far too expensive for most organisations to duplicate the reliability, privacy, security, not to mention the sheer scale of the AWS Cloud. This advantage stems primarily from Amazon's more than a decade and a half of experience in running its own internet based business, which has grown to multi-billions of dollars in sales, the infrastructure Amazon has built to manage said business, and the reasonable anticipation that the Amazon will continue to innovate in managing its network, infrastructure, and services. The global web platform that is AWS provides computing services to a worldwide clientele numbering in the millions of customers, representing billions annually in commerce.

The history of AWS, since its 2006 launch, is filled with examples of new-feature innovation in response to customer feedback all at the same, consistent high-level of reliability and security. The AWS Cloud is the combined result of continual operations monitoring to maintain dependability, the consistent adherence to best IT practices, and the proprietary advances that Amazon continues to pursue, all which is at the disposal of businesses, organisations, and individual who use its services.

SPECIFIC CLOUD SERVICES OFFERED BY AMAZON

As becomes apparent in the ensuing discussion of individual Amazon Cloud services, its platform is comprehensive in covering computing capacity, content delivery, data and application storage, along with various other function. Fundamentally made available in a format equivalent to self-service, AWS has the added advantage for customers that they can access and implement the services as dictated by either timetables within the organisation, or external market forces or other circumstances, without having the delay of having to make arrangements with the service provider in advance [156].

5.5.6 Compute Services

5.5.6.1 Amazon Elastic Compute Cloud

Since its inauguration in August, 2006 as a pay-as-you-go type service for both individual end-users and companies or organisations, Amazon EC2 has provided customers with the ability to do the following [154], [159]:

- ✓ Add cloud computing capacity as needed.
- ✓ Upload their own machine images to be executed on any number of virtualised instances.
- ✓ Keep complete control over and administrative right for both the virtual machine itself and the firewall settings for the network.
- ✓ Access and run numerous virtual Linux servers on demand, along with as many computers as called for web applications and data processing.
- ✓ Fully control each server, including root-level operating system access.
- ✓ Configure firewalls and install any desired software.
- ✓ Permanently save the virtualised image of any set-up Amazon EC2 server.
- ✓ Launch preconfigured servers, read-to-work servers in whatever quantity needed.
- ✓ Shut down servers whenever they are no longer needed.

Moreover, according to [157], Amazon EC2 makes available persistent long-term mountable storage options, in addition to providing various distinct kinds of instance with equally varying performance characteristics. As explained in [159], an API is available to manage server images, creating, starting, and shutting down instances as needed. The essence of the Amazon EC2 Cloud environment is its aggregation of virtualised machines. The term instance is used to refer to each node of Amazon EC2 cloud environment. Charges are calculated on the basis of total CPU time used, in addition to communications among instances, and between those instances and Amazon EC2 external machines. The cost of each instance determines its specification. Purchasing the default plan authorises the user to run up to 20 instances at once. Specifically, at the time of writing, within the US region, Amazon charged $0.085 per hour of operating a default instance, $0.10 for every GB of inbound communication, $0.10 - $0.17 for every GB of outbound communication, and $0.01 for every GB of communication between nodes in different regions. Additionally, there is was no charge for communications among instances inside any given region [160].

5.5.6.2 Amazon Elastic MapReduce

The Amazon Elastic MapReduce (Amazon EMR) makes data processing on a truly gigantic scale both easy and cost efficient. In order to accomplish this feat, Amazon EMR draws on the infrastructure of both Amazon EC2 and Amazon S3, as hosted in a Hadoop framework. Users are freed from the concerns of tuning Hadoop clusters or even planning and implementing elaborate set-ups; instead, they can give their full attention to the issues directly involved in crunching and analysing the data [154]. General applications, such as data mining and warehousing, log file analysis, web indexing, machine learning, as well as discipline specific one such as financial analysis, scientific simulation, and bioinformatics research are among the applications which require high speed processing of enormous quantities of data and for which Amazon EMR is particularly well designed [161].

5.5.6.3 Auto Scaling

Users of the Amazon EC2 are able to set the triggers in their usage patterns which will automatically cause a scaling either up or down as they have pre-directed. This feature means that to the extent any user can anticipate conditions warranting a change in demand, they can direct Amazon EC2 so that it ensures seamless performance in spite of spikes in processing demand, as well as automatic scaling back to minimise wasted usage and the excess cost that would go with it. Among clients whose capacity needs tend to fluctuate by the week, day, or hour, auto scaling is particularly beneficial, all the more so in that it available within Amazon's CloudWatch as an included part of that service for it standard cost [162].

5.5.6.4 Amazon Elastic Load Balancing

Another behind the scenes, but indispensable service is Amazon Elastic Load Balancing (Amazon ELB) which connects to various instances, real or virtual Amazon EC2 in order to distribute application traffic automatically as it arrives. This in turn boosts the fault tolerances significantly for users and their applications, by providing seamless load balancing capacity. Unhealthy instances are discovered within the pool, and the incoming traffic gets rerouted automatically to the healthy instances, until the problem has been corrected. ELB is customisable, either according to an individual availability zone or to numerous zones, and is even available for Amazon Virtual Private Cloud (Amazon VPC) [163].

5.5.7 Content Delivery

Content delivery is provided by Amazon through its CloudFront service.

5.5.7.1 Amazon CloudFront

In addition to being Amazon's principle content delivery vehicle, CloudFront also performs crucial integration functions among other of the provider's web services, enabling middle-users, such as developers and other businesses to achieve rapid data transfer with low latency in distributing content to end-users. CloudFront is capable of delivery everything from the simplest application to the most complex complete website, one that integrates streaming content with both the static and the dynamic. Performance is optimised through automatic routing of content and requests to the closest in a global edge location network [154].

Amazon CloudFront works optimally, not only with Amazon's S3, EC2, ELB, and Route 53, but also, and equally seamlessly, with services and applications on server, even those not created by Amazon. CloudFront accomplishes this in part by storing user files in their original, definitive versions. Amazon CloudFront is provided on the same pay-per-use basis as all other Amazon Cloud-based services [164].

5.5.8 Database

5.5.8.1 Amazon SimpleDB

The Amazon web service interface, SimpleDB, supports users in: 1) creating multiple, distinct and varied data sets, 2) store said data set for easy access at a later point in time, 3) query the data from various sets with ease, and 4) retrieve results efficiently. The automatic data indexing function built into Amazon SimpleDB ensure swift and effortless access to information in response to users' requests. Schemas neither need to be pre-defined before setting up a data set for storage or manipulation not does the user need to change schema in order to add data subsequently. Moreover, scaling-out involves merely the creation of a new domain, as opposed to the construction of entirely new servers [154]. Architecturally, Amazon has organised SimpleDB around a data model consisting of three hierarchically nested concepts: 1) Domains which refer to each collections of objects to be stored, 2) Items, which indicate the objects themselves that will go into the database, and 3) Attributes, which refer to all the defined elements that make up each Item. In terms of nomenclature more widely if imprecisely used, domains are analogous to relational tables, in which items would be inserted into the rows while attributes would go into the columns. Attributes can also be seen as key value pairs, with the understanding that a given key may have more than one value. In this framework, a type string would consist of the data type supported by the key and its values. Within a domain, each item is defined by its having a unique key in contrast to all other items in that domain, obviating the need for the type of rules that are essential to the creation of tables in schema based relational databases [165].

The Amazon SimpleDB avoids several disadvantages of standard, clustered relational database organisation, namely the significant initial capital expenditure required, the complexity and corresponding difficulty in designing it, and the need for significant, on-going administration of an often repetitive nature (e.g., modelling of data, maintaining of indexes, and performance tuning of the system). The simplicity of SimpleDB is impressive in that, it dispenses with the need for schema, it deals with data indexing automatically, and it utilises a simple API to guarantee efficient, high speed storage and access. With the reliability of the Amazon, SimpleDB provides developers and other users with the ability to scale up or down at a moment's notice while taking advantage of all the functionalities of the system on a pay-per-use basis, as well as integrating compatibly with Amazon S3 and Amazon EC2 cloud services. This insures the all-but-flawless cloud-based processing, storage, and querying of data sets in real time. Furthermore, the core functionalities in SimpleDB guarantee and ease of use in the accessing and querying of structurally organised data in an operationally direct, non-complex manner [154].

The simplicity of Amazon SimpleDB structure, as explained above, is well adapted for storing relatively small units of textual information, making the access, management, and modification of such data easy. Users whose applications need only a relatively simple database to function effectively can dispense with traditional relational database such as the relational database management system (RDBMS) servers and the attendant problem of their acquisition and upkeep. The primary goal in the creation of SimpleDB was in fact to reduce complexity and administrative overhead of data management, especially by dispensing with pre-defined schema, allowing users in turn to adjust or reconfigure the SimpleDB's structure and content. Additionally, users' data is indexed for speed and ease in querying, as well as for safe, secure, redundant storage at one of the undisclosed-location data centres run by Amazon [159].

As the name suggests there are some limits to SimpleDB; according to Dewan and Hansdah [166], an individual domain can only accommodate 10 GB, each item being replicated into a set of nodes housed in its Amazon data centre. In SimpleDB, an item can incorporate at most 256 attributes, for a combined maximum item size of 1 MB. Domains are limited to 1 billion attributes in a domain; moreover, each SimpleDB account cannot exceed 100 domains. To the informed user, these limits, however, represent a large quantity of space and capability.

5.5.8.2 Amazon Relational Database Service

For those users who need relational databases and want to use them for cloud computing, Amazon has developed its Relational Database Service (Amazon RDS) with all the features and capacity of MySOL, the well-known database, designed specifically for ease in set up, operation, and scaling either upward or downward. At the same time, Amazon RDS has incorporated cost-efficient, resizable capabilities and the handling of time-consuming administration for the database, so that companies and organisations can devote all their attention to their applications and businesses. The congruence of Amazon RDS with MySQL ensures that users will be able to work seamlessly with the code, applications, and tools with which they are already familiar. While the user gets to define the duration of data retention, the Amazon RDS automatically handles the patching of software, in addition to the backing up and storing of users' databases [154], [167].

The feature which allows the calling up of instances with a single API helps to provide users with enhanced flexibility, as does the replication feature of Amazon RDS which also adds availability, scalability, and reliability to relational databases. As with all AWS, service is provided on a pay-per-use basis [167].

5.5.9 Deployment & Management

5.5.9.1 Amazon CloudWatch

Yet another service in demand by both developers and system administrators is monitoring of the applications run by users, along with the resources consumed in such running; the is the function of Amazon's CloudWatch. The metrics which CloudWatch follows and reports on work in the short run to allow immediate response to glitches in the system that would disrupt its functioning smoothly; beyond that, the service provides insights into systems that in the long run enable innovations and improvements. Not only does CloudWatch automatically monitor other Amazon services, such as Amazon EC2 and instances within Amazon RDS DB, but it also is capable of tracking custom specified metrics from an organisation's applications and services revealing details of their health and performance, as well as the degree of their utilisation of computing resources. As with all other Amazon services, CloudWatch can be deployed, and operating scalable, reliably, and flexibly, can deliver results within minutes, getting rid the chore for organisations of setting up their own monitoring systems. Best of all, CloudWatch can provide any selected level, type, quantity of monitored data, in various formats, including graphs; it can also create alarms and automatically notify clients according to their pre-chosen criteria. Additionally, CloudWatch can inform users of trends usage and act automatically to deal with predetermined circumstances or conditions according to the status of the cloud environment [168].

5.5.9.2 AWS CloudFormation

Both system administrators and developers frequently need to bring together and coordinate the functioning of a variety of the different type of resources which AWS makes available, all in an organised, predictable manner; Amazon's CloudFormation is design to accomplish that specific service. AWS resources can be designated, along with their application's associated dependencies and runtime parameters with the help of provided template for easier organisation. Tasks, such as the order of resource provisioning and dependency control are handled by CloudFormation, while the user remains able to update and modify specifications for AWS resource stacks, even updating the template as needed, using any of the following: the management console, the command line tools, or the APIs. As CloudFormation comes standard with AWS usage, as opposed to being a separate service of AWS, the user incurs no addition charges for drawing on its functions [169].

5.5.10 Application Services

5.5.10.1 Amazon Simple Queue Service

Between computers, traffic of any significant density requires a message handling queue that is both highly reliable and scalable. Amazon SQS ensures that data can be moved between the frequently distributed components of an application regardless of how diverse the tasks are, with no loss of messages whether or not the components involved are currently available. Amazon SQS coordinates tightly with other AWS offerings and infrastructures, Amazon EC2 in particular, to enable the automation of workflow within any operation, not matter how complex or distributed. It does so by taking the AWS web-scale messaging infrastructure and making it transparent and accessible in the form of a Web based service, obviating the need for reconfiguring firewall specification or installing special software. Moreover, with Amazon SQS, components can coordinate even if they are from different networks or created with different technologies, running independently or not simultaneously [154], [170].

5.5.11 Networking

5.5.11.1 Amazon Route 53

The central function of Amazon Route 53 is to transpose names that humans can deal with, hypothetically for instance www.example.com, into the machine readable versions such as the equivalent 192.0.2.1, which computers need to communicate with one another. Employing a global network of Domain Name System servers (DNS), Route 53 processes DNS requests, routing them to the closest DNS server in a rapid, simple to navigate, and cost efficient manner. It is this service that connects the user with other AWS, such as Amazon EC2 instances, Amazon S3 buckets, or Elastic Load Balancers;

on the other hand, Route 53 is just as effective with Amazon external infrastructure. Furthermore, through either the AWS Management Console or user-friendly APIs, an organisation is able to set up and control its public DNS records and can, moreover, regulate organisation internal access and authorisation to manipulating said records, through coordination with AWS's Identity and Access Management (IAM) service. In keeping with the AWS standard charges, users of Route 53 are only charged for the management of domains through the service and for the number of queries answered by Route 53 [171, p. 53].

5.5.11.2 Amazon Virtual Private Cloud

As hybrid cloud computing is a significant part of the cloud environment, Amazon created its Amazon Virtual Private Cloud (Amazon VPC) to connect users with existing infrastructure of their own to selected (i.e., by the user) AWS offerings in a seamless, secure manner in complete isolation through the use of an Amazon virtual private network (Amazon VPN) connection. The organisation has all of its own privacy and security features, such as firewalls, intrusion detection applications, DNS, LDAP, Active Directory features in operation, as a system into which the Amazon VPC fits and with which it conforms. This arrangement is in addition to Amazon's security, privacy, and integrity measures [172].

Through Amazon VPC, the client defines and creates the virtual network of AWS resources and services, including configuring its topology to conform to that of the user's own data centres. This level of user control extends to the ability to set the IP address range, to create subnets, and to configure both network gateways and route tables. A company could, for instance, set up a sub-network with open public-facing Internet access involving their web servers, while for the company's databases or application servers there would exist another private-facing sub-network, which lacked Internet access. In this scenario, access to Amazon EC2 instances could be controlled differently in each subnet [154], [173].

An Amazon VPC user, typically a sizeable business or organisation, is able to establish a hardware VPN and connect the network to the organisation's data centres and integrated AWS services, such as Amazon EC2, Amazon Elastic Block Store (Amazon EBS), and CloudWatch, in effect making AWS an extended arm of the organisation's own data centre, while being charged only for the quantities of resources consumed [173].

5.5.11.3 AWS Direct Connect

The purpose behind AWS Direct Connect is to allow users to create individual dedicated network connections from their physical locations to the AWS, meaning that they set up completely private connectivity with AWS and whatever portion of the users' own infrastructure, be it their office, their data centre, or any collocation environment over which the user have control. The advantages to the user of doing so include: 1) reducing the cost of network operations, 2) increasing bandwidth throughout, and 3) providing a greater degree of consistency in networking than what an Internet connection offers. Beyond all these benefits, AWS Direct Connect enables the user to maintain a dedicated connection, which can be subdivided into numerous logical connections, all specifically linked to one of its AWS Direct Connect physical locations, by means of 802.1q virtual local area networks (VLANs). By means of AWS Direct Connect and its dedicated connections, public resources, e.g., data stored in Amazon S3 may be accessed together with private resources from Amazon EC2 and Amazon VPC all while utilising both public and private IP spaces, preserving network separation. At the same time, reconfiguring is possible at any time in the interest of changing external demands or internal desires [174], [175].

5.5.12 Payments & Billing

5.5.12.1 Amazon Flexible Payments Service

As described earlier, one of the time and hassle saving features among AWS is that developers and other customers who in turn provide cloud-based services to end-users can easily handle accounting, billing, and payment operations, for which Amazon Flexible Payments Service (Amazon FPS) has been designed. For end-users who are also Amazon customers, the intermediary service provider makes use of the existing infrastructure of Amazon's internet retail business for billing, while the end-user makes payments using his or her pre-existing Amazon account just as that user would for any other direct purchase through Amazon. Not only can developers and similar individuals and organisations bill and receive payments for cloud computing services provided, but they can also do the same for any goods and services not provided through the cloud, as well as such activities as soliciting charitable contributions, setting up and implementing recurring payment schedules, and paying other non-Amazon connected vendors [176], [177].

The flexibility of structuring in terms of payment instructions, whether standard and recurring or on an ad hoc basis, including the ability of both senders and recipients of payments to set conditions and constraints provides time and cost saving benefits to both providers and consumers of goods and services alike. For example, a per-week payment limit can be placed on an account by either party, such that the designated recipient alone could withdraw funds and only up to the specified limit. The APIs that Amazon makes available are simple to integrate, lightweight and organised as packages, which Amazon labels Quick Starts; working in tandem with Amazon's SDKs and sample code to create enhanced documentation and ultimately enhanced convenience in making application-based payments [159].

5.5.12.2 Amazon DevPay

Created with developers in mind, the Amazon DevPay service enables mid-users (i.e., developers) to provide cloud- based applications for other end-users by subscription or on-demand without the developer having to create and manage its own billing and account management systems. Automatic customer sign-up and service metering is accompanied by having Amazon handle the billing and collections according to mid-user determined specifications. The fee and payment structure may involve anything from up-front to recurring to usage-based and is made available by Amazon through a simple Web interfacing, providing an advantage to the end-user as well. Both the end-user and the developer profit from the reliability and reputation of trust in Amazon Payments and the infrastructure behind it; moreover, if the end-user has an existing Amazon account, the payment process for them is all the more streamlined [178], [179].

5.5.13 Storage

5.5.13.1 Amazon Simple Storage Service

The goal behind the Amazon S3 is to provide developers with storage connected to and accessible from anywhere on the Web, and to be capable of not only storing but also retrieving and manipulating data not matter how large the quantities, and of course to accomplish this virtually instantly at any time. Amazon S3 carries with it all the features of total scalability, along with the reliability, security, speed and, low cost efficiency because it uses the same infrastructure as developed for the Amazon world-wide retail network [159]. One of the earliest cloud-based services available, as noted in [166], Amazon S3 is configured around an architecture defined by buckets, which compare with directories that can hold one or multiple objects. The objects are in reality up to 5GB of content storage. While both buckets and objects may be freely created by the user, nesting is not permitted, keeping the hierarchy as simple as possible Moreover, although object cannot be updated in situ, they can be freely be moved, renamed, undone, and redone through copying. Through all this replication keeps all objects readily available and accessible.

Not only does Amazon S3 handle any type of data imaginable, but it provides security through all phases of infrastructure creation, maintenance, and back-up, the security needs of which the organisation would otherwise have to deal with. While relieving the user of these concerns, Amazon S3 insures ready access from anywhere on the Web to the data on the part of any application or individual, to whom the organisation grants authorisation. The quantity of data stored, the length of time stored, the bandwidth available for data transfer and publishing all come without limits in as much as Amazon S3 stores data reliably in a distributed fashion across any or all Amazon centres by means of a storage API, designed to work in the simplest for rapid access on a massive scale with complete user control [154].

5.5.13.2 Amazon Elastic Block Store

The potentially large volumes of storage needed in connection Amazon EC2 are provided through the company's Amazon EBS, the storage of which is off-instance, continuing to exist as long as the instance does. The storage volume can be attached to anything running on Amazon EC2, as a device within that instance, making Amazon EBS very efficient for applications necessitating file systems, databases, or block level storage in the raw [180].

Within a given availability zone, volume data in Amazon EBS gets replicated through numerous servers rendering it both reliable and readily accessible. Volume durability is set based on the volume at, as well as the percentage change in data volume since, the last snapshot. The Amazon EBS can claim a ten-fold advantage over commodity hard disk in terms of annual failure rate (AFR), at one tenth to one half a per cent, as compared to four per cent for the hard disk, given less than 20 GB of modified data since previous snapshot [181].

5.5.13.3 AWS Import/Export

The function of the Import/Export component of AWS is to make use of portable storage devises to transport significant quantities of data into and out from its services; this it does by direct transfer to and from the devices through the high-speed internal network of Amazon itself, avoiding the need to migrate by means of the Internet, achieving saving in terms of both time and costs. In the US Standard, US West (Oregon), US West (Northern California), EU (Ireland), and Asia Pacific (Singapore) Regions, Import/Export facilitates data transfer in relation to the company's Amazon S3 buckets, while in the US East (Virginia), US West (Oregon), and US West (Northern California) regions, it functions similarly for the company's EBS snapshots [182].

5.5.14 Workforce

5.5.14.1 Amazon Mechanical Turk

The concept behind Amazon Mechanical Turk (AMT) is to create a marketplace that connects businesses and organisations with a workforce of actual people who are available to perform tasks for which human intelligence is needed over computing. By leveraging this service, a developer or other business client can integrate human services, such as identifying elements in videos or photographs or selecting from several photos the optimal one for a given purpose, transcribing spoken or sung texts or identifying the speaker or performer in audio recordings, eliminating duplications of data, or conducting research into details within the data. Having the AMT eliminates the obstacles of time and expense inherent in utilising a large workforce on a temporary basis [183]. A Human Intelligence Task (HIT) is the designation in this service for what the business or organisation needs to have done, and the APIs of the AMT service provides the access thousands of workers worldwide, who are known in the AMT terms of service as Providers and who in turn provide a high calibre of work, on demand at a low cost. The companies and developers can then, integrate the results into their business systems, processes, products, and services. In order to insure competence, Requesters (AMT terms of service) have the ability to set and verify the qualifications of Providers, by testing, and accept or reject the sample work of a given provider. The workers, for their part, can browse through requests left on open APIs or the AMT requestor site, complete the tasks, and receive payments from the requester through Amazon, much as occurs within other parts of AWS. As fees are set by the requester, Amazon collects a fee amounting to ten per cent of the payment for completed work [184].

5.6 Chapter Summary

Cloud computing is a disruptive innovation. Like the Internet business boom of the 1990s, there is currently a boom in terms of companies of all sizes getting into cloud vendor services. Of particular interest are the business models by which the major established firms in the field are seeking to expand into the cloud. Having introduced cloud computing in the previous chapter (Chapter 4), in this chapter we looked at specific cloud services available in the market today. There are already literally dozens of provider of cloud services from among which a company or organisation may choose. We selected four major cloud players and reviewed the services offered by them; IBM, Microsoft, Google and Amazon were reviewed in Sections 5.1, 5.2, 5.3, and 5.4, respectively. Subsequently, in Section 5.5, we discussed the specific cloud services offered by Amazon in some detail. Amazon cloud services were selected because Amazon is currently among the top cloud services vendor who has made in good detail the services related information available on its website in the public domain. The explanation provided for Amazon cloud services in this Chapter will also be useful in the later chapters where we model and analyse Amazon cloud services.

Section 5.5 was divided into two parts. The first part, comprising Sections 5.5.1 to 5.5.5, discussed Amazon cloud services' characteristics that make it among the top cloud computing vendors. This part focussed on features or characteristics of AWS that distinguish Amazon as a cloud services vendor or providers from others in the field. Principle characteristics, discussed in the first part, that AWS can offer were categorised into Flexibility, Cost Effectiveness, Scalability and Elasticity, Security, and Experience.

Within the discussion of Security, the text analyses four components, namely certifications and accreditations, physical security, secure services, and data privacy.

The specific sets of cloud computing services offered by Amazon, featuring the branch of its business known as AWS, as on May 2012, were reviewed in the second part of Section 5.5, comprising Sections 5.5.6 to 5.5.14. The second part provided a detailed, one-by-one survey of Amazon's different Cloud service offerings. These included: Amazon EC2, Amazon EMR, Amazon CloudFront, Amazon SimpleDB (SDB), Amazon RDS, Amazon CloudFormation, Amazon SQS, Amazon CloudWatch, Amazon Route 53, Amazon VPC, Amazon Direct Connect, Amazon ELB, Amazon FPS, Amazon DevPay, Amazon S3, Amazon EBS, Amazon Import/Export, and AMT. These service offering were grouped according to general types of services they fit into. Amazon EC2 and Amazon EMR are classified as basic computing services while CloudFront, which may not be as directly approached by some users, was labelled under the heading of content delivery. In relation to Amazon SimpleDB and Amazon RDS offerings, the discussion centred on the simplicity of the ways in which these databases are organised. Beyond databases, more general storage is handled by Amazon S3, and Amazon EBS.

There are various components of the described AWS function to keep other services flowing smoothly, both within the Amazon platform and outside on the Web. Among these 'behind-the-scenes services, Amazon SQS, discussed under the subheading of messaging, keeps messages organised and moving appropriately between distributed parts of the system, while Route 53, described under the networking subheading, translates human readable names, such as URL website addresses, into machine readable form, normally numerical strings. Other such services include Amazon ELB, CloudFront (as mentioned above), Import/Export, and CloudFormation. Others, such as CloudWatch which provided numerical and similar types of cloud application monitoring, are engaged by the user as essential, yet indirect aids to the user primary efforts in business or otherwise.

Many of AWS's services are either based upon or directly utilise infrastructure and service originally developed for the Web-based business for which Amazon is world famous. For instance, Amazon FPS, as well as DevPay, capitalise on the infrastructure created for Amazon.com as the billing and payments arm of its global, Internet based retail sales empire. AWS even extends Amazon's model of linking buyers with sellers as part of its on-line retail business by creating AMT to connect cloud computing users with actual people who can provide computer-related service not easily down by machine.

A number of points were repeatedly reinforced throughout the material provided in this chapter, in particular, in Section 5.5, where we discussed Amazon cloud services. First, the financial incentives for organisations to use cloud services include avoiding the necessity and expense of having to obtain hardware and software, to plan out systems, and to install infrastructure, along with all the similar start-up costs of new operations. Beyond these expenses in getting started establishing or upgrading one's own traditional data centre based system, AWS is presented as an alternative which eliminates or greatly reduces the costs in involved with administering and maintaining one's own system and the costs of having underutilised capacity in reserve for atypically large demand, in addition to eliminating the opportunity cost of not being able to move rapidly and flexibly to stay abreast of changing market conditions and opportunities. Moreover, the pay-per-use model which covers all of AWS is stressed throughout the discussion, together with numerous of the individual AWS support services which do not incur additional charges beyond resource usage, as defined by computing power, storage, or querying. Second, the flexibility that AWS offers customers, as the result of instantaneous scalability both up and down on-demand, is described and highlighted in every subsection of the discussion. This feature is tied into the aforementioned adaptability advantages conferred upon users. Third, most of the discussions of individual AWS offerings at least mention, and often highlight the simplicity of their architecture and design. This simplicity is asserted in order to make the case for the ease of their operation by users on one hand, and their integration with different AWS components, as well as Amazon external computing resources on the other. Fourth, all of AWS's offerings are efficient, reliable, and secure because they have the reputation and established track record of the Amazon.com, the parent company, standing behind them.

We are now ready to propose and describe, in the next chapter, a disaster management system which exploits the benefits offered by cloud computing. The material presented in this chapter will also be useful in Chapter 7 where we model Amazon market sectors, applications, and workload.

Chapter 6: An Intelligent Cloud Based Disaster Management System

"Ensuring the success of mass evacuations—The conferees direct the Department of Transportation (DOT), in cooperation with the Department of Homeland Security (DHS), to assess mass evacuation plans for the country's most –high-threat, high-density areas and identify and prioritize deficiencies on those routes that could impede evacuations..."

Departments of Transportation & Housing & Urban Development and Related Agencies Appropriations Act, 2010 Conference Report (111-366) to Accompany HR 3288 & Public Law 111-117, FY 2010 Consolidated Appropriations Act [185]

The importance and scope of emergency response systems have grown tremendously over the past decade in particular after September 11, 2001. Disasters, manmade and natural, are a cause of great economic and human losses each year throughout the world. The overall cost of the recent Japan earthquake and tsunami disaster alone is estimated to have exceeded 300 billion USD. This has driven many new initiatives and programs in countries throughout the world, in particular in the US, Europe, Japan and China.

Transportation and ICT technologies play critical roles in responding to emergencies and minimising disruptions, human and socioeconomic costs. We have witnessed unprecedented advancements in ICT over the last few decades and the role of ICT technologies in Intelligent Transportation Systems is to grow tremendously. Vehicular Ad hoc Networks (VANETs), sensor networks, social networks, Car-to-Car (C2C) and Car-to-Infrastructure (C2I) technologies are enabling transformational capabilities for transportation. Our ability to monitor and manage transportation systems in real-time and at high granularity has grown tremendously due to sensor and vehicular network that generate huge amount of extremely useful data. However, many challenges in realising the Intelligent Transport Systems (ITS) potential remain including the interworking and integration of multiple systems and data to develop and communicate a coherent holistic picture of transportation systems. This is particularly difficult given the lack of data and systems interoperability as well as the business models to develop such an advanced infrastructure which requires coordination between many stakeholders and general public.

Cloud Computing has emerged as a technology, coupled with its innovative business models, which has the potential to revolutionise the ICT and ITS landscape. It is already making a huge impact in all sectors through its low cost of entry and high portability / interoperability. Moreover, the technology allows one to develop reliable, resilient, agile, and incrementally deployable and scalable systems with low boot-up and migration time, and at low costs, while giving users access to large shared resources on demand, unimaginable otherwise.

In this Chapter, we leverage the advancements in the ICT technologies, including ITS, VANETs, social networks, mobile and Cloud computing technologies, and propose an intelligent system for the management of disasters in urban environments. The particular focus of the work presented in this Chapter is on using distributed computing and telecommunication technologies to improve people and vehicle evacuation from cities in times of disasters. By exploiting state of the art technologies, the system is able to gather information from multiple sources and locations, including from the point of incident, is able to make effective strategies and decisions, and propagate the information to vehicles and other nodes in real-time. The Cloud system architecture is described and the traffic models used to provide the transportation intelligence are explained. The effectiveness of the proposed intelligent disaster management system is demonstrated through modelling the impact of a disaster on a real city transport environment. We model two urban scenarios: firstly, disaster management using traditional technologies, and secondly, exploiting our computationally intelligent, VANETs Cloud based disaster management system. The comparison of the two scenarios demonstrates the effectiveness of our system in terms of the number of people evacuated from the city, the improved traffic flow and a balanced use of transportation resources.

Although the work presented in this Chapter focuses on disaster management, our research in this domain is broadly concerned with developing emergency response systems for disasters of various scales with a focus on transportation systems which exploit ICT developments.

The contribution of this research includes a novel Cloud-VANET based distributing computing system architecture, and its associated models, algorithms, technologies and software for the simulation and evaluation of the disaster management system. In this context, the specific contribution of this thesis is the development of a novel multi-disciplinary cloud computing based system, its architecture and system performance evaluation. Further work on the development and evaluation is in progress. The system is being analysed using additional cities, environments and scenarios. This work is continuing to make impact and has resulted into developing international collaborations, one invited (refereed) conference paper [2] and another (refereed) book chapter [3].

This Chapter is organised as follows. Section 6.1 introduces Intelligent Transportation Systems and Vehicular Ad hoc Networks (VANETS). Section 6.2 establishes motivation for research in emergency response systems through a substantive literature review structured into three areas. A historical perspective on emergency response and evacuation initiatives in the US is provided in Section 6.2.1. Policy, Advisory and Survey based approaches to emergency management systems are reviewed in Section 6.2.2, while research focussed on specific technologies is surveyed in Section 6.2.3. Our intelligent disaster management system, its architecture and intelligence layer are described in Section 6.3. This section also describes the city and its transportation network. Section 6.4 presents the analysis of the disaster management system. Finally, the chapter is summarised in Section 6.4.4.

6.1 Intelligent Transportation Systems

The inherent human desire for change, progress, mobility, entertainment, safety and security are leading the way to the development of intelligent transportation systems (ITS). Vehicular ad hoc networks (VANET) are the most prominent enabling technology for ITS. VANETs are formed on the fly by vehicles equipped with wireless communication capability. The participant nodes in VANET interact and cooperate with each other by direct communication with the nodes within range, by hoping messages through vehicles and road side masts. Traditionally, information about traffic on a road is only available through inductive loops, cameras, roadside sensors and surveys. VANETs provide new venues for collecting real-time information from onboard sensors on vehicles and for quick dissemination of information. The information collected through individual nodes participating in VANETs can be integrated together to form a real time picture of the road situation. Many new applications have been enabled through VANETs, though safety and transportation efficiency applications are the most important driver for VANETs. The various ITS stakeholders such as governments, telecommunication companies and car manufacturers are working together to make VANETs based ITS a reality. Hundreds of projects are underway in the US, Europe, Japan, China, Singapore and other countries in the world helping with research, innovation, testing and standardisation activities [186], [187].

6.2 Emergency Response Systems - Literature Review

The importance of emergency and disaster response systems can be evidenced by huge literature that is available in this area. To motivate, we give in Section 6.2.1, a historical perspective on emergency response and evacuation in the US. A lot of the literature found on the web comprises advisory and policy documents developed by various government authorities through surveys, consultations, experiences and other means of research. These are discussed in Section 6.2.2. The other literature on emergency response systems is based on technology-led problem-specific proposals: such proposals are focussed on the use of technologies such as vehicular networks for specific problems. These works look at smaller, a narrowly focussed, problems in isolation with the whole system perspective. These are discussed in Section 6.2.3.

6.2.1 Historical Perspective on Emergency Response Initiatives

In the US, the work on emergency response systems has been underway under different initiatives including the Traffic Incident Management (TIM) program. An 'incident' was defined in 2000 as "any non-recurring event that causes a reduction of roadway capacity or an abnormal increase in demand" [188]. However, major events such as September 11 in 2001 and hurricanes Katrina and Rita in 2005 have broadened the scope of TIM to the broader theme of national preparedness. In 2003, the U.S. Department of Homeland Security (DHS) was given responsibility to develop and administer the National Incident Management System (NIMS), a framework for incident planning and response, at all levels, regardless of cause, size, or complexity. In 2004, the US Federal Highway Administration (FHWA), National Highway Traffic

Safety Administration (NHTSA) and Federal Transit Administration (FTA) collectively launched the Emergency Transportation Operations (ETO) program to improve transportation safety and effectiveness of incident management and evacuation through provisions of tools, information and dynamic partnership across various departments and communities [189]. The ETO functions encompassed six areas including evacuation management and operations, enhanced information sharing, and public access to emergency services.

Evacuation operations happen on a daily basis and may involve a single building, a neighbourhood or an entire city. In the US alone, each year, over 400 tropical storms, hurricanes, tornadoes, and highway hazardous material incidents require evacuation of 1,000 or more people every 2 to 3 weeks. Evacuation management to date remains very high on governments' agenda as natural and manmade disasters continued to hit the world. The US Congress requested last year the department of transportation (DOT), in cooperation with the DHS to assess mass evacuation plans for the country [185].

6.2.2 Policy, Advisory and Survey based Approaches

The Government of Davidson County have conducted a survey to find out the response level of the people in order to support the government and the decision makers in their obligation on the systematic provision of integrated emergency response system. The results from the survey were presented in [190]. The County has been conducting these surveys for several years and this has enabled the survey to become more reliable and sophisticated over the years. The survey aimed to develop an understanding of Public opinion in seven major areas: these are Emergency Situations, Evacuations, Personal Preparedness, Employment, Schools, Knowledge of Government Actions and Demographic information. The findings of the survey were summarised along with visualisations of the data for each of the seven major areas. The majority (99%) of the county citizens stated that they have experienced emergency situations previously they were aware of the associated inconveniences and challenges; and these results were visualised for the type of emergency situations (including tornado, earthquake, vehicle crashes, war, bomb threats, robberies), the citizens' responses and involvement in the emergency situation, the period in which they experienced the emergency situation and the age distributions of those responding to the survey. However, despite the prior experiences of emergency situations by 99% of the citizens, they were not very well-prepared to respond to an emergency situation.

The role of transit in emergency evacuation was studied and reported in [191]. The study, which was conducted by committee established by the Transportation Research Board (TRB), evaluated the role and capacity of transit systems to accommodate the evacuation of people (egress and/or ingress) in emergency situations with a focus on transit systems serving the 38 largest urbanized areas in the US. An extensive review of the relevant literature was carried out. Furthermore, to add to the committee's plan assessment and present recommendations, five case studies were conducted, and the case study sites were selected based on a number of criteria including areas with different types of transit systems, face different types of emergencies, are located in different regions of the country, have a high percentage of special-needs population, and would experience different jurisdictional issues. Transit can play a critical role in emergency evacuation situations particularly for those with special needs and/or lacking private means of transport. The factors that are likely to affect the role of transit in an emergency evacuation were discussed in detail. These factors included, among others, the area characteristics, the nature of emergency, the preparedness and willingness of population to accept orders and use transit, availability of resources, the quality, type and properties the transit systems being used. The committee concluded that most plans for evacuation were inadequate for managing major disasters and the inadequacies were reported. A particular concern was the insufficient integration of transit and other modes of transportation with emergency evacuation strategies. The critical factors in enhancing transit's role in evacuation situations and the limits of transit systems in these situations were given. The recommendations on measure required to increase the capacity and resilience of the overall transportation system were provided. The committee also presented their recommendations on research support and actions required at the federal, state, and local levels.

Buckland and Rahman [192] examined the relationship between community preparedness and response to natural disaster and their level and pattern of community development. They found supporting evidence that the level and pattern of community development affect community capacity to respond to disaster. Communities characterised by higher levels of physical, human and social capital were better prepared and more effective responders to the flood, however, the decision-making processes were complicated for such communities.

Rob Drake, the Mayor of the City of Beaverton, OR, the western neighbour of Portland, reports in [193] on his nightmares due to his worries in relation to the preparations for any potential disasters hitting his city. The Mayor discusses local threats, the hierarchy of emergency preparation, individual and organisational and local government cooperation, and technology cooperation. He emphasises the importance of, and the leadership role that the federal government should take in exploiting interoperable communications technologies in preparing for and responding to disasters. He further notes the critical requirements for development and testing of emergency response systems and the need for teamwork and discipline. He wisely notes:

- ✓ "The only time to successfully build the cooperation that a crisis will demand of us is before it occurs".

6.2.3 Proposals Focussed on Specific Technologies

Buchenscheit et al [194] propose an emergency vehicles warning system that exploits vehicular network technologies. The approaching emergency vehicles could transmit radio signals and detailed route maps to the vehicles and signals in their path in order for those vehicles and infrastructure to take appropriate and timely action. A system prototype has been built and tested in traffic environment comprising emergency vehicles and traffic signals. An approach to disseminate spatio-temporal traffic information in order to reduce chaos in evacuation scenarios using VANETs is presented in [195]. Precisely, their approach provides emergency vehicle path clearing and real-time resource availability to minimise chaos on evacuation and emergency vehicle routes without fully relying on any message relaying infrastructure. This work is further extended by the authors in [196] by exploiting Wireless Fidelity (WiFi) and Worldwide Interoperability for Microwave Access (WiMAX) to provide high end to end network connectivity and minimise network contention and interference. The proposed scheme is evaluated using simulations. Park et al [197] address the non-trivial problem of reliable transmission of multimedia data in VANETs for safe navigation support applications. Their approach is based on network coding and is evaluated using simulations.

Pazzi et al [198] take a distributed systems approach towards emergency preparedness and response systems and discuss the importance of service discovery protocols in these scenarios. In particular, they discuss the discovery of safety and convenience services, and service discovery architectures for VANETs including directory based, directory less and hybrid architectures. Serhani et al [199] propose a service discovery and reservation technique for mobile ad hoc networks (MANET) tailored to support disaster recovery and military operations environments. Their technique locates the resources taking service levels and requirements into account. They build a purpose built simulator to evaluate their technique and report its usefulness in locating and reserving services in varying network density, rate of requests and other operational conditions.

6.3 The Intelligent Disaster Management System

6.3.1 System Architecture

Figure 6.1 depicts the system architecture of the Cloud-enabled vehicular emergency response system. The system consists of three main layers. The Cloud infrastructure layer provides the base platform and environment for the intelligent emergency response system. The Intelligence Layer provides the necessary computational models and algorithm in order to devise optimum emergency response strategies by the processing of the data available through various sources. The System Interface acquires data from various gateways including the Internet, transport infrastructure such as roadside masts, mobile smart phones, social networks etc. As depicted in Figure 6.1, vehicles interact with the gateways through C2C or C2I communications. For example, vehicles may communicate directly with a gateway through Internet if the Internet access is available. A vehicle may communicate with other vehicles, road masts, or other transport infrastructure through point-to-point, broadcast or multi-hop communications.

The emergency response system provides multiple portals or interfaces for users to communicate with the system. The Public Interface allows any individual to interact with the system. The purpose is to interact with the system on one-to-one or group/organisation basis with the system, either to request or provide some information. Of course, an authentication, authorisation and accounting system is expected to be in place to allow and control various activities and functions. The Transport Authorities

Interface provides is a high-privilege interface for the transport authorities to affectively manipulate the system for day to day operational management. The Administrators' Interface provides the highest privilege among the system users and is designed for policy makers and strategists to enable highest level system configuration.

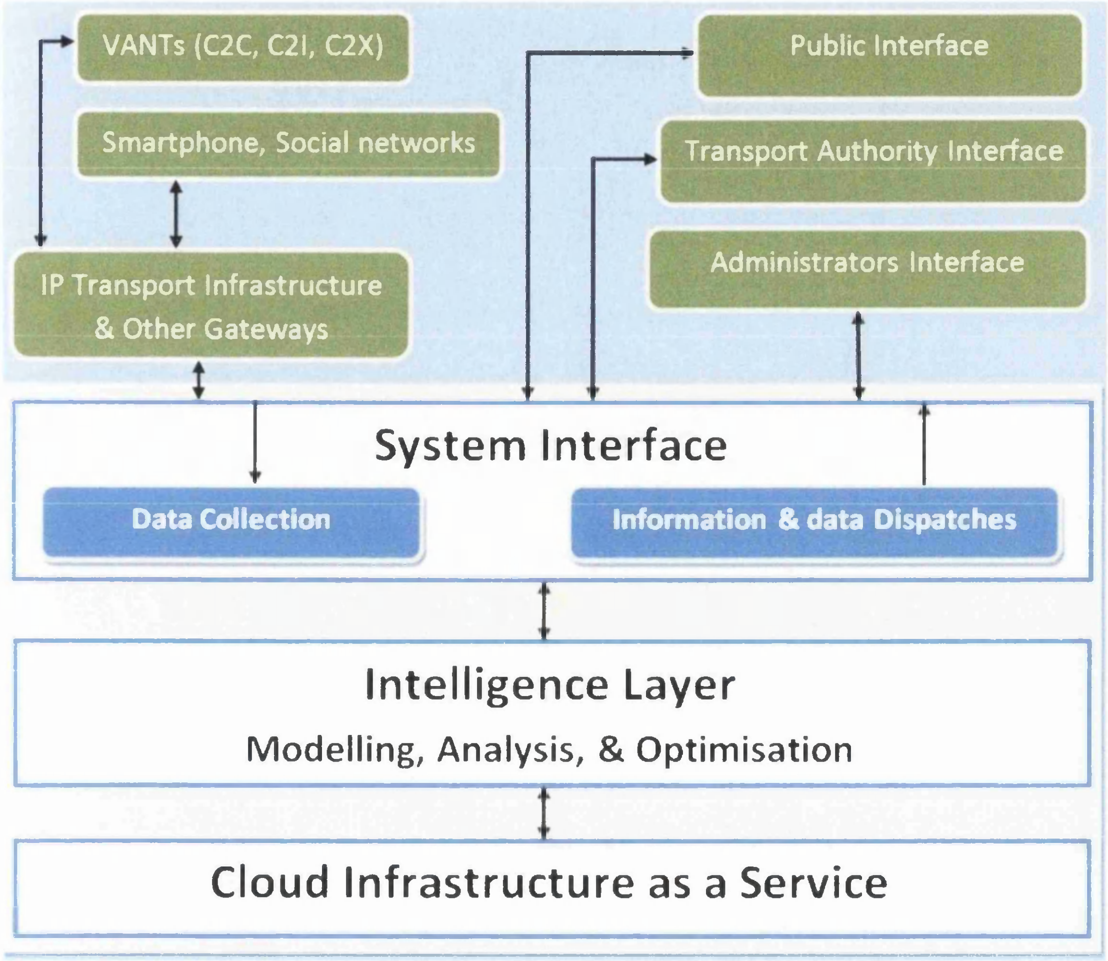

VANTs (C2C, C2I, C2X)
Smartphone, Social networks
IP Transport Infrastructure & Other Gateways
Public Interface
Transport Authority Interface
Administrators Interface
System Interface
Data Collection
Information & data Dispatches
Intelligence Layer
Modelling, Analysis, & Optimisation
Cloud Infrastructure as a Service

Figure 6.1: Emergency Response System - Architecture.

The motivation and background for a Cloud based distributed control system can be found in our earlier work, for example in [200], where we present an architecture for distributed virtualisation using the Xen hypervisor; it allows control and management of a distributed system by posting high-level queries on the system and their validation through real-time monitoring and control of the system. Monitoring relates to the acquisition of data and control relates to despatch of commands and decisions. Also see [201], where a Pervasive Cloud is proposed using the WiMAX broadband technology for railway infrastructure.

6.3.2 The Ramadi City and its Transportation Network

The Ramadi city (Al Ramadi) is the capital of Al Anbar Governorate and is situated at the intersection of the Euphrates River and Al Warrar Channel. The Habbaniya Lake is located a short distance to the south of the City of Al Ramadi. The present population of the city is estimated to be approximately 230,500, although the field survey results of a study in 2009 showed that Al Ramadi population is approximately 355,909.

The General Directorate of Physical Planning of the Ministry of Municipalities and Public Works (MMPW) is preparing for the Development Strategy of Al Ramadi, and it is in line with the development policies of MMPW for other Iraqi cities. The Association of the Canada-based Hydrosult Center for Engineering Planning (HCEP) and the Iraq-based Engineering Consultancy Bureau of Al Mustansiriya University has been commissioned by MMPW to carry out the tasks of this assignment and they have produced a second stage report [202] in November 2009 for the development of the Ramadi city. Iraq is now open to new developments and it is a great opportunity to develop intelligent transportation systems for Ramadi and other Iraqi cities.

Figure 6.2 shows the transportation network map of the Ramadi city, the network consists of zones, nodes, and links. The city is divided into 5 traffic zones; Zone 1 and Zone 5 are in the west side of Al-Warrar River which divides the city into two parts. Zone 1 represents the location of a huge glass factory; Zone 5 represents the west part of the city. The east part of the city contains Zones 2, 3, and 4. Zone 2 represents the old city centre which attracts high number of trips in the morning peak hour.

Note also in Figure 6.2 the two evacuation areas, Evacuation Area 1 and Evacuation Area 2. Their purpose is to provide an appropriate and safe location for the population in case a major disaster strikes the city and people need to be moved out of the city. The two evacuation zones are chosen in the north of the city, because there is an international roadway that joins the Iraqi borders with Syria and Jordan in the west with the capital Baghdad, and the evacuation zones are just a few minutes away from that road. In the south of the city there is the desert only and the connections of roadways in this area are very poor. If there is a need to give medical supplies or transport to injured and affected population, it will be best provided through the international roadway. The area in the east side of the city is mostly is for agriculture land use and is a private property. The west street leads to a nearby city about 30km away, this city has a good hospital but it can be best reached through the northern international roadway. We will discuss the city network and evacuation plan further in Section 6.4.

6.3.3 The Intelligence Layer

We have described the architecture of our proposed emergency response system in Section 6.3.1. Here we give details of the Intelligence Layer (see Figure 6.1). As mentioned in Section 6.3.1, our emergency response system has a System Interface that communicates with various user interfaces and communication gateways. This interface is used to gather and propagate data, information and decisions in order to carry out day to day transport management operations, policy implementations, and emergency response operations. The Intelligence layer consists of various mathematical models, algorithms and simulations, both stochastic and deterministic. These models accept transport related data received from various sensors such as inductive loops, intra-vehicular sensor networks, VANETs and C2I communications, and user interfaces. The data received from various sources goes through an internal validation layer before it is accepted by the modelling and analysis layer. The modelling or simulation algorithm is used for a particular activity based on the nature of the activity. In some cases, it is necessary and/or affordable to employ microscopic traffic models, for example in developing transport policies and procedures; this is due to the demands on higher accuracy and greater flexibility on the available time for decision, optimisation and analysis. In other cases, microscopic simulations may not be necessary, or may not be possible, due to the real-time nature of operations such as day to day transport management operations

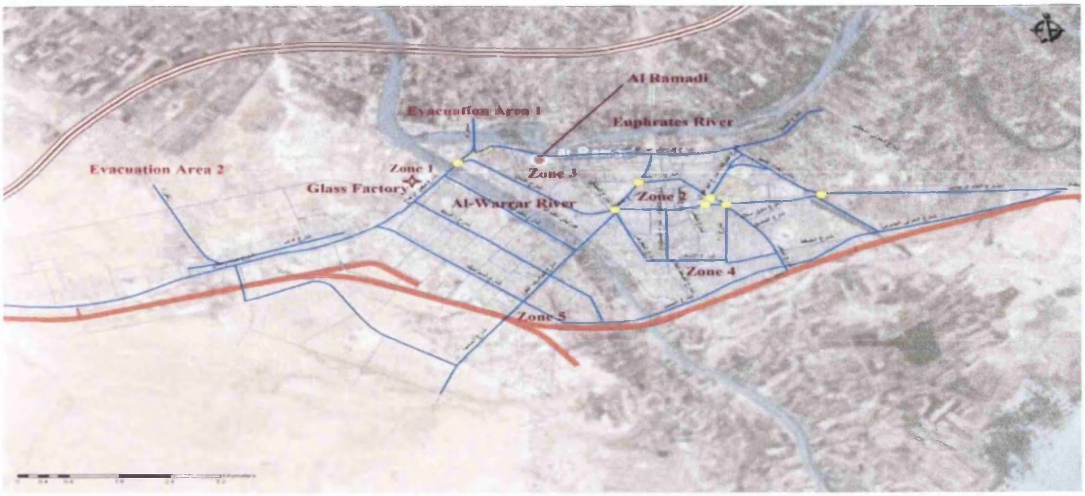


Figure 6.2: Transportation network of the Ramadi city.

Emergency response systems are an extreme example because, firstly, the availability of real-time data may be greatly limited due to the unavailability of many communication sources due to a disaster (e.g. broken communication links), and secondly, the time period in which the system has to act would be short. In such cases, macroscopic models which require relatively small computational time and resources may be the only option. Which models to invoke in a particular situation is an area of our on-going investigation and we will continue to improve on our automatic model selection algorithm. We will also be looking at ways of enhancing our distributed algorithms so we could invoke the most precise models for real-time critical situation such as great disasters. It is important to note here that we envisage a Cloud infrastructure which is virtualised and flexible to exist, or moved, outside the area affected by the disaster. This is possible considering the capability of Cloud technology. We focus on the topic of this Chapter, i.e. emergency disaster management, and as an example present here some details of a macroscopic model that is used for emergency situations where time to act is short and real time information is limited due to possibly broken communication links. For our work on other types of modelling, in general, see [186] for vehicular grid networks, [203], [87] and [204] for Markov modelling of large systems, and [205] for 3D virtual reality microscopic simulator. We consider the Lighthill-Whitham-Richards (LWR) model [206], [207], a macroscopic model, to represent the traffic in the city. The LWR model can be used to analyse the behaviour of traffic in road sections, and describe the dynamic traffic characteristics such as speed (u), density (ρ), and flow (q). The model is derived from the conservation law (first order hyperbolic scalar partial differential equation) by using the following equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} = 0,$$

where ρ is the traffic density in vehicle/km, and u is the traffic velocity according to distance x and time t. By using Greenshield traffic model, the relation between ρ and u could be as follows:

$$u\,(\rho) = u_{max} \times \left(1 - \frac{\rho}{\rho_{max}}\right),$$

where u_max is the maximum speed, and ρ_max is the maximum density. The fundamental relationship between flow, density, and speed is given by:

$$q = \rho \times u.$$

In this Chapter, we will use flow and volume interchangeably. Basically, the LWR model and the equations given above are able to model and predict the road traffic; the Origin-Destination (O-D) matrix (see next section) provides realistic data to start an LWR-based simulation which in turn stepwise provides traffic for the next time step. Traffic conditions from each discrete time step are fed into the next time step to predict the future evolution of the traffic under certain constraints (see Section 6.4.2.2 for various constraints). Having given here the mathematical description of the macroscopic model that we employ, in the next section, we will describe our approach for city evacuation in a disaster emergency situation.

6.4 System Evaluation

The transportation network of the Ramadi city consisting of zones, nodes, and links has been depicted earlier in Figure 6.2. We now make use of our earlier discussions in Section 6.3 and describe the disaster scenario in Section 6.4.1. Subsequently, in Section 6.4.2, we present analysis of the system and establish its usefulness as a disaster response system.

6.4.1 The Disaster Scenario

Consider Figure 6.2 which divides the Ramadi city into 5 zones. In Table 6-1, we quantify transportation trends of the city in terms of an Origin-Destination (O-D) matrix between the five city zones. The numbers of trips in the O-D matrix shown are in the mid-week period with natural conditions. These trips are calculated using the Fratar model. Note that the highest rate of trips is toward destination Zone 2 in the city centre.

Table 6-1: An O-D Matrix of the Transportation Network In Ramadi City

	zone 1	**zone 2**	**zone 3**	**zone 4**	**zone 5**
zone 1	0	0	0	0	0
zone 2	82	0	172	935	228
zone 3	172	2757	0	1171	108
zone 4	343	10026	381	0	248
zone 5	272	4835	358	1699	0

In Zone 1 lies a glass factory and beside it is Al-Warrar dam; both of these pose major risks to the city. We consider in this Chapter the risks related to the glass factory and Al-Warrar dam as a case study in order to describe and evaluate our emergency response system. The related potential risks for disaster events in Zone 1 are outlined below:

- ✓ Fire hazards at the main factory
- ✓ Technology failure due to shutdown of power plants that feed the city
- ✓ Explosion of hazardous materials in the glass factory
- ✓ Terrorist attack in the area of the factory
- ✓ The collapse of Al-Warrar dam adjacent to the factory

The above listed disaster events except the last one may last several hours before it will be controlled; for transportation planning purposes, special care is required to handle the emergency situation and saving peoples' lives.

We now focus on a disaster event which could happen in the glass factory. This event could be any one of the potential risks listed earlier in this section, e.g. fire or explosion of hazardous materials in the glass factory. The details of the event are as follows.

6.4.1.1 Timing of the Disaster Event

The traffic conditions in a city typically vary during the course of the week. We consider that the event happens during the mid-week period, say on Tuesday. Our methodology is independent of a particular day/time, although the traffic situation would vary depending on the day and time of the disaster event. Usually the most critical condition in the traffic network is in the morning peak (herein between 7:30 am to 8:30 am) and evening peak hour (2:00 pm to 3:00 pm). These peaks are for official commuters but the commercial activity in the city centre usually begins after 9:00 am, at this time the peak hour are somehow relieved. We consider that the incident happens at 9:30 am. The event causes the network to be closed in the Zone 1 and some nearby road links. An emergency response system is required, at this stage, to coordinate the city transport, communicate with the city population, and lead people out of the city to a safe location. In this case of the Ramadi city, people will be moved to the two evacuation areas (see Figure 6.2; we have already discussed the justification of the two evacuation areas in Section 6.3.2). The emergency response systems are discussed and evaluated next.

6.4.2 Results and Discussion

We consider and compare two scenarios for emergency response system. Firstly, the traditional emergency response system where people will gain awareness of the disaster situation and response procedure through media such radio, television, telephones (given that such means are still accessible), and through their physical environment (e.g. interacting with the people who are in the nearby area). Secondly, our VANET and Cloud-based intelligent emergency response system, which automatically collects data; intelligently processes the data; and, devises and propagates effective strategies and decisions based on the real-time situation, in line with appropriate policies and procedures already in place in the system. We evaluate the two systems and compare their performance.

6.4.2.1 A Traditional Disaster Response Scenario

A disaster usually causes most people who are in its vicinity to move away from the disaster location. The panic sets off and people start pushing each other without any effective coordination. The situation with vehicles becomes no different, as in the absence of any effective coordination, the roadway sections around a disaster area are blocked, and the incident will spill over like a shockwave over the entire network system. Such a reaction for the Ramadi city caused by a major disaster in the glass factory is depicted in Figure 6.3. Note in the figure that the roads, near the factory, that connect Zone 1 with Zones 3 and 5, and with Evacuation Area 1, all are blocked (depicted by the roads coloured in black). Also note that the roads connecting Zones 2, 3 and 4 with each other, all have very low volume below 500 vehicles per hour (depicted by the roads coloured in red). We further note a couple of roads near Zone 2, nearer the outer boundaries of the city, with volumes between 500 and 1000 vehicles per hour (represented by roads coloured in blue). Furthermore, we note that the road which are located at the outer boundaries of the Ramadi city are coloured in brown and green, depicting higher volumes, between 1000 and 1500 (brown), and greater than 1500 vehicles per hour (green), respectively.

The Vehicle volumes that we have computed using our models amounts to 660 vehicles per hour (400 vehicles in the first 30 minutes) after the glass factory incident for Evacuation Area 1, and 2200 vehicle per hour (1000 vehicle in the first 30 minutes) for Evacuation Area 2. Clearly, there are many more vehicles (almost 4 times) reaching Evacuation Area 2 compared to Evacuation Area 1.

The traffic situation painted in the city network of Figure 6.2 and described in the paragraph above is calculated using the macroscopic model described in Section 6.3.3; it represents a snapshot taken at 10am, i.e. half an hour after the disaster incident has taken place. The 30 minutes period after the incident gives some opportunity to people to start heading towards, and reaching, safe places (such as evacuation areas) outside the boundaries of the city. This period also gives time for transmission of the information so most of the road users know what is happening and where they should be heading.

A final note on the evacuation process: the public transportation vehicles in Ramadi city consists of buses only. The public transportation vehicles will be involved, where possible, in the cases of emergencies, although it will need a decision from the City Council authorities. There is however a plan in place for emergency events in the Ramadi city that 50% of the public vehicles will be involved in transporting people to safe areas. Since public transportation vehicles have a high passenger capacity, these will be useful in the evacuation process. Thirty minutes after the event, when all the drivers in the event area have received the information about the event by VANET, is an appropriate period to make such a decision of incorporating these public vehicles in the evacuation process.

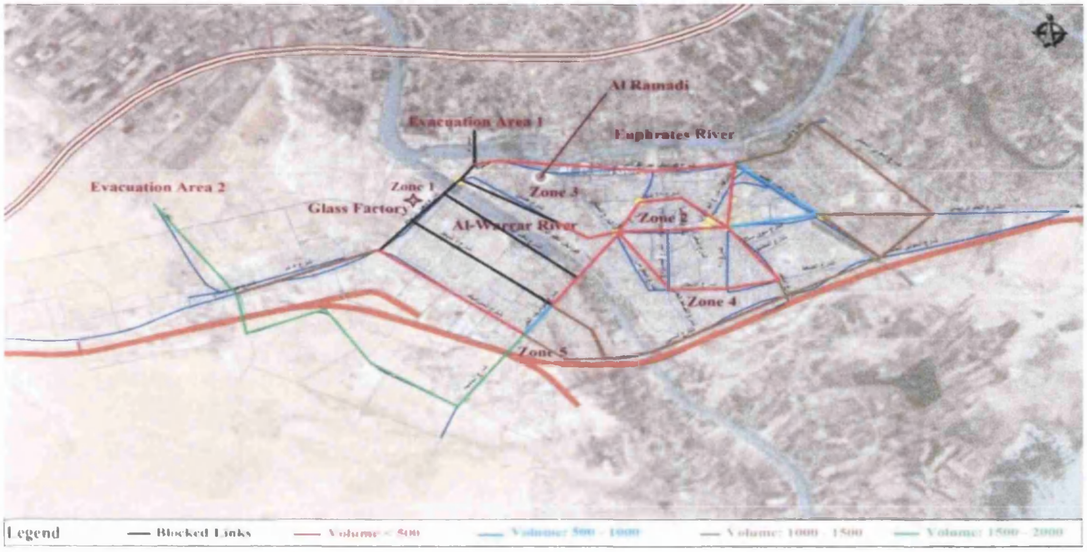


Figure 6.3: transportation network of the Ramadi city with traditional disaster response system.

6.4.2.2 Intelligent VANET Cloud Emergency Response System

We now evaluate our proposed VANETs and Cloud based disaster response system. All the disaster scenario conditions are same as in the previous section including the role of public transportation in the evacuation process. The difference lies in the ability of the system to

1) Acquire real-time data, and establish communication through VANETs, smartphones and social networks,
2) Process the data and devise an optimum strategy by data analysis, and
3) Coordinate and control road traffic and other efforts through dissemination of information and management of the available transport infrastructure (e.g. controlling traffic signals if possible, sending a route map to the traffic navigators and other Global Positioning System (GPS) enabled devices etc).

These three steps are iterative and can provide a periodic update to take any real-time changes into account. For instance, the macroscopic modelling in the Intelligence Layer could be used to periodically compute an O-D matrix depending on the type of disaster and real-time traffic conditions. The O-D matrix then can form the basis of information that is propagated to the transport infrastructure and authorities, individuals and groups. Furthermore, in order to achieve a certain desired traffic control program, certain boundary conditions will be enforced on the city through traffic management systems and authorities, such as

1) There will be no entry in the city,
2) No entry in the area of event,
3) Many routes will be changed into one way flow outside the event area etc.

The road traffic network situation after the disaster hits the Ramadi city is depicted in Figure 6.4, this time though we have exploited our proposed disaster management system to curtail the disaster impact. As in Figure 6.3, the network represents a snapshot taken at 10am, half an hour after the disaster incident has taken place. Take a quick look at the two figures, Figure 6.3 and Figure 6.4, and note the differences – do you see in Figure 6.4 less black and red and more roads in green? Note also that the roads leading to both evacuation areas are now green representing clear roads and high flow (1500-2000 vehicles per hour). Moreover, Note that a greater part of the city centre has roads with free flow (i.e. in green colour) except the roads between Zone 1 and Zone 5 which are coloured in red, representing low flow at less than 500 vehicles per hour. The road next to the glass factory is coloured black and represents a broken link. Also, a few roads near Zone 2, nearer the outer boundaries of the city, with low (red), medium (blue) and medium high (brown) volumes. The low volume (less than 500), we believe, is because of the use of alternative roads available in this case towards Evacuation Area 1.

Based on the computations and our models, 2660 vehicles per hour (1260 vehicles in the first 30 minutes) are being evacuated to Evacuation Area 1, and 2860 vehicle per hour (1530 vehicle in the first 30 minutes) are evacuated to Evacuation Area 2. The evacuation volume per hour is almost similar for both evacuation areas. This is clearly a balanced use of the two evacuation areas, an improvement over the traditional disaster management approach reported earlier where the use of Evacuation Area 1 was significantly smaller.

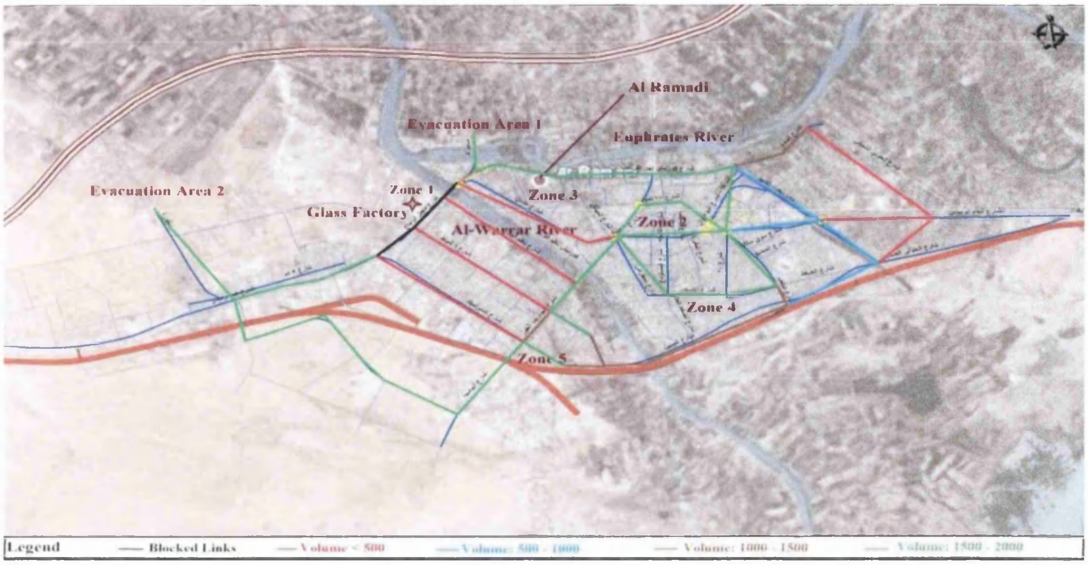


Figure 6.4: Transportation network of the Ramadi city with our disaster management system.

6.4.3 Dynamic VANET based Traffic Sensing and Control

The previous section presented a static approach to the intelligent VANET Cloud based traffic control for evacuation management during major disasters. The approach is static because the Cloud based intelligence layer computes the disaster strategy only once and propagates the devised control plan to the vehicles so that the vehicles could found their routes to the evacuation areas based on their locations. In this section we enhance our earlier approach by extending the earlier static methodology to a dynamic one.

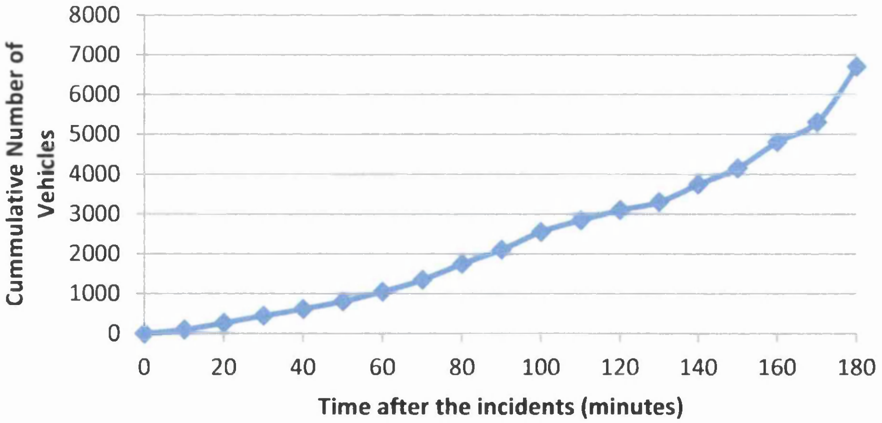


Figure 6.5: The cumulative number of vehicles against time in the east side of the city

The dynamic approach exploits the fact that the road and disaster condition can be sensed periodically, in real time, through vehicular networks and other sources, such as road traditional sensors (inductive loops, traffic counters etc), cameras, social networks, traffic authorities etc (if these are still functional after the disaster). We sense the traffic condition periodically every 10 minutes, giving sufficient time to sensing the traffic, compute a suitable traffic assignment strategy, and propagate the computed strategy through vehicular networks, traffic control signals (if these are functional) and other dissemination and control sources. All the other settings, except dynamic sensing and control, described in the previous section remain the same. This approach allows any transient affects in the city traffic to be taken into account in real-time (every 10 minutes in this case but the time can be decreased or increased to suit the disaster situation) and have a real-time automated control over the evacuation plan. The results for the dynamic modelling approach are presented in Figure 6.5, Figure 6.6, and Figure 6.7.

Figure 6.5 shows the cumulative number of vehicles against time on the Al-Am Street in the east side of the city (See map and streets in Figure 6.4). Figure 6.6 shows the cumulative number of vehicles against on the Ceramic Street in the west side of the city. As mentioned earlier the data is collected, the control strategy is computed and propagated at 10 minutes intervals (see data points in the two figures).

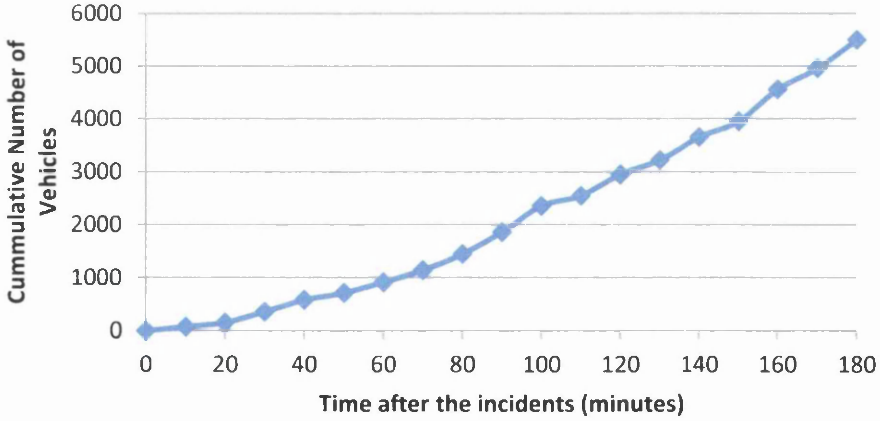


Figure 6.6: The cumulative number of vehicles against time in the west side of the city

Figure 6.7 depicts the time dependent process of evacuation of the residents of the Ramadi city. The figure shows that after 30 minutes of the incident there are only 12% of the people have been evacuated. This is mainly due to social habits and characteristics of the population (an extensive study of the demographics, regional, topographic, environmental and economical characteristics of the Ramadi city, along with its infrastructure and major development issues can be found in [202]). About 50% of people that are in the risk area are indoors. The evacuation of the people indoors mainly begins after 30 minutes from the time of the disaster struck the city. This is done using the private vehicles as well as the public transport as described in Sections 6.4.2.1 and 6.4.2.2 .

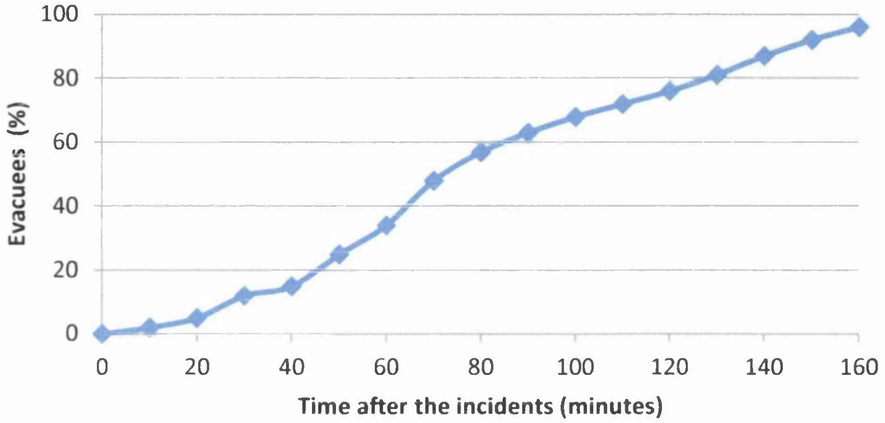


Figure 6.7: The percentage of people evacuated from the city against time

6.4.4 Cloud Computing and Vehicular Networks

As a closing note, we would like to reiterate the advantages here of using cloud computing and vehicular ad hoc networks (VANETs) over the traditional IT and networking systems. Cloud computing (see Chapter 4) is based on virtualisation of resources and is characterised by ubiquitous, convenient, on-demand network access to shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. Moreover, cloud computing also provides utility computing based pay as you go models. Therefore, it provides the right opportunity and models to develop the ICT infrastructure as described in this chapter. Such an infrastructure can be deployed dynamically on a per need basis or permanently. In disaster situations, or in other situations as needed, the software, the data and the computing (software in execution) can be saved (time snapshot) and moved to another (safer) location in short periods of time. Usually, companies and government organisations due to disaster recovery planning compliance require off-site data protection which means that critical data should also be backed up out of the main location. Since September 2001, and many natural disasters in the recent times, organisations are keeping backups of important data in a place increasingly farther away to avoid data losses in case a disaster hits a large geographical area. One would therefore expect that civil emergency planning organisations would also create backups of disaster management data and other related algorithms and software in various geographically spread places. Permanent data items including software etc would require execution in a safe location in case a disaster hits an area and disaster related intelligent decision making is required. However, permanent items would already be available in safe locations due to disaster recovery compliance and legislations. Therefore, only dynamically created items and some real time data requires moving from disaster location to a safe location where the intelligence software can be safely executed. This data would be relatively small compared to the case if all data required moving (including the permanent one). The data can be moved using any of the available networks such as satellite links, cellular networks (e.g. high speed 4G broadband links), metropolitan area networking (MAN) technologies such as WiMAX, mesh networks possibly composed of WiFi LANs, and/or dedicated network links used by the traffic authorities. In the worst case scenario where none of these network links are available, vehicular ad hoc networks and/or other ad hoc networks, such as formed by the public mobile phones and mobile emergency stations can be used to move the computations to a safer location and to propagate back the navigation and evacuation information to the public. Such networks, as the name says are "ad hoc" and "opportunistic" and can be formed on the fly.

6.5 Chapter Summary

The importance of emergency response systems cannot be overemphasised due to the many manmade and natural disasters in the recent years. A greater penetration of ICT in ITS will play a critical role in disaster response and transportation management in order to minimise loss of human life, economic costs and disruptions. In this Chapter, we exploited ITS, C2X, VANETS, mobile and Cloud computing technologies and proposed an intelligent disaster management system. The system architecture and components were described. The system was evaluated using modelling and simulations and its effectiveness was demonstrated in terms of improved disaster evacuation characteristics. Furthermore, the relevant technologies were introduced and explained and a fairly detailed literature review on emergency response systems was provided. The work presented in this chapter has shown that the use of the state of the art ICT technologies enables great advantages in disaster situations. Cloud computing can be used to provide dynamic decision making in transportation and disaster management situations for traffic control and city evacuation purposes, including the possibilities of moving, in quasi-real-time, a virtual computing infrastructure and decision software out of a disaster zone. Moreover, as explained in Section 6.4.4, ad hoc networks (VANETs and/or other ad hoc networks, such as formed by the public mobile phones and mobile emergency stations) can be used to move the computations to a safer location and to propagate back the navigation and evacuation information to the public. Such networks as the name says are “ad hoc” and “opportunistic” and can be formed on the fly.

The contribution of the research presented in this chapter includes a novel Cloud-VANET based distributing computing system architecture, and its associated models, technologies and software for the simulation and evaluation of the disaster management system. In this context, the specific contribution of this thesis is the development of a novel multi-disciplinary cloud computing based system, its architecture and system performance evaluation. Further work on the development and evaluation is in progress. This work is continuing to make impact and has resulted into developing international collaborations and publications.

The future work in this domain will focus on further analysis and validation of the disaster management system, and on broadening the scope of this work to real-time operational and strategic management of transport infrastructure using a range of modelling and control methodologies as mentioned in Section 6.3.3.

Chapter 7: Analysis of Cloud Market Sectors, Applications, and Workload

We have introduced cloud computing, its architecture, deployment and service models in Chapter 4. In Chapter 5, we reviewed the cloud services offered by four major cloud vendors, IBM, Microsoft, Google and Amazon; the specific cloud services offered by Amazon were discussed at length. An application of cloud computing in transportation and digital economy was developed, discussed and analysed in Chapter 6.

We are now ready, in this chapter, to model and analyse Amazon market sectors, applications, and workload. The contribution of this research is in the identification of the major applications and market sectors where cloud computing is being adopted as well as in understanding cloud computing workloads.

The research presented in this Chapter is specific to Amazon. However, Amazon is arguably the top and among the largest cloud computing vendors, and therefore this study is also representative of the cloud computing landscape in general.

The motivation for this study is established through a detailed review of the literature comprising over 200 papers. The literature review sources included conference papers [17], [22], [23], [29], [30], [32], [41], [51], [99], [103], [105], [134], [135], [139], [140], [143], [144], [166], [187], [194], [197], [200], [205], [208]–[249], journal papers [9], [12], [14], [15], [20], [21], [24], [25], [28], [36], [39], [44], [45], [47], [48], [100], [114], [130], [137], [250]–[292], white papers [16], [35], [84], [111], [115], [116], [120]–[124], [132], [138], [142], [151]–[153], [156]–[158], [173], [184], [293]–[304], market research reports [19], [40], [56], [57], [87], [93], [95]–[97], [119], [125], [126], [128], [131], [147], [185], [188], [191], [305]–[327] and books [3], [10], [13], [18], [26], [27], [38], [43], [49], [92], [98], [101], [102], [106], [108]–[110], [112], [117], [136], [141], [145], [146], [148], [159], [186], [198], [201], [328]–[348]. This extensive literature review has helped us to present a comprehensive review (and bibliography) of the cloud computing landscape as presented in Chapter 4 and Chapter 5. A great deal of literature in cloud computing has focussed on its various facets including architecture, infrastructure, deployment and service models, security, management, non-functional requirements and essential requirements. However, we have found no effective research on characterising and modelling its applications, market sectors and workloads with a focus on capacity management. Some cloud vendors do offer capacity management products; however, little information on the design aspects of these products is available in the public space.

Mostly, it is the information on the functionality of the products that you can find. More importantly, no compiled information is available on the applications, market sectors and workload of a major cloud computing provider. Such a study and information can be of great benefit in studying capacity management, risk management and other interesting aspects of this exciting and rapidly evolving field of cloud computing. Furthermore, modelling and analysing such aspects of cloud computing providers is vital because collapse of a big cloud vendor due to its inability to understand the variations in its applications, market sectors and workloads could lead to severe impacts not only on the cloud provider and its customers but also on the national and global economies.

This Chapter is organised into seven sections as follows. Section 7.1 introduces the methodology that has been used in this chapter to model and analyse cloud computing applications and market sectors. Section 7.2 presents the analysis of cloud computing applications and market sectors.

The cloud computing landscape is analysed based on Amazon Products and Services, Amazon IaaS Packages, and Amazon Solutions in Section 7.3, Section 7.4, and Section 7.5, respectively. Section 7.6 presents the cloud computing workload analysis in terms of computations, RAM and secondary storage. Finally Section 7.7 summarises and concludes the chapter.

Table 7-1: List of Amazon's cloud computing customer organisations

Bankinter	DNAnexus	Model Metrics	Trendsmap.com	Junta de Andalucia
Bioproximity	DreamFactory	Monografias.com	TriSys	Kenwood
Cycle Computing	Eagle Genomics	Morph	TweetDeck	LocateTV
Fraunhofer ITWM	Educations.com	Napera	Unfuddle	Marcellus
Harvard Medical School	Egis Technology	Nextpoint	Urbanspoon	MediaPlatform
NASA Jet Propulsion Lab	Encoding.com	Nimbus Health	Urmystar	PBS
Pathwork Diagnostics	Envoy Media Group	nuTsie	uSwitch.com	Pixamba
Scribd	Ericsson	Nealab Technologies	VMLogix	Playfish
Washington Post	Eton Digital	One Hour Translation	VSC Technologies	SmugMug
Yelp!	European Space Agency	OpenCrowd	Xignite	Sonico.com
3scale	Excelsoft Technologies	Papaya Mobile	Wowza	Soundtrckr
6waves 6 Waves Limited	FindTheBest.com	ParkVu	WSO2	Tubaah/NDTV
99designs	FlyCast	Peritor	Zeba Consulting	Twistage
Abaca	Forward3D	photoWALL	Zen Engineering Network Inc.	U.S. Department of State
Active.com	fruux	PicTranslator	37signals	Alexa
Active Interview	ftopia	Ping.sg	Altexa	Hanzo
Actual Analytics	GatherSpace.com	PIXNET	Amazon.com	MiraiBio
AdaptiveBlue	GoAnimate	PostRank	Cloudberry Lab	SearchBlox Software
Advanced Innovations	GoSquared	Praekelt Foundation	ElephantDrive	Acquia
Airbnb	Guardian News & Media	QlikTech	Ipswitch	Amsterdam Museum Night Foundation
Appirio	gumi	Queue-it	Jungle Disk	CITYTECH
Arcus Global	HashCube	Raven	MediaSilo	CloudAngels
Assay Depot	HostedFTP.com	RedBubble	Moonwalk	CozyCot
Autodesk Seek	InstallFree	redBus	Sonian	Digitaria
BackType	InvisibleHand	Ripplex	TeamEXtension	G.ho.st
BigDoor Media	Issuu	Roambi	Vembu	Gumiyo
BrowserMob	Jitscale	rPath	WebServius	iTwin Pte. Ltd.
BuildFax	Kehalim	sambaash	Zmanda	JoomlArt
Cantina consulting	Kingnet Technology	Skifta	AF83	OpenEI.org
Channel Dynamix	Kooaba	Smartsheet	AiCache	RATB
Cirrhus9	kununu.com	Smowtion	EnterpriseDB	ShareThis
Classle	LabSlice	SOASTA	July Systems	Swiftrank
CloudPrime	LIFEPLAT	Sorenson Media	Linden Lab	Zoopla
Cmune	Litmus	StarPound	36Boutiques	Virgin Atlantic Airways
Conduit	Live Talkback	SundaySky	Net Applications	VivaReal
Croop-LaFrance	LiveLeader	Suunto	Zoomii	
CSS Corporation	LiveMocha	swisstopo	Enlighten Designs	
Datapipe	LOUD3R	Tal.ki	fotopedia	
DS3	M-Dot	Tapjoy	GigaVox Media	
DdsWebLink	Mahindra Satyam	Techout.com	Hitachi Systems	
DigitalChalk	MarketSimplified	TellApart	Hungama	
directthought	MileSplit	TicketLea	Indianapolis 500	

Table 7-2 Applications being used in the Amazon cloud

Financial models	service technology	File Transfer Protocol	social network website	technology solutions and services	public health services
statistical language model	social learning platform	access applications remotely	social web engagement analytics	financial services	solutions for entertainment industries
Monte Carlo Simulations	**cloud-based messaging services.**	Online best-price finder	mobile technology solutions	media server software	personal guide to TV viewing
simulations	3D Social Games Publisher	Online document publication	business intelligence software platform	middleware vendor	cloud-based platform
web log analysis	Platform for branded apps	on-demand IT Infrastructure as a Service (IaaS)	Virtual queue SaaS	consulting firm	Webcasting
solutions for industry	Software as a Service	Affiliate Platform	Marketing Tools	network systems design	educational entertainment
genetic testing models	technology solutions	develops games	Art Photography Creative Community	software services	solutions for international stock
process satellite images	application management	services for mobile visual search	travel agency	offsite backup	Social games
drug design process	authentication servers	delivers employment website	infrastructure of internet communication	eCommerce	Photo & Video Sharing
develops high-value diagnostic tests	Web-based application products	Virtual Lab Management solution	software for Mobile Communications	Cloud Storage Tools	Voice and digital entertainment websites
marketing agency	learning system	Social-network	Cloud computing and app management	cloud backup and storage	geosocial Internet radio
social reading site	solutions for healthcare & pharmaceutical	testing service	manages cloud-based communities	Secure File Transfer Software	media house
IT architectures	DNA Sequence Data management	Live participation & voting	Multi Media Streaming	data storage and backup	video platform
education services	web applications for cloud platforms	live chat solution	project management	video workflow management	international online community
Local search and review site	bioinformatics services & software	Language Learning	Ad Network	data management system	commerce data services
API Management Solutions	Learning community	SemContent Services	Cloud-Based Testing	Email hosting	Crowdsourcing Transcription
gaming applications	fingerprint biometrics	Digital Coupon Redemption Processing	Video Delivery Network & Encoding	Java Development and Maintenance	software applications
connect designers	Video transcoding service	automotive components	Open Source Communication Engine	Backup Software	Adaptive Learning Platform
Spam blocking application	marketing firm	SaaS mobile financial applications	automated video solution	API management and monetization	Online Video Curation Platform
online sports community	Telecom services	Network web sites	designs and manufactures precision instruments	backup and recovery software	Photo Sharing and Archiving
video interviewing	web 2.0 digital design agency	Cloud Computing Services & Solutions	Geographical reference data and products.	social networking solution	create high quality datasets
solutions for video content analysis	data provider	community-built content site	free forum product	Application Acceleration	web services
social entertainment network	software platforms & applications	deployment, delivery and management system	platform for mobile developers	Relational Database Management System	IT Consultancy
electronics product management solutions	tool for decisions making	network management services	web-based business transactions and SOAP/REST Web Services	Mobile Internet and Advertising	IT infrastructure migration and integration services
Cloud Computing Solution	Mobile App Development Platform	management solutions for VoIP.	analytics platform	technology platform	digital marketing and technology
software development and services	Digital search marketing agency	Medical Records Software	ticketing platform	e-commerce	free virtual computer
Research Services	system preference panel,	Internet and mobile Radio	mapping of word trends on Twitter	applications for webmasters and eMarketers	mobile commerce
design software & service	File Sharing	develops and manages social Web applications	recruiting software	Online video ecommerce	Remote Access
social media analytics	Cloud Based management solutions	translation and proofreading services	personal dashboard	online bookstore	Open source knowledge sharing platform
Gamification API for site or app	creativity and ideas	custom Rich Internet Applications	hosted software development environment	website design	tools to track Recovery funds
load testing	Real-time analytics	social networking and games for mobile devices	restaurant social network	Wikipedia for photos	sharing network, social influence
Property History Reporting	multimedia business	mobile application	television advertising solutions	podcast and videocast production and publishing	
technology consulting	social gaming	high performance web applications	comparison Website	Mobile Broadcast Solution	

7.1 Methodology

As mentioned in the previous section that the aim of the research presented in this chapter is to model and characterise cloud computing applications, market sectors and workload. The research contributes towards the identification of the major applications and market sectors where cloud computing is being adopted as well as towards understanding cloud computing workloads. Towards this purpose, we focus on Amazon as the cloud provider and build a detailed ontology of its cloud computing space using over two hundred (203, to be specific) case studies. These case studies relate to the 203 customer organisations that purchase from Amazon a mix of its various cloud service offerings. For the list of organisations, see Table 7-1.

The information provided in this chapter is taken from the Amazon website, as on may 2012, [349]. It had been a daunting task to collect information on cloud applications, market sectors and workload. This is due to the fact that the information presented on the Amazon website is textual without clearly quantifying the cloud resources purchased. In most cases information about applications was not available and it had to be extracted from the text through calculations. In certain cases where sufficient information was unavailable, we have tried to contact Amazon and its customer organisations. In most cases, however, we did not get a response and therefore we made assumptions based on the other information available.

The information that we have collected from Amazon revealed that there are around 500 applications being used on Amazon cloud by the 203 organisations. These applications are listed in Table 7-2. An analysis of these 203 organisations and 500 applications revealed that the applications are being used in a total of 14 industry or market sectors. These market sectors are listed below:

1. Social
2. Video
3. Data
4. Marketing
5. Analytics
6. Banking & Finance
7. Hosting

8. Design
9. Mobile
10. Search Engines
11. Learning
12. Games
13. Entertainment
14. Healthcare

Note that this is not a classification of applications, rather of market sectors. An application, such as Monte Carlo Simulations may be used by multiple organisations and market sectors, such as Analytics as well as Banking and Finance. We have used multiple visualisation tools to present information including keywords visualisation tool Tagxedo and Wordle [350], [351], and Microsoft Excel. The keyword visualisation works by displaying each keyword in sizes proportional to its frequency of occurrence. So for example, if a word 'cloud' appears most number of times in a list of words; it will be displayed in the largest font. This technique can help quickly identify major trends in data.

7.2 Applications and Market Sectors

In this section, we analyse cloud computing applications and market sector using keyword visualisations. We first carry out a micro-level analysis of all the applications to identify the trend. This will be followed by the visualisation of applications on market sector level.

Although we have carried out visualisation of all the 14 market sectors, we will only present and discuss two largest market sectors, Social Applications and Video in Section 7.2.1 and Section 7.2.2, respectively.

Figure 7.1 visualises the keywords contained in the set of cloud computing applications that we have compiled using the Amazon case studies. There are a number of keywords which are prominent in the figure including Software, Management, Platform, Services, Technology, Web and Solutions. But these words are common words used in many applications. These perhaps can be ignored. There are other prominent words too in the figure such as Social, Video, Mobile and Data. These words represent the real applications that we wish to identify in the cloud of applications.

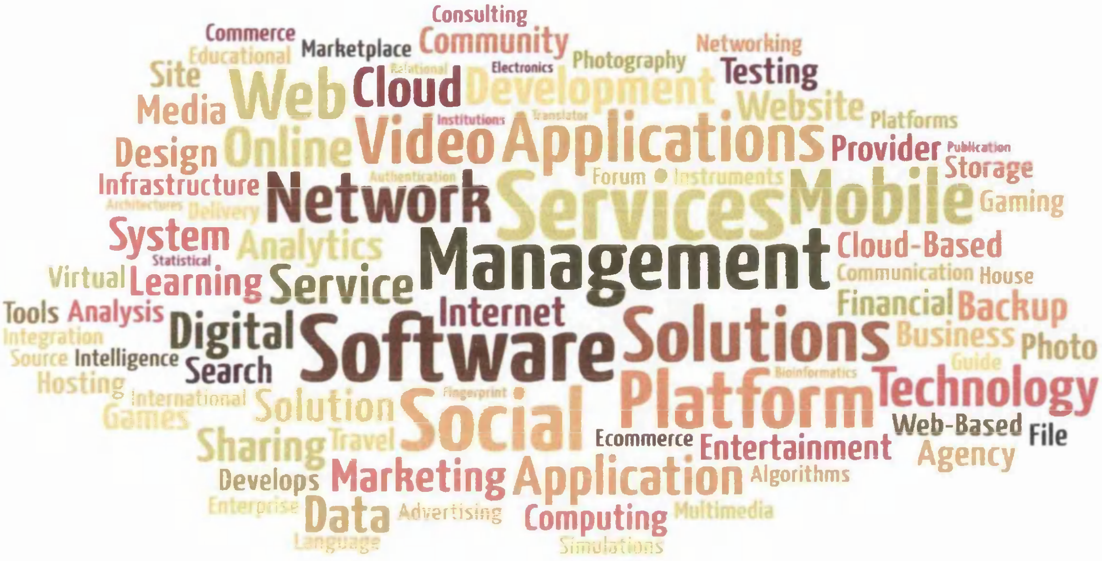


Figure 7.1 Keywords in cloud computing applications

In order to improve our analysis, we modified the application names by removing common keywords from many of the application names. The common keywords included software, platform, and management. The new set of application names are visualised in Figure 7.2. The figure illustrates that social is the most prominent word of all which means that social is using Amazon cloud services across the different cloud market sectors and industries. The second most prominent word in the figure is mobile, and it is followed by video, services, marketing, data, etc.

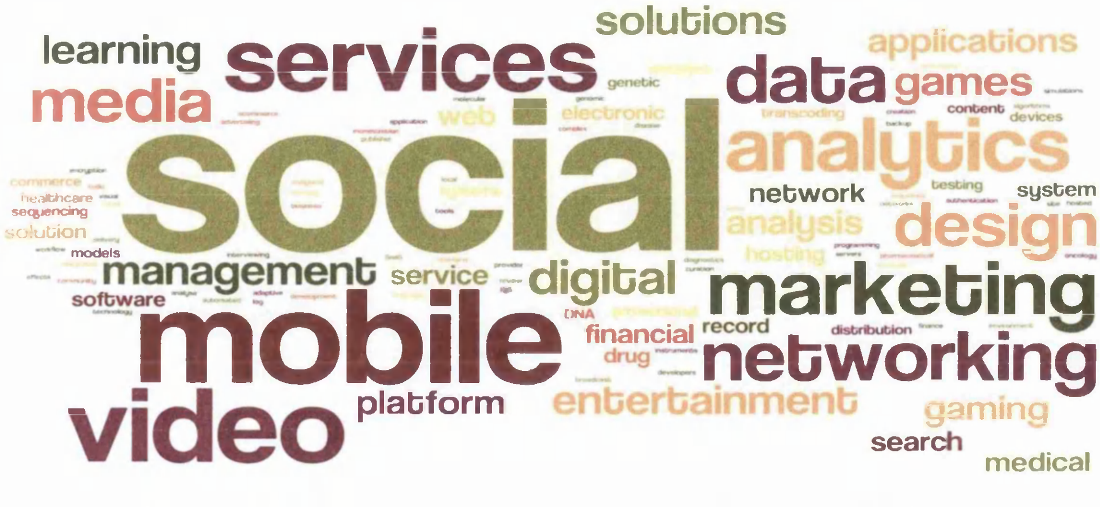


Figure 7.2: Keywords in the modified set of cloud computing applications

In order to improve our analysis, we modified the application names by removing common keywords from many of the application names. The common keywords included software, platform, and management. The new set of application names are visualised in Figure 7.2. The figure illustrates that social is the most prominent word of all which means that social is using Amazon cloud services across the different cloud market sectors and industries. The second most prominent word in the figure is mobile, and it is followed by video, services, marketing, data, etc.

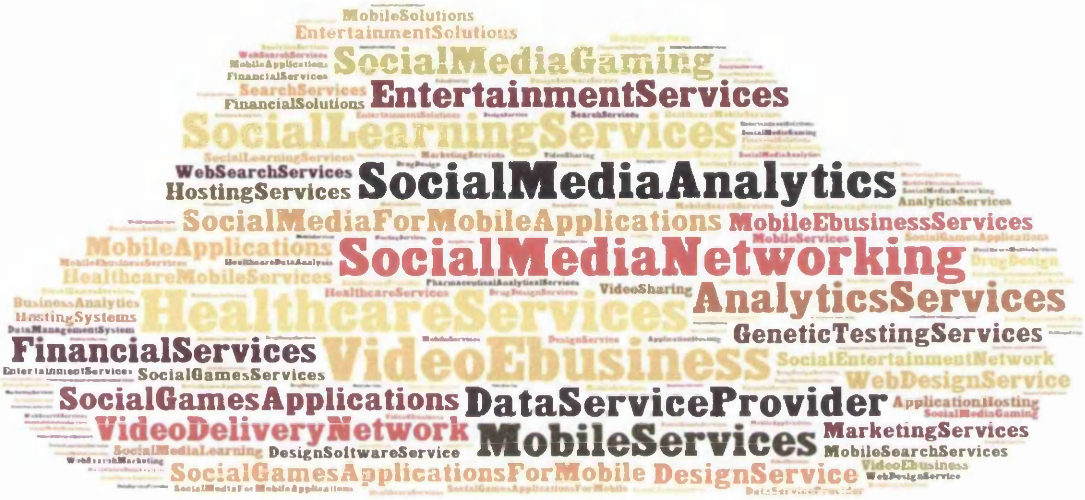


Figure 7.3: Keywords in cloud computing applications (no spaces in application names)

In order to further understand the nature of cloud computing applications, we further modified the keywords in the set of cloud computing applications by removing spaces between each application name. For example, 'Mobile Services' was replaced with 'MobileServices'. Figure 7.3 plots this new set of application keywords. The figure depicts Healthcare, Video, Marketing and Social Media as prominent applications. Social Media although appears the most number of times in the set of applications, its size is reduced due to it appearing in multiple keywords including Social Media Analytics, Social Media Networking, Social Media Learning and so on.

7.2.1 The Social Applications Market Sector

We now look at the Social Applications sector internally without looking at the applications in other sectors. Figure 7.4 plots the keywords in the application names in the Social Applications Sector. We note that cloud computing services have been widely used in (social) networking followed by gaming, analytics, media, and learning, respectively. 19 companies out of 203 have used Amazon Cloud Services for social application sector. Among those 19 companies, 9 of them have used ACS for social networking, 5 of them have used ACS for social gaming, 4 of them have used ACS for social analytics, and only 1 of them have used ACS for social learning.

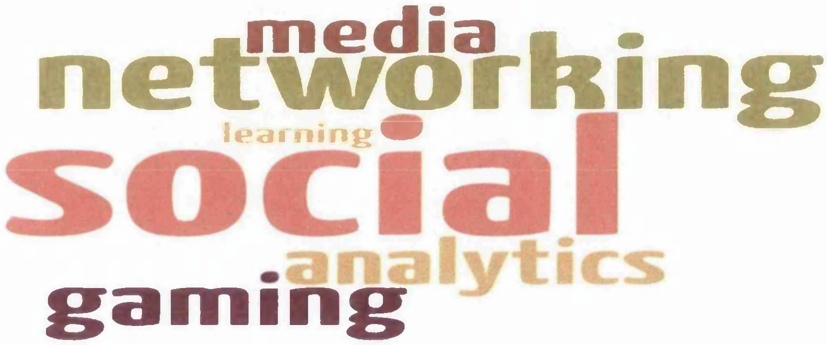


Figure 7.4 Keywords in cloud computing applications within the Social Applications Sector

To further confirm these findings, we visualised the keywords in the application names within the Social Applications sector in Figure 7.5, but this time with no spaces. It is evident in the figure that Social Media Networking is the most used application in the Social Applications market sector followed by Social Media Gaming, Social Media Analytics and Social Media Learning.

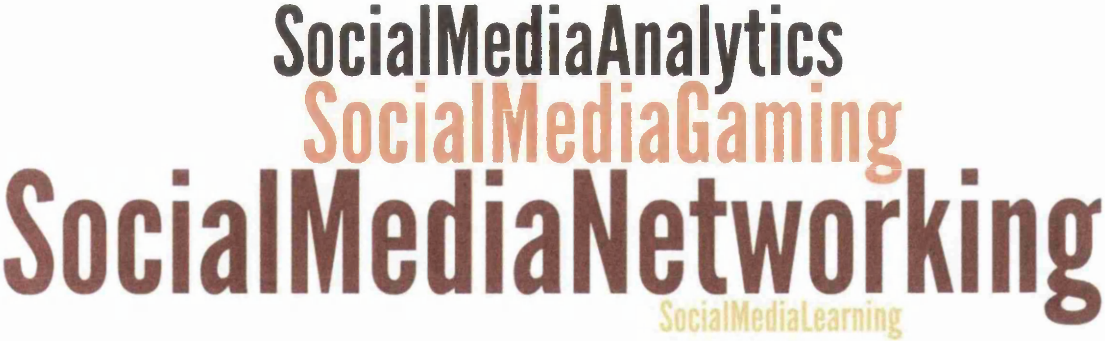


Figure 7.5: Applications within the Social Applications Sector (no spaces)

7.2.2 The Video Applications Market Sector

Subsequent to the Social Applications sector in the previous subsection, we now look at the Video Applications market sector internally without looking at the applications in other sectors. Figure 7.6 plots the keywords in the application names in the Video Applications Sector. We note that within the Video market sector, cloud computing applications are widely used in the (video) network, delivery, solution and ecommerce applications. To further confirm these findings, we visualised the keywords in the application names within the Video Applications sector in Figure 7.7, but this time with no spaces. It is clear in the figure that Video Delivery Network is the most used application in the Video Applications market sector followed by Video Solution applications, Video eCommerce, Video Content Analysis, Video Networking, Video Interviewing, Video Curation and Video Workflow Management.

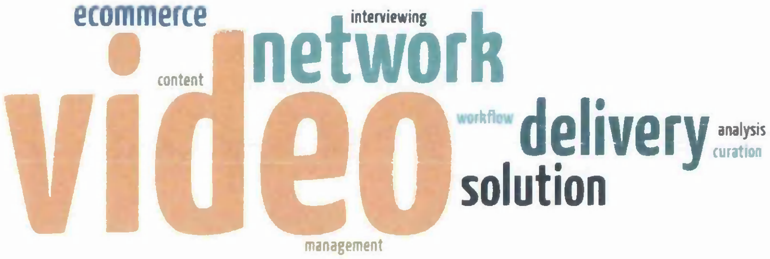


Figure 7.6: Keywords in cloud computing applications within the Video Application Sector

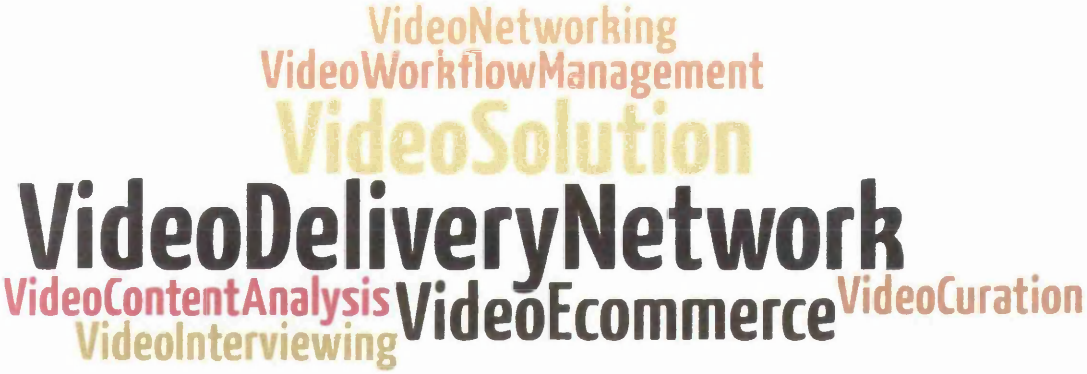


Figure 7.7: Applications within the Video Applications Sector (no spaces)

7.2.3 Summary

This section presented an analysis of cloud computing applications and market sector using keyword visualisations. A micro-level analysis of all the applications was carried out in order to identify the trend for application domains.

We first presented a visualisation of the keywords contained in the set of cloud computing applications that we have compiled using the Amazon case studies. The visualisation identified and highlighted a number of keywords; the most prominent and relevant keywords to this study are Social, Video, Mobile and Data. These words represent the real applications in the cloud of applications. In order to gain further insight into cloud applications, we modified the keywords by removing certain common words from the keywords data. These modifications revealed that "social applications" are the lead applications (in terms of number) that are employing Amazon cloud services across the different cloud market sectors and industries. The social applications are followed by "mobile" applications, "video", "services", "marketing", and "data". Subsequently, further changes were made to gain additional insight into the cloud applications and identify specific application domains which are the major users of cloud computing; these investigations revealed that Healthcare, Video, Marketing and Social Media are the major application customers of cloud computing.

Further to identifying the major applications domains of cloud computing, we selected two (social and video applications) of these domains for further investigation into relevant market sectors. The investigations into the social applications revealed that Social Media Networking is the most used application in the Social Applications market sector followed by Social Media Gaming, Social Media Analytics and Social Media Learning. A similar study into the video applications for cloud revealed that Video Delivery Network is the most used application in the Video Applications market sector followed by Video Solution applications, Video eCommerce, Video Content Analysis, Video Networking, Video Interviewing, Video Curation and Video Workflow Management.

In conclusion, the investigations in this section revealed that Healthcare, Video, Marketing and Social Media are the major application customers of cloud computing. Social media is being employed in a number of sectors where Social Media Networking is the lead market sector within social media applications. Similarly, many Video applications are employing cloud computing, where Video Delivery Network is the lead market sector within the various video applications. These findings agree with the general trend of the increasing use of social media and video applications. In context of cloud computing, it also shows that the use of social media and video applications adapt well with cloud computing, and that cloud computing, due to its flexibility and dynamic business models, may well be among the leading driver of the increasing use of social media and video applications.

7.3 Analysis based on Amazon Products and Services

Amazon AWS has 13 Products and among these Products there are 25 Cloud Services available for customers. Table 7-3 lists the 13 Products and their associated Cloud Services.

Table 7-3: Illustration of the AWS Products and their associated services.

AWS Products	Associated Cloud Services
Compute	EC2, Elastic MapReduce, Auto Scalling
Messaging	SQS, SNS, SES
Storage	S3, EBS, AWS Import/Export
Content Delivery	CloudFront
Monitoring	CloudWatch
Support	AWS Premium Support
Database	SimpleDB, RDS
Networking	Route 53, VPC, ELB
Web Traffic	Alexa Web Information Service, Alexa Top Sites
Deployment & Management	AWS Elastic Beanstalk, AWS CloudFormation
Payments & Billing	FPS, DevPay
Workforce	Mechanical Turk
E-Commerce	FWS

As it can be seen in Figure 7.8 and Figure 7.9 the most used services is EC2, and from the analysis of the case studies, EC2 has been used by 185 companies, making it the most used services of all. EC2 is computing services and this makes it obvious that using Cloud Computing Technology has enabled many organisations to meet their required demands for computing power in a very cost effective and time saving ways due to the fact that Amazon offers its services in ways such as pay per use method which allow organisation to pay for the services only for the time they use them and not stuck with long term contracts. S3 service is the second most used services; it has been used by 164 companies. S3 is storage Cloud Service that Amazon provides to organisation in pay per use method making it very useful for them to use the service and pay for their usage only. EBS has been used by 52 companies making the third most used service. EBS is another storage Service that Amazon provides and it is mainly suited for applications that require a database, file system, or access to raw block level storage. ELB is a networking service, it has been used by 45 companies and it is mainly used to distribute incoming application traffic across multiple EC2 instances automatically.

S3
ELB
simpleDB
SQS
EC2
Mechanical-Turk
EBS
Auto-Scaling
RDS
CloudWatch
Elastic-MapReduce
CloudFront

Figure 7.8: A Visualisation of the AWS Services that have been used by the Companies

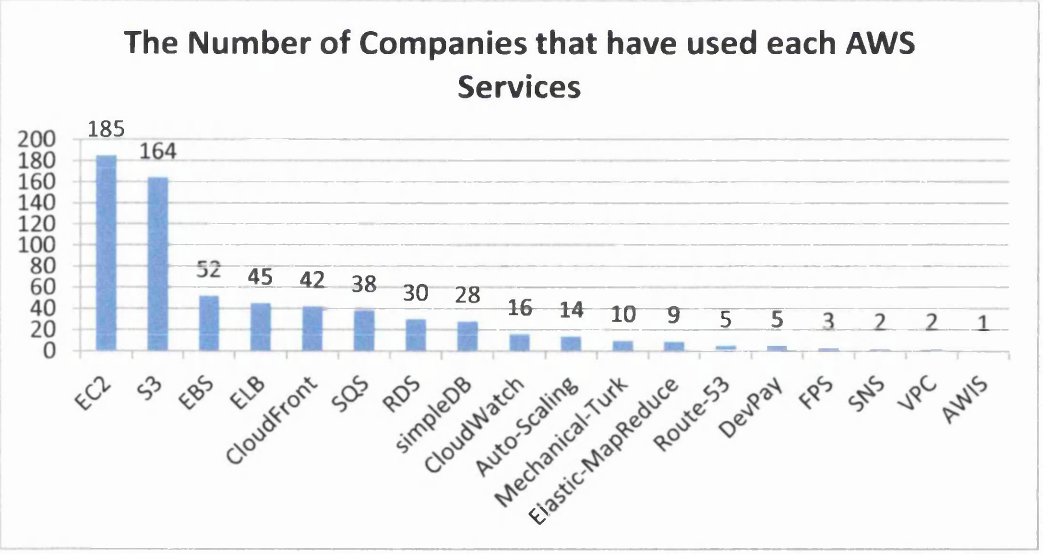


Figure 7.9: A bar chart illustrating the AWS Services that have been used by the Companies.

CloudFront is the fifth most used service as it has been used by 42 companies and it is a web service for content delivery. CloudFront allows businesses to easily distribute content to end users with low latency, high data transfer speeds, and no commitments. SQS is a messaging service and 38 companies have used it. SQS is mainly used by developers to move data between distributed components of their applications which perform different tasks; this is done without losing messages or requiring each component to be always available.

Amazon RDS is a database service; it has been used by 30 companies. Many companies use Amazon RDS due to the fact that it allows them to easily set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing time-consuming database administration tasks. SimpleDB is another database services, 28 companies have used it making it the eighth most used service. It is a flexible and scalable non-relational data store that offloads the work of database administration. CloudWatch is a monitoring service that allows organisations to easily be able to monitor the AWS cloud services and the applications they run on AWS. By using CloudWatch, Developers and system administrators can collect and track metrics, gain insight, and react immediately to keep their applications and businesses running smoothly.

Table 7-4: Illustration of the AWS Services types, Products Types, Number of companies used them and the rate of each used services.

Service Type	Product Type	# of Companies (Case Studies)	Usage Rate
EC2	Compute	185	28.42%
S3	Storage	164	25.19%
EBS	Storage	52	7.99%
ELB	Networking	45	6.91%
CloudFront	Content Delivery	42	6.45%
SQS	Messaging	38	5.84%
RDS	Database	30	4.61%
simpleDB	Database	28	4.30%
CloudWatch	Monitoring	16	2.46%
Auto-Scaling	Compute	14	2.15%
Mechanical-Turk	Workforce	10	1.54%
Elastic-MapReduce	Compute	9	1.38%
Route-53	Networking	5	0.77%
DevPay	Payments & Billing	5	0.77%
FPS	Payments & Billing	3	0.46%
SNS	Messaging	2	0.31%
VPC	Networking	2	0.31%
AWIS	Web Traffic	1	0.15%

Table 7-4 illustrates the types of Amazon's services; the products types, the number of companies used the services and the total rate of the usage. From the table we can see that EC2 which is a compute has been widely used by 185 companies leaving it on the top of the most used Amazon services with a rate of 28.41%. S3 which is a storage service came second on the list of the most used services as it has been used by 164 companies leaving it with a rate of 25.19%. SQS which is a messaging service has been used by 38 companies and it is number six in terms of the most used Amazon service with a rate of 5.84%. Route-53 and DevPay are in place 14 of the most used Amazon services with only 5 companies used them each leaving them with a rate of less than 1%.

7.3.1 Summary

This section presented an analysis of the use of cloud computing services across various companies. We looked into Amazon AWS which provides 25 Cloud Services to customers within its 13 sets of Products. The data about the use of all services by various companies was provided in tabular, graphical and keywords visualisation forms. The investigations revealed that the largest service (in terms of the number of companies) which is being used by the Amazon cloud customers is the EC2 services; 185 companies, out of the total 203 companies, use Amazon EC2. The second largest service is the S3 storage service, being used by 164 companies. The EBS service is the next largest, being used by 52 companies.

7.4 Analysis based on Amazon IaaS Packages

We first present some information about Amazon IaaS Packages and IaaS Service offerings and then use this information to present the analysis of cloud customer landscape. AWS Cloud Services are provided in three different packages and these are defined below.

7.4.1 On-Demand Instances

On-Demand Instances is a payment package that allows clients to pay for the compute capacity they use by the hour with no long-term commitments. This option gives clients the freedom of the costs and complication of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs.

7.4.2 Reserved Instances

Reserved Instances is a payment package that allows clients to pay a low, one-time payment for each instance they want to reserve and in turn receive a significant discount on the hourly usage charge for that instance. After the one-time payment for an instance, that instance is reserved for the clients and there is no further obligation.

7.4.3 Spot Instances

Spot Instances is a payment package that allows clients that enable clients to bid for unused Amazon EC2 capacity. Instances are charged the Spot Price, which is set by Amazon EC2 and fluctuates periodically depending on the supply of and demand for Spot Instance capacity. To use Spot Instances, clients place a Spot Instance request, specifying the instance type, the Availability Zone desired, the number of Spot Instances they want to run, and the maximum price they are willing to pay per instance hour. To determine how that maximum price compares to past Spot Prices, the Spot Price history for the past 90 days is available via the Amazon EC2 API and the AWS Management Console. If the clients' maximum price bid exceeds the current Spot Price, then their request is fulfilled and their instances will run until either they choose to terminate them or the Spot Price increases above their maximum price (whichever is sooner).

7.4.4 Infrastructure as a Service offerings from Amazon

Amazon offers IaaS in one of the six 6 Instance types, these are explained below. This information is used in calculating the organisational spending on cloud services as well for workload modelling.

1. Standard Instances

The Standard Instances are offered in three types and they are defined as:

- ✓ **Small Instance (Default):** 1.7 GB of memory, 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), 160 GB of local instance storage, 32-bit platform

- ✓ **Large Instance:** 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform
- ✓ **Extra Large Instance:** 15 GB of memory, 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform

2. **Micro Instances**
   - ✓ **Micro Instance:** 613 MB of memory, Up to 2 EC2 Compute Units (for short periodic bursts), EBS storage only, 32-bit or 64-bit platform
3. **High-Memory Instances**
   - ✓ **High-Memory Extra Large Instance:** 17.1 GB memory, 6.5 ECU (2 virtual cores with 3.25 EC2 Compute Units each), 420 GB of local instance storage, 64-bit platform
   - ✓ **High-Memory Double Extra Large Instance:** 34.2 GB of memory, 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform
   - ✓ **High-Memory Quadruple Extra Large Instance:** 68.4 GB of memory, 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform
4. **High-CPU Instances**
   - ✓ **High-CPU Medium Instance:** 1.7 GB of memory, 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each), 350 GB of local instance storage, 32-bit platform
   - ✓ **High-CPU Extra Large Instance:** 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform
5. **Cluster Compute Instances**
   - ✓ **Cluster Compute Quadruple Extra Large Instance:** 23 GB of memory, 33.5 EC2 Compute Units, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet
6. **Cluster GPU Instances**
   - ✓ **Cluster GPU Quadruple Extra Large Instance:** 22 GB of memory, 33.5 EC2 Compute Units, 2 x NVIDIA Tesla “Fermi” M2050 GPUs, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet

On-Demand-Instances-S
Reserved-Instances-S
Reserved-Instances-A

Figure 7.10: A Visualisation of the Payment packages that have been used by the Companies.

In Figure 7.10 we can see that Reserved-Instances-A being the most visible keyword due to the fact that it has been used by 115 companies. The second most visible keyword is On-Demand Instances-S as it has been used by 44 companies. Reserved-Instances-S came third is the list of the most visible keywords because it has only been used by 21 companies. Spot Instances-S, Spot Instances-A, and On-Demand Instances-S can hardly been seen because they have been used by low number of companies.

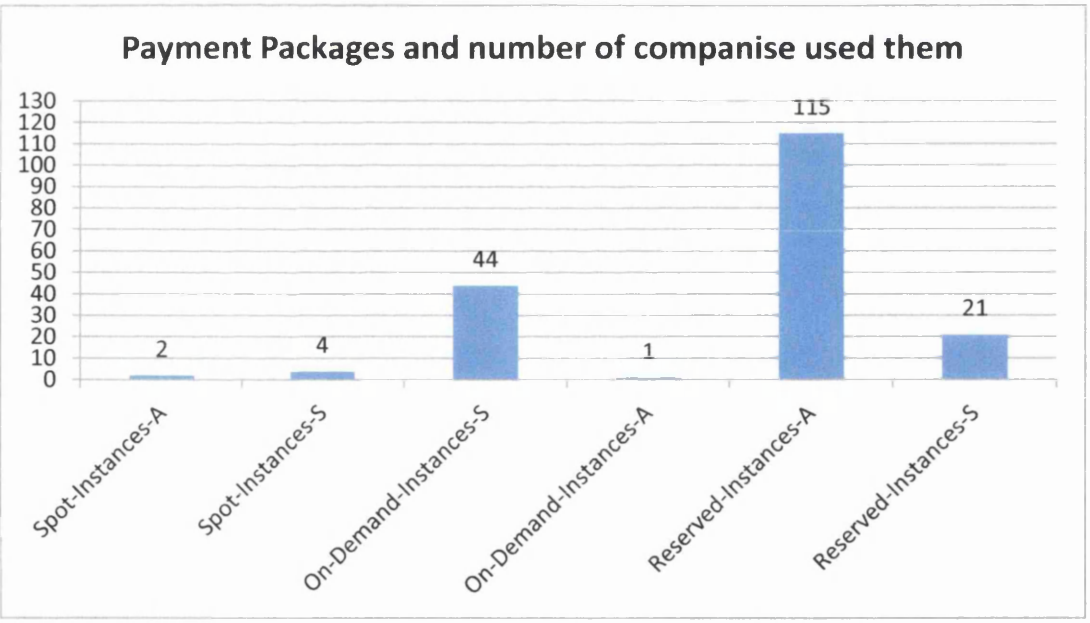


Figure 7.11: A bar chart illustrating the Payment Packages and number of companies used them.

Amazon defines the payment packages in three different ways and they are called; On-Demand Instances, Reserved Instances, and Spot Instances. Figure 7.11 illustrates the number of organisations used each package. Not all of the specific payment packages used by companies were available on Amazon, so we had to come up terms for each package and the terms are A and S where A means assumed because we assumed that

that specific package was used based on the available information, and S means stated and this is based on the actual information obtained from Amazon. From the figure we can note that Reserved Instances-A has been widely used by the largest number of companies as it has been used by 115 companies. It is then followed by On-Demand Instances-S and it has been used by 44 companies. Reserved Instances-S came third as it has been used by 21 companies. The least used packages are Spot-Instances-A and On-Demand- Instances-A where they have been used by only 2 and 1 companies respectively.

7.4.5 Summary

The previous section presented an analysis of the use of cloud computing services offered by Amazon across various customer companies. This section analysed deeper into the cloud customer landscape based on the various Amazon IaaS (Infrastructure as a Service) Packages and IaaS Service offerings. The on-demand, spot, reserved, standard, micro, cluster GPU and other instances offered by Amazon were explained. The use of various types of instances by the set of 203 companies was analysed through keywords visualisation and bar charts. Majority of the companies (115 out of 203) used reserved instances which means that most of the companies reserve instances for lower risks and high predictability at the cost of higher costs. The total number of companies who used reserved instances were 136 (115 + 21); we deduced from the available informed in the 115 cases that the used instances were reserved while in 21 cases the use of reserved instances were explicitly stated. Quite a few companies (44) used on-demand instance. The trend shows that companies are using both on-demand and reserved instances where former allows lower costs at increased risk while the latter allows higher costs with lower business risks.

7.5 Analysis based on Amazon Solutions

Amazon has divided their Cloud Computing Services into 9 Solution areas as below:

- ✓ Application Hosting
- ✓ Backup and Storage
- ✓ Content Delivery
- ✓ E-Commerce

- ✓ High Performance Computing
- ✓ Media Hosting
- ✓ On-Demand Workforce
- ✓ Search Engines
- ✓ Web Hosting

After reviewing the 203 case studies and visualising the solutions that have been used in the case studies, it becomes clear that the most used solution is the Application Hosting (see Figure 7.12, Figure 7.13, and Figure 7.14).

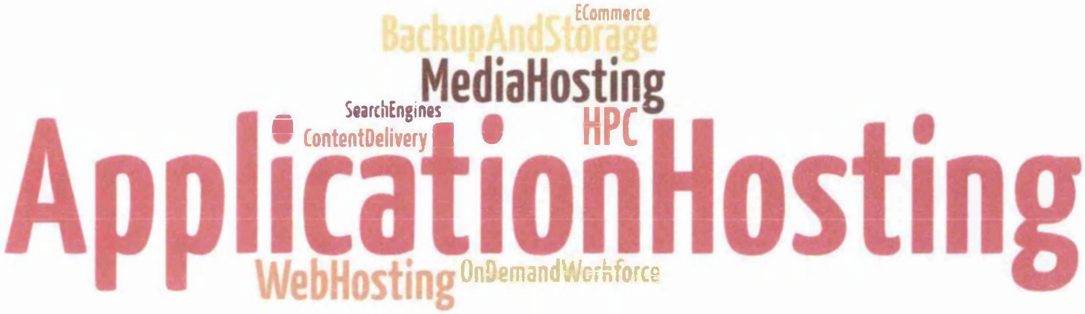


Figure 7.12: A Visualisation of Amazon Classifications.

Out of the 203 companies, 130 used the Amazon Application Hosting services, leaving it to be the most used solution. This indicates that there is a great demand for Cloud Application Hosting Services. Media-Hosting came second in terms of the most used solution, but with only 20 companies using it. Web-Hosting and HPC solutions came third and fourth respectively with 17 companies using Web-Hosting solution and 16 companies employing HPC solutions. Backup Storage solution is the 5th most used solution as it has been used by 14 companies. On-Demand Workforce solution has been used by just 7 companies making number 6 in the most used solution. Content Delivery solution came 7th in the list of the most used solution as it has been used by just 5 companies. Finally E-Commerce and Search Engines came 8th and last on the list of the most used solutions with only 4 companies for each.

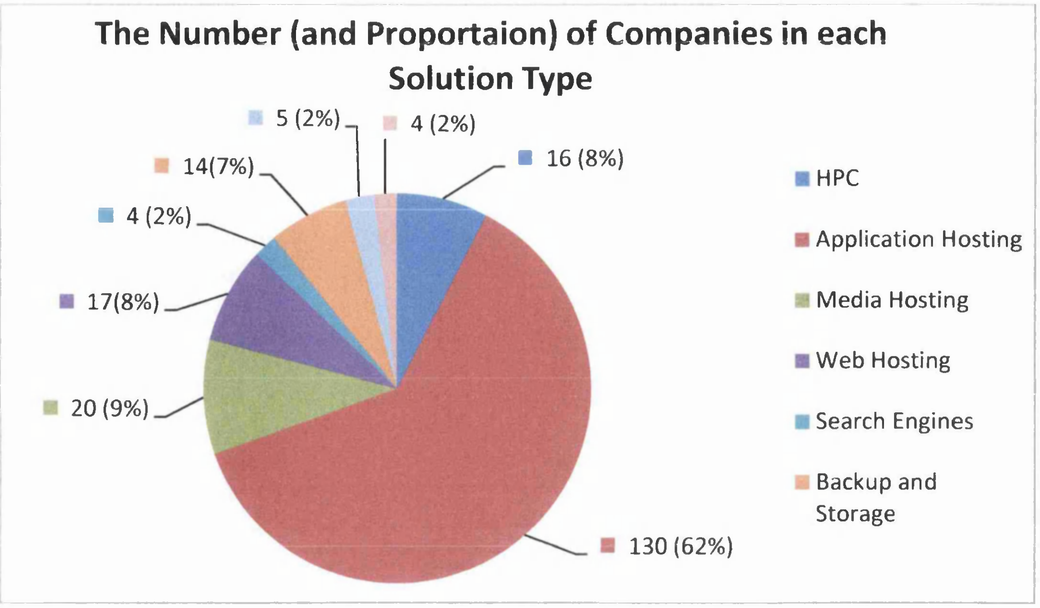


Figure 7.13: A Pie chart illustrating the number of companies that have used each Class.

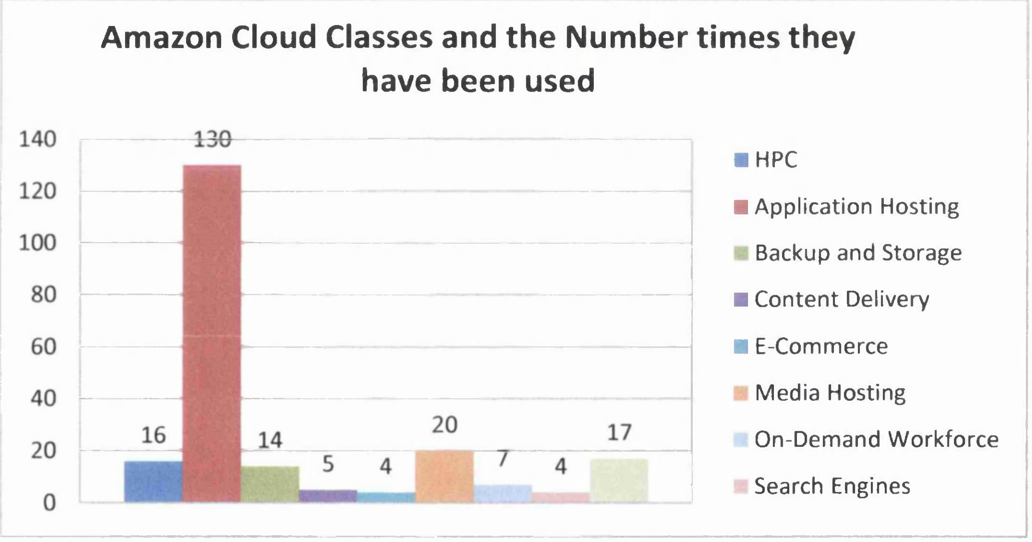


Figure 7.14: A Bar chart illustrating the number of companies that have used each Class

7.5.1 Analysis of Organisational Spending

After analysing the AWS case studies, a calculation of spending on Cloud services and saving was carried out. Analysis of the data was based on total spending and saving on companies in each class type.

HPC companies were the biggest spender as well as saver in millions as their spending was $98579919.19 and their saving $394319676.8. HPC companies spending went for services such as EC2, S3, Elastic MapReduce, EBS, and SQS. Application Hosting companies came second as their spending in millions reached $11556508.56, and their saving reached $46163110.64.

The services used in Application Hosting are EC2, S3, EBS, ELB, CloudFront, SQS, RDS, simpleDB, CloudWatch, Auto-Scaling, Elastic-MapReduce, Route-53, DevPay, SNS, and VPC. Media Hosting came third in terms of the most spending and saving companies, their total spending is $1192475.96, and their total saving is $4769903.84, they used services such as EC2, S3, CloudFront, Elastic MapReduce, SQS, RDS, EBS, and ELB.

Table 7-5: Spending and saving of companies with the associated Solution types.

Solution Type	Number of Companies	Total Spending	Total Saving
HPC	16	$98579919.19	$394319676.8
Application Hosting	130	$11556508.56	$46163110.64
Media Hosting	20	$1192475.96	$4769903.84
Web Hosting	17	$1334667	$5338668.00
Search Engines	4	$155316.5	$621266.00
Backup and Storage	14	$633660.25	$2534641.00
Content Delivery	5	$345207.32	$1380829.28
E-Commerce	4	$137474.16	$549896.64

With 17 companies, Web Hosting Class came forth with total spending of $1334667 and total saving of $5338668.00, many companies in the Web Hosting Class used services such as EC2, S3, RDS, Route 53, SimpleDB, CloudFront, SQS, EBS, Auto Scaling, Cloud Watch, and ELB. Four companies in the Search Engines spent $155316.5 on Cloud Services and saved $621266.00 making Serch Engines Class number five on the list of the most spent and saved Classes by the use of Cloud Services.

Companies in Search Engines Class used Cloud Services such as SimpleDB, EC2, S3, and SQS. With $544544.32 spent and $2178177.28 saved, Backup and Storage companies came sixth and they used services such as EC2, S3, EBS, CloudFront, SQS, RDS, simpleDB, DevPay, and FPS. Companies in Content Delivery class spent $431860.25 and saved $1727441.00 and they used services such as EC2, S3, and CloudFront.

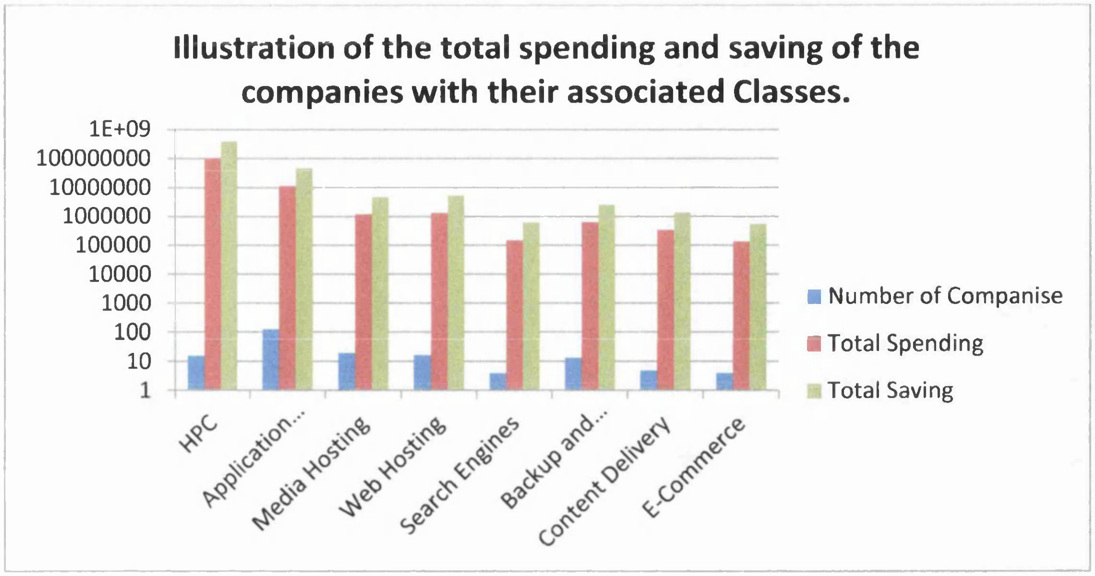


Figure 7.15: A bar chart illustrating the total spending and saving of the companies with their associated Classes.

Figure 7.15 illustrates the total spending and saving of companies based on the Amazon classes. It can be noted that 16 HPC companies has spent more than $98500000 on Amazon services and they saved more than $394300000. Furthermore, 130 Application Hosting companies have spent $11500000 on Amazon services and saved more than $46000000. 17 Web Hosting companies came third in terms of spending and saving on Amazon services. They spent more than $1300000, and saved a total of more than $5300000. The lowest spending and saving companies are the 4 E-Commerce companies where they spent less than $140000 on Amazon services and saved less than $550000.

Table 7-6 illustrates Amazon classes' types and the services that have been used by each class and the rate of the most services used by each class. From the table we can see that Application Hosting is on the top of the list as it has 15 of the Amazing services which are equivalent to 80% of the Amazon Services. Web Hosting came second on the list as it used 11 of the Amazing services which represent 62% of the Amazon Services. Backup & Storage and E-Commerce came third on the list using 9 services each equivalent to 50% of the Amazon Services. Content Delivery came last on the list as it has used only 3 Amazon services which represent only 17% of the Amazon Services.

Table 7-6: Solution types, the services used by each Solution type and the rate of usage

Class Type	Services used	Rate
HPC	EC2, S3, Elastic MapReduce, EBS, SQS	34%
Application Hosting	EC2, S3, EBS, ELB, CloudFront, SQS, RDS, simpleDB, CloudWatch, Auto-Scaling, Elastic-MapReduce, Route-53, DevPay, SNS, VPC	80%
Media Hosting	EC2, S3, CloudFront, Elastic MapReduce, SQS, RDS, EBS, ELB	45%
Web Hosting	EC2, S3, RDS, Route 53, SimpleDB, CloudFront, SQS, EBS, Auto Scaling, Cloud Watch, ELB	62%
Search Engines	SimpleDB, EC2, S3, SQS	23%
Backup and Storage	EC2, S3, EBS, CloudFront, SQS, RDS, simpleDB, DevPay, FPS	50%
Content Delivery	EC2, S3, CloudFront	17%
E-Commerce	EC2,S3, ELB, EBS, SQS, CloudWatch , Auto Scaling, FPS, CloudFront	50%

Figure 7.16 is an illustration of the organisations' spending on the Amazon cloud computing services. Harvard Medical School is most visible keyword in the figure making it the top spender on Amazon cloud services. Other visible keywords are Fraunhofer ITWM the German research institute, the European Space Agency, NASA Jet Propulsion Lab, Ericsson, DNAnexus, Pathwork Diagnostics, and Advanced Innovations. Such visibility means that these organisations have seen the great benefit of cloud computing and as a result they have spent large sums and of course their saving is much higher too.

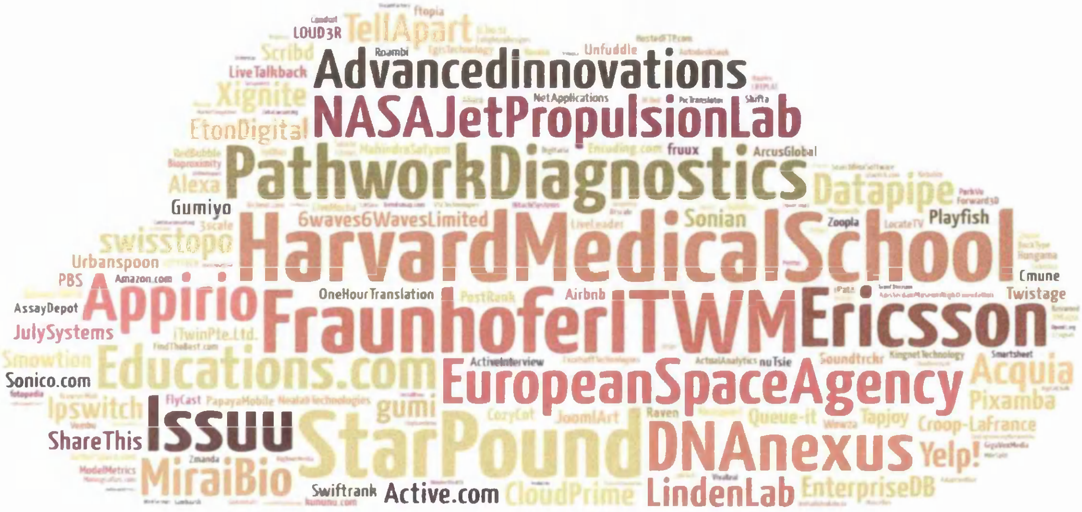


Figure 7.16: Amazon Cloud computing customer organisations in terms of their spendings

7.5.2 Summary

Amazon provides their Cloud Computing Services into 9 Solution areas including Application Hosting, Backup and Storage, Content Delivery, E-Commerce, High Performance Computing, Media Hosting, On-Demand Workforce, Search Engines, and Web Hosting. This section presented an analysis of the 203 company activities across these nine solution types. The various Amazon cloud services used by the set of 203 companies and the organisational spending across these nine solution areas were analysed using tabular, graphical and keywords visualisations.

The analysis revealed that the most widely used solution area by the companies in terms of their number (130 companies out of 203) is the Application Hosting solution by Amazon. The usage of the solutions areas was followed by Media Hosting solutions, Web Hosting, HPC Backup Storage, On-Demand Workforce, Content Delivery, E-Commerce and Search Engine solutions respectively. Subsequently, we analysed the overall spending and respective savings by the companies on cloud computing across the nine solution areas. In spending, HPC was ranked first, i.e. the overall collective spending by the companies on HPC solutions was approximately $100 million, exceeding the spending on any other solutions areas. The second highest collective spending by the companies was on the Application Hosting solutions, followed by Media Hosting, Web Hosting, Search Engines, Backup and Storage, Content Delivery and E-Commerce respectively.

In summary, our analysis revealed that Application Hosting is the most widely used cloud computing solution area, naturally because purchasing and maintenance of business applications is tedious and expensive; cloud computing saves efforts and costs allowing the businesses to focus on their core businesses. Moreover, HPC was found as the second most widely used area and the leading area in cloud solution where companies spent finances. This is due to the fact that HPC resources usually require huge capital investments and maintenance costs. Many companies in the past were unable to use HPC to improve their business processes, product quality and expand their portfolio. Cloud computing has now enabled businesses, particularly small medium enterprises to employ HPC technologies.

7.6 Workload Analysis

7.6.1 Computation

Figure 7.17 illustrates the Organisations' usages of Computation per day. The maximum Computation value that has been used per day was 192000 core-hours as a Computation requirement for Fraunhofer ITWM. Furthermore, the minimum Computation value used per day was less than 50 core-hours and this is the daily requirement of several organisations such as Datapipe (40 core-hours), Ping.sg (48 core-hours), and 36Boutiques (48 core-hours). The figure plotted more than 200 Organisations and due to the fact that the figure is designed using algorithm we are unable to display all of the 200 Organisations.

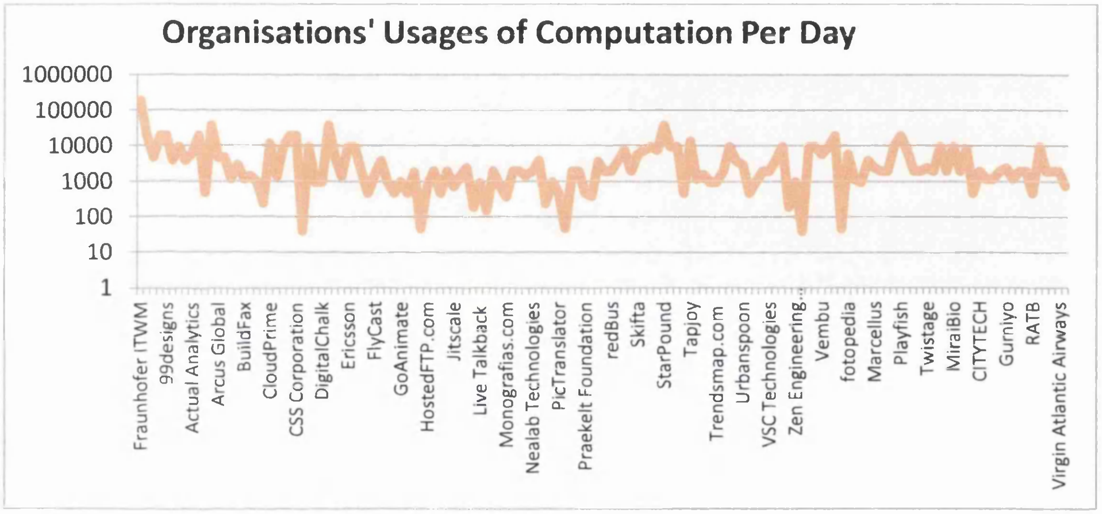


Figure 7.17: Illustration of the Organisations' Usages of Computation per Day (Core-Hours)

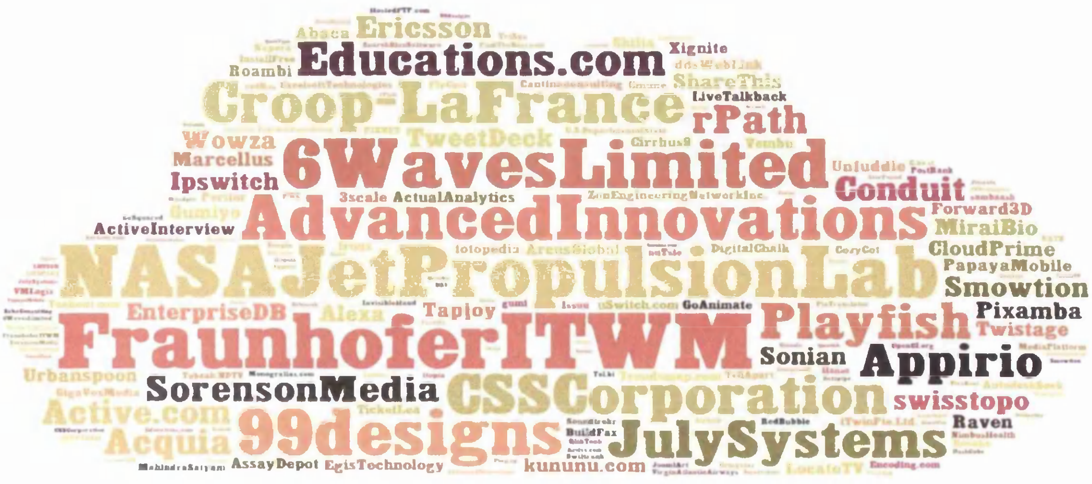


Figure 7.18: Visualisation of Organisations' daily usage of computation.

Figure 7.18 presents a visualisation of the organisations in terms of their daily usage of computation. It can be noted that Fraunhofer ITWM and NASA Jet Propulsion Lab are the most visible keywords and this is due to the fact that their daily usage of computation is greater than others. However, we can also see that organisations such as 6Waved Limited, Advanced Innovations, 99designs, July Systems, and Croop-LaFrance are also very visible as their daily usage of computation is high comparing to organisations such as Forward3D, LocateTV, Build Fax, and Giga You Media which require far less daily computation usage and can hardly been seen in the figure.

7.6.2 Random Access Memory

Figure 7.19 illustrates the Organisations' usages of RAM (GB) per day. The maximum RAM value that has been used per day was 168000 GB (~168 TB) as a requirement for Fraunhofer ITWM. Furthermore, the daily RAM's requirement for NASA Jet Propulsion Lab is less than 66000 GB (~66 TB). The minimum value was 40.8 GB, and this is the daily requirement of Ping.sg. The figure plotted more than 200 Organisations and due to the fact that the figure is designed using algorithm we are unable to display them all.

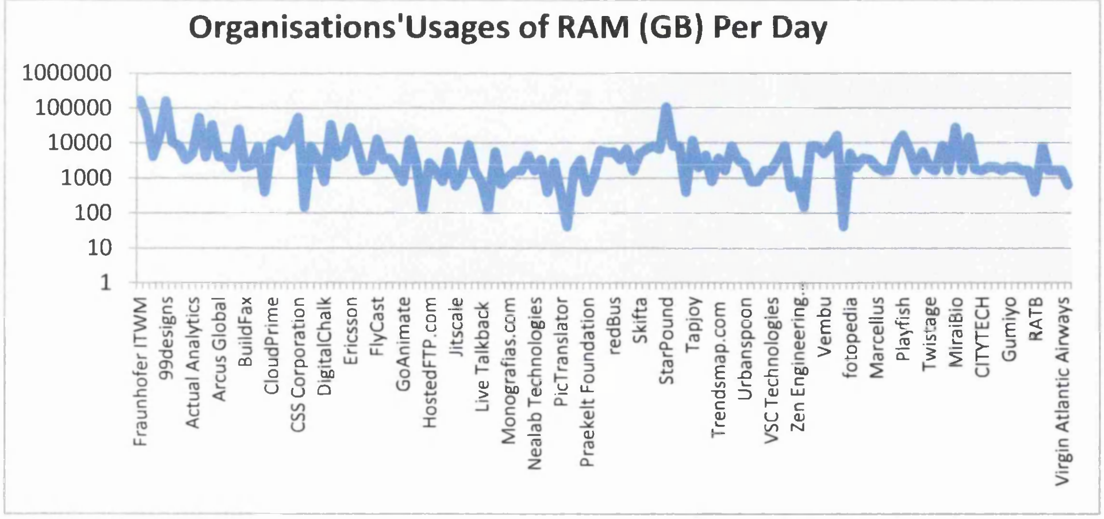


Figure 7.19: Organisations' Usages of RAM (GB) per Day.

Figure 7.20 illustrates a visualisation of the organisations daily usage of RAM in GB. From the figure, keywords such as Fraunhofer ITWM, 99designs, NASA Jet Propulsion Lab, CSS Corporation are the most visible keywords of all and the reason behind that is high daily usage of RAM. Furthermore, Advanced Innovations, Star Pound, 6Waved Limited, July Systems, and Croop-LaFrance are also visible, but not as visible as the organisations mentioned earlier because their daily usages of RAM is less. Additionally, organisations such as Red Bubble, Roambi, gumi, and Peritor are not visible and can hardly be seen because their daily usages of RAM are very low.

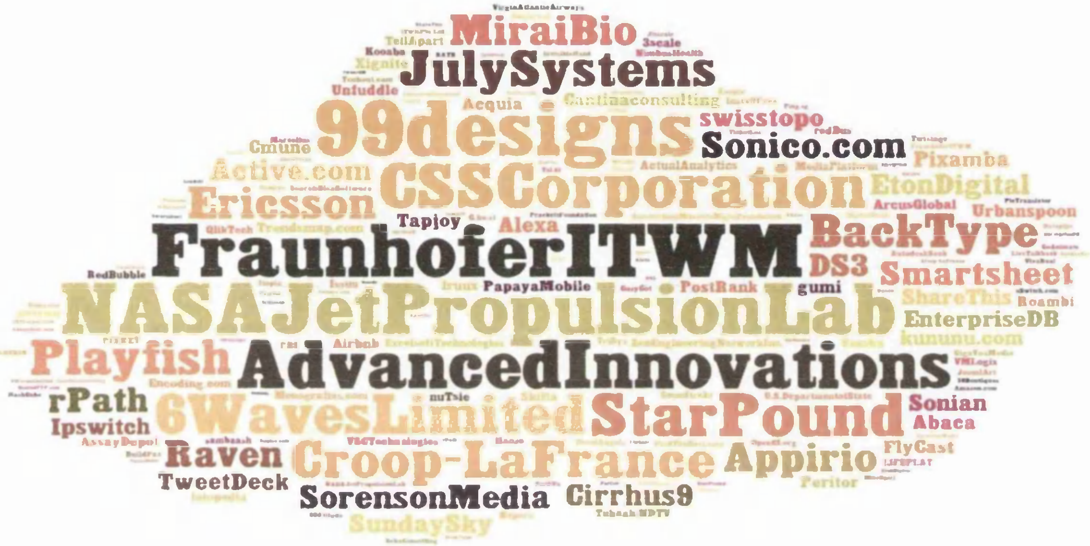


Figure 7.20: Visualisation of Organisations' daily usage of RAM (GB).

7.6.3 Disk Space

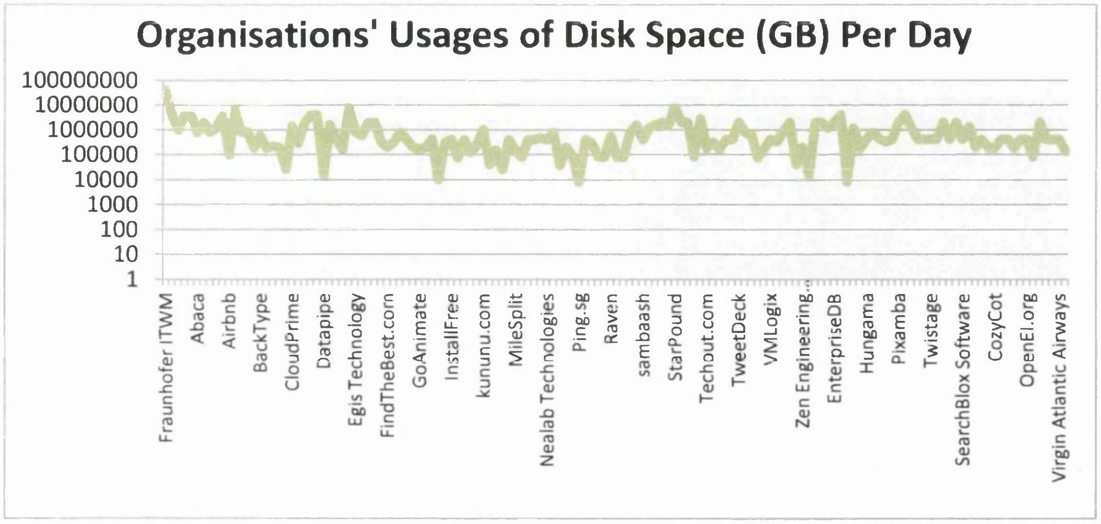


Figure 7.21: Organisations' Usages of Disk Space (GB) per Day

Figure 7.21 shows the Organisations' usages of Disk Space (GB) per day. From the figure it can be noted that the maximum Disk Space value that has been used per day was 40,560,000 GB (40.56 PetaBytes), and that was required by Fraunhofer ITWM. Additionally, the minimum value was 8400 GB, and this is the daily requirement of Ping.sg as well as 36Boutiques. The figure plotted more than 200 Organisations and due to the fact that the figure is designed using algorithm we are unable to display them all.

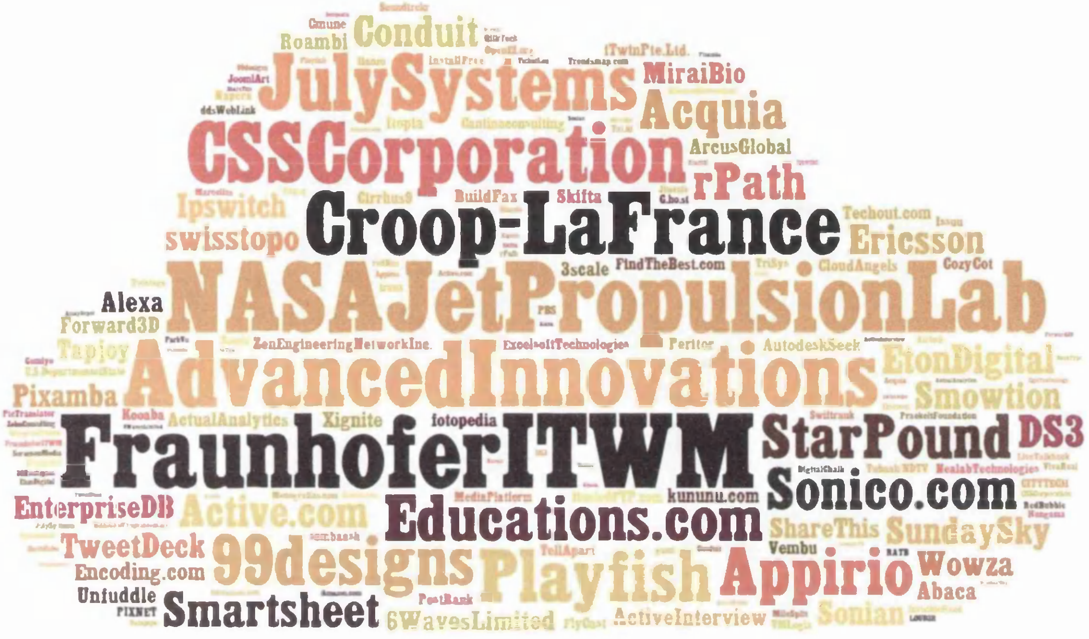


Figure 7.22: Visualisation of Organisations' daily usage of Disk Space (GB).

Figure 7.22 illustrates a visualisation of the organisations daily usage of Disk Space in GB. It can be noted that in the figure there are several visible keywords such as Fraunhofer ITWM, Advanced Innovations, NASA Jet Propulsion Lab, Croop-LaFrance, CSS Corporation, and July Systems. They are the most visible keywords of all due to the fact that they require high daily usage of disk space. Moreover, organisations such as Sonic.com, 99designs, Play Fish, Appirio, Education.com, and Star Pound are visible but their daily requirements of disk space are less than the organisations mentioned earlier and as a result they are not very visible.

7.6.4 Summary

This section presented an analysis of Amazon cloud workload in terms of computations, RAM and disk space usage across the set of 203 companies. We highlighted the major companies which used the highest core-hours, RAM and disk storage. The statistics about the minimum and maximum usage per day was also presented. Most importantly, we computed the overall workload of a leading cloud vendor. This analysis identified that cloud use across IT customers varies greatly and all sizes of companies, from SMEs to leading Universities, government research labs and multinationals are using cloud services.

7.7 Chapter Summary

This chapter presented a study on the analysis of Amazon market sectors, applications, and workload. This study contributed towards the identification of the major applications and market sectors where cloud computing is being adopted as well as in understanding cloud computing workloads.

Section 7.1 introduced the methodology that was used in this chapter to model and analyse cloud computing applications and market sectors. Section 7.2 presented an analysis of cloud computing applications and market sector using keyword visualisations. The investigations revealed that Healthcare, Video, Marketing and Social Media are the major application customers of cloud computing. Social media and video applications are being employed in a number of market sectors; Social Media Networking and Video Delivery Network are the lead market sector in these areas respectively. This insight gained through our analysis agrees with the general trend of the increasing use of social media and video applications. In context of cloud computing, it also shows that the use of social media and video applications adapt well with cloud computing, and that cloud computing, due to its flexibility and dynamic business models, may well be among the leading driver of the increasing use of social media and video applications. Similarly, Healthcare applications generally are increasingly employing ICT to provide services for the digital economy era. The findings of Section 7.2 validate also this general trend.

Section 7.3 presented an analysis of the use of cloud computing services across various companies. The analysis revealed that the largest service used by the Amazon cloud customers is the EC2 compute service followed by S3 storage service and the EBS (Elastic Block Store) service. It confirms the fact that raw computing power, storage capacity and services are the most demanded commodity in ICT.

Section 7.4 analysed further into the cloud customer landscape based on the various Amazon IaaS Packages and Service offerings such as on-demand, spot, reserved, micro, and cluster GPU instances. This analysis revealed that the companies are using both on-demand and reserved instances where former allows lower costs at increased risk while the latter allows higher costs with lower business risks.

Section 7.5 explored the various activities of the set of 203 companies across the nine Solution areas offered by Amazon. These investigations revealed that Application Hosting is the most widely used cloud computing solution area. HPC was found as the second most widely used area and the leading area in cloud solution where companies spent finances. These findings confirm that cloud computing is being dominantly adopted in these areas because it relieves the businesses from tedious efforts, large capital and maintenance costs and enables them to focus on their core business functions.

The analysis presented in Section 7.6 identified that cloud computing use across IT customers varies greatly and all sizes of companies, from SMEs to leading Universities, government research labs and multinationals are using cloud services. Most importantly, we computed and provided the overall workload of a leading cloud vendor.

The research presented in this Chapter is specific to Amazon but is also representative of the cloud computing landscape in general. The motivation for this study was established through a detailed review of the literature comprising over 200 papers. A great deal of literature in cloud computing has focussed on its various facets including architecture, infrastructure, deployment and service models, however, no effective research on analysing and characterising its applications, market sectors and workloads with a focus on capacity management is available. This study is of great benefit in studying capacity management, risk management and other interesting aspects of this exciting and rapidly evolving field of cloud computing, and is vital because collapse of a big cloud vendor could lead to severe impacts on the cloud provider, its customers, as well as on the national and global economies.

Chapter 8: Conclusions and Future Work

The fields of nanotechnologies, electronics, computing and communications have seen unprecedented developments over the past few decades. These developments have transformed the IT systems landscape. The IT systems have evolved from desktop and tightly coupled mainframe computers of the past to modern day highly complex distributed systems. These ICT systems interact with humans at a much advanced level than what was envisaged during the early years of computer development. The ICT systems of today have gone through various phases of developments by absorbing intermediate and modern day concepts such as networked computing, utility, on demand and autonomic computing, virtualisation and so on. We now live in a ubiquitous computing and digital economy era where computing systems have penetrated into the human lives to a degree where these systems are becoming invisible. The price of these developments is in the increased costs, higher risks and higher complexity.

There is a compelling need to study these emerging systems, their applications, and the emerging market sectors that they are penetrating into. Motivated by the challenges and opportunities offered by the modern day ICT technologies, we aimed in this thesis to explore the major technological developments that have happened in the ICT systems during this century with a focus on developing techniques to manage applied ICT systems in digital economy. In the process, we wished to also touch on the evolution of ICT systems and discuss these in context of the state of the art technologies and applications. We identified the two most transformative technologies of this century, grid computing and cloud computing, and two application areas, intelligent healthcare and transportation systems, and contributed in the following areas.

- ✓ A workload model of a grid-based ICT system in the healthcare sector was proposed and analysed was presented in Chapter 3. The work demonstrated the potential of computational grids for its use in healthcare organisations to deploy diverse medical applications. Several organisational and application scenarios for grid deployment in the healthcare sector were considered including four different classes of healthcare applications and 3 different types of healthcare organisations.

✓ In Chapter 6, an intelligent disaster management system for urban environments was proposed by exploiting the advancements in the ICT technologies, including ITS, VANETs, social networks, mobile and Cloud computing technologies. The particular focus of the work was on using distributed computing and telecommunication technologies to improve people and vehicle evacuation from cities in times of disasters. The effectiveness of the proposed intelligent disaster management system was demonstrated through modelling the impact of a disaster on a real city transport environment. The specific contribution of this work was the development of a novel multi-disciplinary cloud computing based system, its architecture and system performance evaluation.

✓ A study on the analysis of Amazon market sectors, applications, and workload was presented in Chapter 7. The contribution of this research is in the identification of the major applications and market sectors where cloud computing is being adopted as well as in understanding cloud computing workloads. This research is specific to Amazon but since Amazon is the top and among the largest cloud computing vendors (see e.g. [4], [5]), this study is also representative of the cloud computing landscape in general. This study is of great benefit in studying capacity management, risk management and other interesting aspects of this exciting and rapidly evolving field of cloud computing. Furthermore, modelling and analysing such aspects of cloud computing providers is vital because collapse of a big cloud vendor due to its inability to understand the variations in its applications, market sectors and workloads could lead to severe impacts not only on the cloud provider and its customers but also on the national and global economies.

✓ Three chapters (Chapter 2, Chapter 4, and Chapter 5) were devoted to explore in detail the landscape of the two technologies (grid and cloud computing) using over 300 sources.

✓ The work contributed in multidisciplinary fields involving healthcare, transportation, mobile computing, vehicular networking, grid, cloud, and distributed computing.

The material on the historical developments, technology and architectural details of grid computing (see Chapter 2) served to understand the reasons grid computing was seen in the past as the global infrastructure of the future. It explains that grid computing pioneered and helped develop the concepts, science and technologies for the fundamental ingredients of cloud computing such as dynamic resource sharing, collaborations and multi-institutional virtual organisations. This material on grid computing (Chapter 2) formed the basis which we subsequently used to explain the cloud computing landscape (see Chapter 4). The background chapters on grid and cloud computing, collectively, provided an insight into the evolution of ICT systems over the last 50+ years, from mainframes to microcomputers, internet, parallel computing, distributed computing, cluster computing, World Wide Web, and computing as a utility and service.

The existing and proposed applications and realisations of grid and cloud computing in healthcare and transport (see Chapter 3, Chapter 6, and Chapter 7) were used to further elaborate the two technologies (i.e. the state of the art in ICT systems) and the ongoing ICT developments in digital economy. The workload model for the grid-based healthcare ICT system (see Chapter 3) provided an insight into the possibilities for resource sharing, collaborations and virtual organisations enabled through grid computing. This workload model can be used by researcher and practitioners for developing shared resources and collaborations, and for resource management of ICT systems. An innovative application of cloud computing for digital economy was proposed in the area of intelligent transportation systems (see Chapter 6); it demonstrated an innovative use of cloud computing to provide dynamic decision making in transportation and disaster management situations for traffic control and city evacuation purposes, including the possibilities of moving, in quasi-real-time, a virtual computing infrastructure and decision software out of a disaster zone. Examples of real cloud services available in the market today were given (see Chapter 5) and subsequently used in Chapter 7 to analyse Amazon market sectors, applications, and workload. As mentioned earlier, this analysis of Amazon cloud space is useful for capacity and risk management of ICT systems.

It was explained that the grid workload modelling study (Chapter 3) is also applicable to cloud computing systems and can be applied to extend the work on cloud workload analysis (Chapter 7). We intended to apply the grid-based healthcare model to the Amazon cloud study of Chapter 7 using the real data collected from Amazon but were unable to due to the time limitations. As mentioned that during the course of this PhD we had initially focussed on grid computing because, by the start of the PhD, the concept of modern day cloud computing was not popular and had not really been taken up by the industry. Grid computing, at that time, was the state of the art for the ICT industry developing technologies for dynamic collaborations, large-scale resource sharing and virtual organisations. By 2010, cloud computing demonstrated high potential to become the future of computing infrastructure and consequently the industry shifted its focus toward cloud computing. Accordingly, we had also shifted the focus of this PhD toward cloud computing technology and applications.

The research developed through this PhD thesis has resulted in four refereed publications: three conference papers, two of these invited, and one book chapter published by Springer. The work that is produced during the PhD, we believe, can produce three or more descent quality journal publications. However, this has not been possible due to the unexpected circumstances as given in Chapter 1; I am hopeful that the research contributed towards this thesis will continue to improve resulting in a number of high quality publications in the near future.

8.1 Future work

We discuss below the directions for future work.

- ✓ Chapter 3 presented our work on a quantitative model to evaluate the suitability of computational grids for pervasive medical applications deployment in healthcare organisations. For a range and mix of medical applications, and three classes of healthcare organisations, we computed steady state probability distributions for the healthcare grid system. This study has quantitatively demonstrated the potential of computational grids for their use in the healthcare area by evaluating their performance for a range of diverse applications and organisations. Future work will focus on improving the models, its analysis and validation by including in the model more detailed parameters and details of Grid-based healthcare systems. We have acquired detailed workload information from the Amazon study presented in Chapter 7. We will find healthcare related workload data from the Amazon study of Chapter 7 and will use it to improve the grid-based healthcare system model.

- ✓ Chapter 6 presented our work on an intelligent disaster management system. Further work on the development and evaluation is in progress. This work is continuing to make impact and has resulted into developing international collaborations and publications. We have acquired more real data from three more cities including two UK cities. The data is being modelled using multiple modelling approaches. We are looking into developing the cloud model further by introducing more detail into it including multiple wireless technologies and social network data. We are also looking into introducing details of middleware in the system architecture. Further work on modelling vehicular network is also in progress. The future work will focus on further analysis and validation of the disaster management system, and on broadening the scope of this work to real-time operational and strategic management of transport infrastructure using a range of modelling and control methodologies as mentioned in Section 6.3.3.

- ✓ In Chapter 7, we presented analysis of the Amazon market sectors, applications, and workload. Future work will include a similar study of one or more other cloud computing vendors in order to make a comparison, and, more importantly, to extend the study to a general case. It is also helpful to understand how cloud computing markets are evolving. Future work will focus on making periodic studies of this kind to understand the evolution of market space, e.g the number and types of cloud applications, the market sectors and the changing computational demands of the ICT industry. The study in Chapter 7 will also be used to model and analyse cloud computing workload for resource and capacity management.

Glossary

Amazon CloudFront: CloudFront is an Amazon feature that is capable of delivery everything from the simplest application to the most complex complete website, one that integrates streaming content with both the static and the dynamic.

Amazon CloudFormation: CloudFormation is design to allow administrators and developers to bring together and coordinate the functioning of a variety of the different type of resources which AWS makes available, all in an organised, predictable manner.

Amazon CloudWatch: Amazon CloudWatch is the monitoring of the applications run by users, along with the resources consumed in such running.

Amazon DevPay: Amazon DevPay service enables mid-users (i.e., developers) to provide cloud- based applications for other end-users by subscription or on-demand without the developer having to create and manage its own billing and account management systems.

Amazon EC2: Amazon EC2 makes available persistent long-term mountable storage options, in addition to providing various distinct kinds of instance with equally varying performance characteristics.

Amazon Elastic Block Store: Amazon Elastic Block Store provides users with block level storage volumes that can be used with Amazon EC2 instances.

Amazon ELB: Amazon ELB is an Amazon feature that connects to various instances, real or virtual Amazon EC2 in order to distribute application traffic automatically as it arrives.

Amazon EMR: Amazon EMR makes data processing on a truly gigantic scale both easy and cost efficient. In order to accomplish this feat, Amazon EMR draws on the infrastructure of both Amazon EC2 and Amazon S3, as hosted in a Hadoop framework.

Amazon Mechanical Turk: Amazon Mechanical Turk (AMT) creates a marketplace that connects businesses and organisations with a workforce of actual people who are available to perform tasks for which human intelligence is needed over computing.

Amazon Route 53: Amazon Route 53 is used to transpose names that humans can deal with, hypothetically for instance www.example.com, into the machine readable versions such as the equivalent 192.0.2.1, which computers need to communicate with one another.

Amazon S3: Amazon S3 is storage service for the Internet and It provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web.

Amazon SimpleDB: Amazon SimpleDB supports users in creating multiple, distinct and varied data sets, store said data set for easy access at a later point in time, query the data from various sets with ease, and retrieve results efficiently.

Amazon SQS: Amazon SQS ensures that data can be moved between the frequently distributed components of an application regardless of how diverse the tasks are, with no loss of messages whether or not the components involved are currently available.

Amazon VPC: Amazon VPC connects users with existing infrastructure of their own to selected (i.e., by the user) AWS offerings in a seamless, secure manner in complete isolation through the use of an Amazon virtual private network (Amazon VPN) connection.

Applications Layer: Applications layer and it is what is visible to the end user since here reside the applications that the user creates, deploys, reconfigures, and operates, depending on the service.

Auto Scaling: Auto Scaling is an Amazon feature that allows users to scale their Amazon EC2 capacity up or down automatically according to conditions they define.

AWS: AWS is considered to be a very powerful and complete cloud services platform. Such platforms enable businesses to have a range of services such as compute power, storage, content delivery, as well as functionalities with the aim of allowing businesses to organise and deploy applications and services that can be achieved at reduced cost with better flexibility, scalability, and reliability.

AWS Direct Connect: AWS Direct Connect allows users to create individual dedicated network connections from their physical locations to the AWS, meaning that they set up completely private connectivity with AWS and whatever portion of the users' own infrastructure, be it their office, their data centre, or any collocation environment over which the user have control.

AWS Import/Export: AWS Import/Export make use of portable storage devises to transport significant quantities of data into and out from its services.

CDS: Clinical Decision Support CDS: is an important part of the knowledge management technology used in Healthcare organisations.

Cloud Auditor: Cloud Auditor is a party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation.

Cloud Broker: Cloud Broker is an entity that manages the use, performance and delivery of cloud services and negotiates relationships between cloud providers and cloud consumers.

Cloud Computing: Cloud computing can be defined as web applications and server services that users pay for in order to access rather than software or hardware that the users buy and install themselves.

Cloud Service Consumer: Cloud Service Consumer is an organisation, a human being or an IT system that consumes service instances delivered by a particular cloud service.

Collective Layer: Collective layer handles resource management on a global or system-wide level, particularly the interaction between distinct resource collections or groupings.

Community Cloud: Community Cloud model is particularly attractive for those groups who want to combine the best of the private model with the public model and its advantages.

Connectivity Layer: The Connectivity layer, which is the heart of the protocols for the communication and authentication, such protocols are very essential for the network transactions which are specified in grid.

EHR: Electronic Health Record EHR: is a record which holds a patient's entire health information, such as medical history, tests results, diagnoses and treatments.

Fabric Layer: the Fabric layer consists of all the resources which are part of the grid and have a physical instantiation.

Globus Toolkit (GT): Globus Toolkit (GT) is an open source technology which is essentially allowing the use of the technology for grid. GT enables users across businesses, organisations, and geographic boundaries to share computing power, databases, and other tools securely online.

GRAM: Grid Resource Access and Management, which is known as (GRAM) is a protocol with functionalities that include the distribution and allocation of computational resources as well as controlling and monitoring such resources.

Grid Computing: grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.

Grid Infrastructure: Grid infrastructure means the grid middleware together with the hardware for which it is designed working in conjunction to create the virtual integrated infrastructure.

Grid Middleware: Grid middleware refers to that software which is designed for the specific purpose of functioning to make it possible to pool and share among diverse types of resources, in the process creating VOs.

GridFTP: GridFTP, and it mainly used as a management tool to manage the data access. In GT, client-side C and Java APIs and SDKs are classified and defined for each one of these protocols.

GRIP: Grid Resource Information Protocol (GRIP), this protocol is primarily use for identifying the resource information protocol standard as well as the model of the associated information.

GTCP: Grid TeleControl Protocol (GTCP)) is a GT4 component that is deployed in managing instrumentation.

Hybrid Cloud: hybrid cloud model features a combination of two or even three of the other models; hence, hybrid clouds can differ markedly from one another in their configuration.

Infrastructure as a Service (IaaS): Infrastructure as a Service (IaaS) means providing virtualised resources on demand, including servers, operating systems, and software stacks.

LotusLive: LotusLive is a platform that provides clients with online services that are delivered using the SaaS model.

MIP: Medical Imaging and Image Processing MIP: Medical imaging is being used by specialists and doctors in healthcare organisations to diagnose and examine patients. Image processing provides doctors with consistent support when it comes to the analysis and treatment of the patients.

OGSA: Open Grid Services Architecture OGSA can be described as distributed communication and computing architecture based in the region of services, with the aim to assure the interoperability on heterogeneous systems so that diverse kinds of resources can communicate and share information effortlessly.

Platform as a service (PaaS): Platform as a service (PaaS) PaaS enables the creation and deployment of applications, using a platform which offers practically unlimited resources, such as processors, memory, and data storage.

Private Cloud: Private cloud exists strictly within a particular organisation, and by definition, access by a user is restricted to that group's designated membership.

Public Cloud: Public cloud is available to basically anyone who is paying to use it, with no stipulations on location, access point, or intended purpose.

Resource Layer: The Resource layer consists primarily of management and information protocols which serve to enable the connectivity layer's communication and security protocols to do their job of securely negotiating, initiating, monitoring, as well as the accounting of and paying for what is employed from among individual resources.

RLS: Service Orchestration is the composition of system components to support the Cloud Providers activities in arrangement, coordination and management of computing resources in order to provide cloud services to Cloud Consumers.

Software as a Service (SaaS): Software as a Service (SaaS) the software being hosted by the service provider while the user connects to the Internet and uses the needed software, as it is, basically, without reconfiguring it or integrating it with other applications or program.

Utility Computing: Utility computing refers to grid computing, along with the applications which it supports, when they are made available on a pay-for-what-you-use basis in one of two ways, first, via an open grid utility service to multiple users or second, in the form of a hosting solution to benefit and individual group, business, organisation or VO.

VANETs: VANETs provide new venues for collecting real-time information from onboard sensors on vehicles and for quick dissemination of information.

Virtual Organisation: Virtual Organisation to mean the group, whether made up of persons, organisations, or any other conceivable collection of entities, which bands together to set the circumstances for and limits on resource sharing in a given network context.

Windows Azure: Windows Azure is a platform that allows users to have the ability with the use of Microsoft data centres to build, host and scale web applications.

WMS: Workspace Management Service (WMS) is a GT4 component which is used for the dynamic allocation of Unix accounts as a simple form of sandbox.

Bibliography

[1] S. Altowaijri, R. Mehmood, and J. Williams, "A Quantitative Model of Grid Systems Performance in Healthcare Organisations," in *ISMS '10 Proceedings of the 2010 International Conference on Intelligent Systems, Modelling and Simulation*, Liverpool, United Kingdom, 2010, pp. 431 –436.

[2] Z. Alazawi, S. Altowaijri, R. Mehmood, and M. B. Abdljabar, "Intelligent Disaster Management System based on Cloud-enabled Vehicular Networks," in *2011 11th International Conference on ITS Telecommunications (ITST)*, St. Petersburg, Russia, 2011, pp. 361 –368.

[3] Z. Alazawi, M. Abdljabar, S. Altowaijri, A. Vegni, and R. Mehmood, "ICDMS: An Intelligent Cloud Based Disaster Management System for Vehicular Networks," in *Communication Technologies for Vehicles*, vol. 7266, A. Vinel, R. Mehmood, M. Berbineau, C. Garcia, C.-M. Huang, and N. Chilamkurti, Eds. Springer Berlin / Heidelberg, 2012, pp. 40–56.

[4] M. Semilof, "Top 10 cloud computing providers of 2011," *Cloud computing information, news and tips - searchCloudComputing.com*, 2011. [Online]. Available: http://searchcloudcomputing.techtarget.com/feature/Top-10-cloud-computing-providers-of-2011#slideshow. [Accessed: 30-Oct-2012].

[5] searchCloudComputing, "Introduction - Top 10 cloud computing providers of 2012," *Cloud computing information, news and tips - searchCloudComputing.com*, 2012. [Online]. Available: http://searchcloudcomputing.techtarget.com/photostory/2240149038/Top-10-cloud-providers-of-2012/1/Introduction. [Accessed: 30-Oct-2012].

[6] Z. Alazawi, S. Altowaijri, M. B. Abdljabar, R. Mehmood, and O. Alani, "An Intelligent Disaster Management System with Cloud Computing and Vehicular Networks," in *3rd CSE Doctoral School Postgraduate Research Conference*, Salford, United Kingdom, 2012.

[7] Globus, "About the Globus Toolkit," 2012. [Online]. Available: http://globus.org/toolkit/about.html. [Accessed: 11-Oct-2012].

[8] B. Tierney, W. Johnston, J. Lee, and M. Thompson, "A data intensive distributed computing architecture for grid applications," *Future Gener. Comput. Syst.*, vol. 16, no. 5, pp. 473–481, 2000.

[9] J. Joseph, M. Ernest, and C. Fellenstein, "Evolution of Grid Computing Architecture and Grid Adoption Models," *IBM Systems Journal*, vol. 43, no. 4, pp. 624–645, Oct. 2004.

[10] I. Foster and C. Kesselman, Eds., *The Grid 2: Blueprint for a New Computing Infrastructure*, 2nd Revised edition. Burlington, Massachusetts, USA: Morgan Kaufmann, 2003.

[11] I. Foster and C. Kesselman, "Computational Grids," in *Vector and Parallel Processing — VECPAR 2000*, vol. 1981, J. Palma, J. Dongarra, and V. Hernández, Eds. Berlin Heidelberg, Germany: Springer-Verlag, 2001, pp. 3–37.

[12] I. Foster, C. Kesselman, and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *Int. J. High Perform. Comput. Appl.*, vol. 15, pp. 200–222, Aug. 2001.

[13] J. Rittinghouse and J. Ransome, *Cloud Computing: Implementation, Management, and Security*, 1st ed. CRC Press, 2009.

[14] C. Jews, R. Ahmad, and D. H. Surman, "IBM Parallel Sysplex clustering: technology options for continuous availability," *IBM Syst. J.*, vol. 47, no. 4, pp. 505–517, Oct. 2008.

[15] I. Foster, "What is the Grid? - a three point checklist," *GRIDtoday*, vol. 1, no. 6, Jul. 2002.

[16] E. Castro-Leon and J. Munter, "Grid Computing Looking Forward." Intel white paper, Intel Solution Services, 2005.

[17] T. Weishaupl, F. Donno, E. Schikuta, H. Stockinger, and H. Wanek, "Business in the Grid: BIG Project," in *Proceedings of the 2nd International Workshop on Grid Economics & Business Models (GECON2005) at GGF13*, Seoul, Korea, 2005.

[18] K. Stanoevska-Slabeva, T. Wozniak, and S. Ristol, *Grid and Cloud Computing: A Business Perspective on Technology and Applications*, 1st ed. Springer, Berlin, 2009.

[19] Insight Research, "Grid Computing: A Vertical Market Perspective 2006-2011," The Insight Research Corp, Mountain Lakes, NJ, USA, Market Analysis and Statistics Report, 2006.

[20] M. A. Rappa, "The utility business model and the future of computing services," *IBM Systems Journal*, vol. 43, no. 1, pp. 32–42, Jan. 2004.

[21] R. Buyya, D. Abramson, and S. Venugopal, "The Grid Economy," *Proceedings of the IEEE, Grid Computing*, vol. 93, no. 3, pp. 698 –714, Mar. 2005.

[22] A. Vahdat, T. Anderson, M. Dahlin, E. Belani, D. Culler, P. Eastham, and C. Yoshikawa, "WebOS: Operating System Services for Wide Area Applications," in *HPDC '98 Proceedings of the 7th IEEE International Symposium on High Performance Distributed Computing*, Chicago, Illinois, USA, 1998, pp. 1–16.

[23] C. Catlett, "The Philosophy of TeraGrid: Building an Open, Extensible, Distributed TeraScale Facility," in *CCGRID '02 Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*, Berlin, Germany, 2002, p. 8.

[24] F. Gagliardi, B. Jones, F. Grey, M.-E. Begin, and M. Heikkurinen, "Building an infrastructure for scientific Grid computing: status and goals of the EGEE project," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 363, no. 1833, pp. 1729–1742, 2005.

[25] I. Foster, "Globus Toolkit Version 4: Software for Service-Oriented Systems," *Journal of Computer Science and Technology*, vol. 21, no. 4, pp. 513–520, 2006.

[26] R. Buyya, J. Broberg, and A. M. Goscinski, *Cloud Computing: Principles and Paradigms*, 1st ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.,, 2011.

[27] R. Buyya and S. Venugopal, *Market-Oriented Computing and Global Grids: An Introduction, in Market Oriented Grid and Utility Computing*, 1st ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2009.

[28] K. Keahey, I. Foster, T. Freeman, and X. Zhang, "Virtual workspaces: Achieving quality of service and quality of life in the Grid," *Scientific Programming*, vol. 13, no. 4, pp. 265–275, Oct. 2005.

[29] I. Foster, "Globus toolkit version 4: software for service-oriented systems," in *NPC'05 Proceedings of the 2005 IFIP international conference on Network and Parallel Computing*, Beijing, China, 2005, pp. 2–13.

[30] I. Foster and C. Kesselman, "The Globus project: a status report," in *Proceedings of the 7th Heterogeneous Computing Workshop (Hcw'98)*, Orlando, Florida, U.S.A, 1998, pp. 4–18.

[31] R. Butler, D. Engert, I. Foster, C. Kesselman, S. Tuecke, J. Volmer, and V. Welch, “Design and Deployment of a National-Scale Authentication Infrastructure,” *Network, IEEE*, vol. 33, no. 12, pp. 60–66, Dec. 2000.

[32] I. Foster, C. Kesselman, G. Tsudik, and S. Tuecke, “A security architecture for computational grids,” in *CCS ’98 Proceedings of the 5th ACM conference on Computer and communications security*, San Francisco, California, United States, 1998, pp. 83–92.

[33] M. Gasser and E. McDermott, “An Architecture for Practical Delegation in a Distributed System,” in *Proceedings of the 1990 IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, California, USA, 1990, pp. 20–30.

[34] J. Howell and D. Kotz, “End-to-end authorization,” in *OSDI’00 Proceedings of the 4th conference on Symposium on Operating System Design & Implementation - Volume 4*, San Diego, California, USA, 2000, vol. 4, pp. 1–14.

[35] V. Berstis, “Fundamentals of Grid Computing.” IBM RedBooks Paper. IBM Corporation, 2002.

[36] D. S. Meliksetian, J.-P. Prost, A. S. Bahl, I. Boutboul, D. P. Currier, S. Fibra, J.-Y. Girard, K. M. Kassab, J.-L. Lepesant, C. Malone, and P. Manesco, “Design and Implementation of an Enterprise Grid,” *IBM Systems Journal*, vol. 43, no. 4, pp. 646 –664, 2004.

[37] M. Smith, T. Friese, and B. Freisleben, “Model Driven Development of Service-Oriented Grid Applications,” in *AICT-ICIW ’06 Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services*, Guadeloupe, French Caribbean, 2006, pp. 139–145.

[38] L. Ferreira, V. Berstis, J. Armstrong, M. Kendzierski, A. Neukoetter, M. Takagi, R. Bing-Wo, A. Amir, R. Murakawa, O. Hernandez, J. Magowan, and N. Bieberstein, *Introduction to Grid Computing with Globus*. North Castle, New York, USA: IBM Redbooks publication, IBM Press, 2003.

[39] I. Foster and S. Tuecke, “Describing the Elephant: The Different Faces of IT as Service,” *ACM Queue, New York, NY, USA*, vol. 3, no. 6, pp. 27–34, 2005.

[40] B. Aiken, J. Strassner, B. Carpenter, I. Foster, C. Lynch, J. Mambretti, R. Moore, and B. Teitelbaum, “Network Policy and Services: A Report of a Workshop on Middleware,” IETF, Fremont, CA, USA, Technical Report 2768, 2000.

[41] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster, “The Globus Striped GridFTP Framework and Server,” in *SC ’05 Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, Seattle, WA, USA, 2005, pp. 1–11.

[42] A. Rodriguez, D. Sulakhe, E. Marland, V. Nefedova, N. Maltsev, M. Wilde, and I. Foster, “A Grid-Enabled Service for High-Throughput Genome Analysis,” in *Proceedings of the 10th Global Grid Forum, Workshop on Case Studies on Grid Applications*, Berlin, Germany, 2004, pp. 1–11.

[43] R. Wolski, J. Brevik, J. S. Plank, and T. Bryan, “Grid Resource Allocation and Control Using Computational Economies,” in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman and G. Fox, Eds. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2003, pp. 747–771.

[44] Z. Li, C. Cheng, and F. Huang, “Utility-driven solution for optimal resource allocation in computational grid,” *Comput. Lang. Syst. Struct.*, vol. 35, no. 4, pp. 406–421, 2009.

[45] B. Allcock, A. Chervenak, I. Foster, C. Kesselman, and M. Livny, "Data Grid tools: Enabling Science on Big Distributed Data," *Journal of Physics: Conference Series*, vol. 16, no. 1, pp. 571–575, Jan. 2005.

[46] D. Bernholdt, S. Bharathi, D. Brown, K. Chanchio, M. Chen, A. Chervenak, L. Cinquini, B. Drach, I. Foster, P. Fox, J. Garcia, C. Kesselman, R. Markel, V. Nefedova, L. Pouchard, A. Shoshani, A. Sim, G. Str, D. B. K. Chanchio, S. Bharathi, A. Chervenak, C. K. A. Usc, D. Brown, L. Cinquini, P. Fox, J. Garcia, D. Middleton, G. Str, B. Drach, D. Williams, and L. L. National, "The Earth System Grid: Supporting the Next Generation of Climate Modeling Research," *Proceedings of the IEEE, Grid Computing*, vol. 93, no. 3, pp. 485–495, Mar. 2005.

[47] E. Cody, R. Sharman, R. H. Rao, and S. Upadhyaya, "Security in Grid Computing: A Review and Synthesis," *Decision Support Systems*, vol. 44, no. 4, pp. 749–764, Mar. 2008.

[48] A. Chakrabarti, A. Damodaran, and S. Sengupta, "Grid Computing Security: A Taxonomy," *IEEE Security and Privacy*, vol. 6, no. 1, pp. 44–51, Jan. 2008.

[49] H. Jin, D. Reed, and W. Jiang, Eds., *Network and Parallel Computing*, vol. 3779. Berlin Heidelberg, Germany: Springer, 2005.

[50] I. Foster, K. Czajkowski, D. Ferguson, J. Frey, S. Graham, T. Maguire, D. Snelling, and S. Tuecke, "Modeling and Managing State in Distributed. Systems: The Role of OGSI and WSRF," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 604–612, 2005.

[51] W. Benger, I. Foster, J. Novotny, E. Seidel, J. Shalf, W. Smith, and P. Walker, "Numerical Relativity in a Distributed Environment," in *Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing*, San Antonio, Texas, USA, 1999, pp. 1–11.

[52] U. Schwiegelshohn, R. M. Badia, M. Bubak, M. Danelutto, S. Dustdar, F. Gagliardi, A. Geiger, L. Hluchy, D. Kranzlmüller, E. Laure, T. Priol, A. Reinefeld, M. Resch, A. Reuter, O. Rienhoff, T. Rüter, P. Sloot, D. Talia, K. Ullmann, R. Yahyapour, and G. von Voigt, "Perspectives on grid computing," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1104–1115, Oct. 2010.

[53] C. Mitton and C. Donaldson, "Health care priority setting: principles, practice and challenges," *Cost Effectiveness and Resource Allocation*, vol. 2, no. 1, pp. 1–8, 2004.

[54] C. P. Roth, Y.-W. Lim, J. M. Pevnick, S. M. Asch, and E. A. McGlynn, "The Challenge of Measuring Quality of Care From the Electronic Health Record," *American Journal of Medical Quality*, vol. 24, no. 5, pp. 385–394, Sep. 2009.

[55] A. Majeed, J. Car, and A. Sheikh, "Accuracy and completeness of electronic patient records in primary care," *Fam. Pract.*, vol. 25, no. 4, pp. 213–214, Aug. 2008.

[56] H. I. U., Clinical Standards Department, Royal College of Physicians HIU, "A clinician's guide to record standards – Part 1: why standardise the structure and content of medical records?," Academy of Medical Royal Colleges, Academy of Medical Royal Colleges, London, United Kingdom, Service Guidance 4275b, 2008.

[57] H. I. U., Clinical Standards Department, Royal College of Physicians HIU, "A Clinician's Guide to Record Standards– Part 2: Standards for the structure and content of medical records and communications when patients are admitted to hospital," AoMRC, London, United Kingdom, Clinician Guidance Report 4275a, 2008.

[58] *Keynote (IHS): Healthcare Information Systems - Requirements and Vision.* .

[59] P. Donachy, T. Harmer, and R. Perrott, "GeneGrid - a Grid Based Virtual Bioinformatics Laboratory," presented at the UK eScience All Hands Meeting, Environmental eScience Applications, 2003,111-116., East Midlands Conference Centre Nottingham, 2003, pp. 111–116.

[60] B. J. Liu, M. Z. Zhou, and J. Documet, "Utilizing Data Grid Architecture for the Backup and Recovery of Clinical Image Data," *Comput Med Imaging Graph*, vol. 29, no. 2–3, pp. 95–102, 2005.

[61] A. Joch, "Grid gets down to business: Early enterprise adopters of grid computing praise benefits such as ultra-speedy processing for heavy-duty applications.," *Network World*, 2004. [Online]. Available: http://www.networkworld.com/power/2004/122704techgrid.html. [Accessed: 11-May-2013].

[62] D. Goldstein and P. Groen, "Grid Computing, Health Grids, and EHR Systems, Virtual Medical Worlds," 2006. [Online]. Available: http://www.hoise.com/vmw/07/articles/vmw/LV-VM-01-07-36.html. [Accessed: 24-May-2013].

[63] V. Koufi, F. Malamateniou, and G. Vassilacopoulos, "A Mediation Framework for the Implementation of Context-Aware Access Control in Pervasive Grid-Based Healthcare Systems," in *Advances in Grid and Pervasive Computing*, N. Abdennadher and D. Petcu, Eds. Geneva, Switzerland: Springer Berlin Heidelberg, Germany, 2009, pp. 281–292.

[64] H. M. Phung, D. B. Hoang, and E. Lawrence, "A Novel Collaborative Grid Framework for Distributed Healthcare," in *9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009. CCGRID '09*, Shanghai, China, 2009, pp. 514–519.

[65] T. Savel, K. Hall, B. Lee, V. McMullin, M. Miles, J. Stinn, P. White, D. Washington, T. Boyd, and L. Lenert, "A Public Health Grid (PHGrid): Architecture and value proposition for 21st century public health," *International Journal of Medical Informatics*, vol. 79, no. 7, pp. 523–529, Jul. 2010.

[66] H. Song, J. J. Dong, C. Han, W. Jung, and C.-H. Youn, "A SLA-Adaptive Workflow Integrated Grid Resource Management System for Collaborative Healthcare Services," in *Third International Conference on Internet and Web Applications and Services, 2008. ICIW '08*, Athens, Greece, 2008, pp. 702–707.

[67] R. Kamal, N. H. Tran, and C. Hong, "Event-based middleware for healthcare applications," *Journal of Communications and Networks*, vol. 14, no. 3, pp. 296–309, 2012.

[68] J. Zhang, K. Zhang, Y. Yang, J. Sun, T. Ling, G. Wang, Y. Ling, and D. Peng, "Grid-based implementation of XDS-I as part of image-enabled EHR for regional healthcare in Shanghai," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 2, pp. 273–284, 2010.

[69] V. Koufi and G. Vassilacopoulos, "HDGPortal: A Grid portal application for pervasive access to process-based healthcare systems," in *Second International Conference on Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008*, Tampere, Finland, 2008, pp. 121–126.

[70] A.-S. Hosam, M. Abbas, S. Ahmad, and A. Mustafa, "Intelligent Agent System Architecture for Presenting Health Grid Contents from Complex Database," in *2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, Liverpool, United Kingdom, 2010, pp. 38–42.

[71] H. Viswanathan, E. K. Lee, and D. Pompili, "Mobile grid computing for data- and patient-centric ubiquitous healthcare," in *2012 First IEEE Workshop on Enabling Technologies for Smartphone and Internet of Things (ETSIoT)*, Seoul, Korea, 2012, pp. 36–41.

[72] J. Calvillo, I. Román, S. Rivas, and L. M. Roa, "Privilege Management Infrastructure for Virtual Organizations in Healthcare Grids," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 316–323, 2011.

[73] T. Boyd, T. Savel, G. Kesarinath, B. Lee, and J. Stinn, "The Use of Public Health Grid Technology in the United States Centers for Disease Control and Prevention H1N1 Pandemic Response," in *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, Perth, Australia, 2010, pp. 974–978.

[74] N. Preve, "Ubiquitous Healthcare Computing with Sensor Grid Enhancement with Data Management System (SEGEDMA)," *Journal of Medical Systems*, vol. 35, no. 6, pp. 1375–1392, 2011.

[75] H. K. Huang, "Utilization of medical imaging informatics and biometrics technologies in healthcare delivery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 1–2, pp. 27–39, May 2008.

[76] A. Naseer and L. K. Stergioulas, "Web-Services-Based Resource Discovery Model and Service Deployment on HealthGrids," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 838–845, 2010.

[77] A. Stell, R. Sinnott, O. Ajayi, and J. Jiang, "Designing Privacy for Scalable Electronic Healthcare Linkage," in *International Conference on Computational Science and Engineering, 2009. CSE '09*, 2009, vol. 3, pp. 330–336.

[78] P. Hu, L. Sun, and E. Ifeachor, "A Framework for Bioprofile Analysis Over Grid," *IEEE Systems Journal*, vol. 3, no. 4, pp. 520–535, 2009.

[79] A. Naseer, L. K. Stergioulas, S. Hammoud, and M. Li, "Grid-based Semantic Integration and dissemination of medical information," in *Third International Conference on Digital Information Management, 2008. ICDIM 2008*, 2008, pp. 272–278.

[80] S.-J. Oh and C.-W. Lee, "u-Healthcare SensorGrid Gateway for connecting Wireless Sensor Network and Grid Network," in *10th International Conference on Advanced Communication Technology, 2008. ICACT 2008*, 2008, vol. 1, pp. 827–831.

[81] D. Kyriazis, K. Tserpes, G. Kousiouris, A. Menychtas, G. Katsaros, and T. Varvarigou, "Data Aggregation and Analysis: A Grid-Based Approach for Medicine and Biology," in *International Symposium on Parallel and Distributed Processing with Applications, 2008. ISPA '08*, 2008, pp. 841–848.

[82] W. M. Ahmed, B. Bayraktar, A. K. Bhunia, E. D. Hirleman, J. P. Robinson, and B. Rajwa, "Rapid Detection and Classification of Bacterial Contamination Using Grid Computing," in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007. BIBE 2007*, 2007, pp. 286–293.

[83] U. Barcaro, M. Righi, P. P. Ciullo, E. Palanca, K. Cerbioni, A. Starita, S. Di Bona, and D. Guerri, "A Decision Support System for the Acquisition and Elaboration of EEG Signals: the AmI-GRID Environment," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007*, 2007, pp. 4331–4334.

[84] M. Olive, H. Rahmouni, and T. Solomonides, "A European Healthgrid Roadmap", Share Project - Supporting and Structuring Healthgrid Activities &

Research in Europe: Developing a Roadmap." The Publications Office of the European Union, 2008.

[85] T. Solomonides and K. Dean, "Review of HealthGrid 2008: 'Global HealthGrid: eScience meets Biomedical Informatics'," in *CBMS '08, Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*, Jyvaskyla, Finland, 2008, pp. 342–342.

[86] J. N. Tsitsiklis, "A comparison of Jacobi and Gauss-Seidel parallel iterations. Applied," *IEEE Trans. Aut. Control*, vol. 2, no. 2, pp. 325–332, 1989.

[87] R. Mehmood and J. Crowcroft, "Parallel Iterative Solution Method for Large Sparse Linear Equation Systems," University of Cambridge, Cambridge, United Kingdom, Technical Report UCAM-CL-TR-650, Oct. 2005.

[88] W. J. Stewart, *Introduction to the numerical solution of Markov chains*. Princeton, N.J.: Princeton University Press, 1994.

[89] Y. Saad and H. A. V. D. Vorst, "Iterative Solution of Linear Systems in the 20-th Century," *Journal of Computational and Applied Mathematics*, vol. 123, pp. 1–33, 2000.

[90] Y. Saad, *Iterative methods for sparse linear systems*, 2 edition. Philadelphia, Pa, USA: Society for Industrial and Applied Mathematics, 2003.

[91] B. Philippe, Y. Saad, and W. J. Stewart, "Numerical Methods in Markov Chain Modelling," *Operations Research*, vol. 40, pp. 1156–1179, 1996.

[92] B. J. S. Chee and C. Franklin Jr, *Cloud Computing: Technologies and Strategies of the Ubiquitous Data Center*, 1st ed. London, United Kingdom: CRC Press, Inc., 2010.

[93] D. C. Plummer, D. W. Cearley, and D. M. Smith, "Cloud Computing Confusion Leads to Opportunity," Gartner Inc., Stamford, CT, USA, Market Analysis and Statistics Report G00159034, Jun. 2008.

[94] F. Gens, "IT Cloud Services User Survey, pt.2: Top Benefits & Challenges," *IDC eXchange & Blog Archive & IT Cloud Services User Survey, pt.2: Top Benefits & Challenges*, 2008..

[95] W. Fellows, "Partly Cloudy – Blue-Sky Thinking About Cloud Computing," The 451 Group, London, United Kingdom, Business Report, Jun. 2008.

[96] S. A. Mertz, C. Eschinger, T. Eid, and B. Pring, "Dataquest Insight: SaaS Demand Set to Outpace Enterprise Application Software Market Growth," Gartner Inc., Stamford, CT, USA, Market Analysis and Statistics Report G00150222, Aug. 2007.

[97] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Electrical Engineering and Computer Sciences University of California at Berkeley, Berkeley, CA, USA, Technical Report UCB/EECS-2009-28, Feb. 2009.

[98] G. Reese, *Cloud Application Architectures: Building Applications and Infrastructure in the Cloud*, 1st ed. O'Reilly Media, 2009.

[99] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared," in *Proceedings of the Grid Computing Environments Workshop, 2008. GCE '08*, Austin, Texas, USA, 2008, vol. abs/0901.0, pp. 1–10.

[100] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, pp. 50–55, 2009.

[101] V. (J. R.. Winkler, *Securing the Cloud: Cloud Computer Security Techniques and Tactics*, 1st ed. Waltham, Massachusetts, USA: Syngress, 2011.

[102] I. Menken and G. Blokdijk, *Cloud Computing - The Complete Cornerstone Guide to Cloud Computing Best Practices: Concepts, Terms, and Techniques for Successfully Planning, Implementing, and Managing Enterprise IT Cloud Computing Technology*, 2nd ed. Brisbane, Queensland, Australia: Emereo Publishing, 2009.

[103] C. Vecchiola, S. Pandey, and R. Buyya, "High-Performance Cloud Computing: A View of Scientific Applications," in *ISPAN '09 Proceedings of the 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks*, Aohsiung, Taiwan, 2009, pp. 1–13.

[104] I. Menken, *Cloud Computing - The Complete Cornerstone Guide to Cloud Computing Best Practices: Concepts, Terms, and Techniques for Successfully Planning, Implementing, and Managing Enterprise IT Cloud Computing Technology*, 1st ed. Brisbane, Queensland, Australia: Emereo Publishing, 2008.

[105] R. Buyya, Chee Shin Yeo, and S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities," in *HPCC 08' Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications*, Dalian, China, 2008, pp. 5–13.

[106] N. Antonopoulos and L. Gillam, *Cloud Computing: Principles, Systems and Applications*, 1st Edition. Berlin Heidelberg, Germany: Springer, 2010.

[107] R. L. Krutz and R. D. Vines, *Cloud Security: A Comprehensive Guide to Secure Cloud Computing*, 1st ed. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2010.

[108] E. A. Marks and B. Lozano, *Executive's Guide to Cloud Computing*, 1st ed. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2010.

[109] M. Miller, *Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online*, 1st ed. Indianapolis, Indiana, USA: Que Publishing, 2008.

[110] E. Ramos, K. Acker, R. Green, and S. Llaurency, *IBM Redbooks | Cloud Computing and the Value of zEnterprise*. North Castle, New York, USA: IBM Redbooks publication, IBM Press, 2011.

[111] IBM, "Seeding the Clouds: Key Infrastructure Elements for Cloud Computing." White Paper, IBM Corporation, 2009.

[112] T. Velte, A. Velte, and R. Elsenpeter, *Cloud Computing, A Practical Approach*, 1st ed. McGraw-Hill Osborne Media, 2009.

[113] Amazon Web Services, "Amazon Simple Storage Service (Amazon S3)," *Amazon Simple Storage Service (Amazon S3*, 2012. [Online]. Available: http://aws.amazon.com/s3/. [Accessed: 08-Jul-2012].

[114] M. Jayalal, R. Jehadeesan, S. Rajeswari, and S. Satya, "Moving From Grid to Cloud Computing: The Challenges in an Existing Computational Grid Setup," *IJCSC*, vol. 1, no. 2, pp. 415–418, 2010.

[115] E. Kourpas, "Grid Computing: Past, Present and Future – An Innovation Perspective." White Paper, IBM Corporation, 2006.

[116] T. B. Winans and J. S. Brown, "Cloud computing - A collection of working papers." Deloitte Development LLC, 2009.

[117] M. Böhm, S. Leimeister, C. Riedl, and H. Krcmar, "Cloud Computing – Outsourcing 2.0 or a new Business Model for IT Provisioning?," in *Application*

Management, F. Keuper, C. Oecking, and A. Degenhardt, Eds. Wiesbaden: Gabler, 2011, pp. 31–56.

[118] J. Buck, "Cloud Confusion Amongst IT Professionals," *Cloud Confusion Amongst IT Professionals - Version One*, 2009. [Online]. Available: http://www.versionone.co.uk/news/cloud-of-confusion-amongst-it-professionals.php. [Accessed: 18-Jun-2012].

[119] B. Lesieure and C. Baroudi, "Business Adoption of Cloud Computing: Reduce Cost, Complexity and Energy Consumption," Aberdeen Group Inc, Boston, MA, USA, Market Analysis and Statistics Report, Sep. 2009.

[120] M. Behrendt, B. Glasner, P. Kopp, R. Dieckmann, G. Breiter, S. Pappe, H. Kreger, and A. Arsanjani, "Introduction and Architecture Overview IBM Cloud Computing Reference Architecture 2.0." IBM Press, 2011.

[121] HP, "Understanding the HP CloudSystem Reference Architecture." Hewlett-Packard Development Company, White paper, 4AA3-4548ENW, 2011.

[122] Microsoft Corporation, "Hyper-V Cloud Fast Track Program, Reference Architecture Technical White Paper." Microsoft Corporation, 2011.

[123] DMTF Informational, "Architecture for Managing Clouds: A White Paper from the Open Cloud Standards Incubator." Distributed Management Task Force, White paper, DSP-IS0102, 2010.

[124] DMTF Informational, "Use Cases and Interactions for Managing Clouds : A White Paper from the Open Cloud Standards Incubator." Distributed Management Task Force, White paper, DSP-IS0103, 2010.

[125] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, "NIST Cloud Computing Reference Architecture: Recommendations of the National Institute of Standards and Technology," NIST, Gaithersburg, MD, USA, Technical Report NIST SP - 500-292, 2011.

[126] P. Mell and T. Grance, "The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology," NIST, Gaithersburg, MD, USA, Technical Report 800-145, 2011.

[127] OASIS, "OASIS Privacy Management Reference Model (PMRM) Technical Committee," *OASIS | Advancing open standards for the information society*, 2012. [Online]. Available: https://www.oasis-open.org/committees/pmrm/charter.php. [Accessed: 14-Oct-2012].

[128] S. Ried, H. Kisker, P. Matzke, C. Mines, T. Mendel, J. R. Rymer, and M. Lisserman, "The evolution of cloud computing markets," Forrester Research, Inc, Cambridge, MA, USA, Technical Report 57232, Jul. 2010.

[129] L. Youseff, M. Butrico, and D. Da Silva, "Toward a Unified Ontology of Cloud Computing," in *Proceedings of the 2008 Grid Computing Environments Workshop (GCE)*, Austin, Texas, USA, 2008, vol. 1, pp. 1–10.

[130] B. Hayes, "Cloud computing: As software migrates from local PCs to distant internet servers, users, and developers alike go along for the ride," *Communications of the ACM*, vol. 51, no. 7, pp. 9–11, Jul-2008.

[131] S. A. Mertz, T. Eid, C. Eschinger, H. H. Swinehart, C. Pang, and B. Pring, "Market Trends: Software as a Service, Worldwide, 2007-2012," Gartner Inc., Stamford, CT, USA, Market Analysis and Statistics Report G00160847, Sep. 2008.

[132] Appistry Inc., "Unlocking the Promise of Cloud Computing for the Enterprise: Achieving scalability, agility and reliability with cloud application platforms." White Paper, Appistry Inc., St. Louis, Missouri, USA, 2009.

[133] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Virtual Infrastructure Management in Private and Hybrid Clouds," *IEEE Internet Computing*, vol. 13, no. 5, pp. 14–22, Oct. 2009.

[134] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The Eucalyptus Open-Source Cloud-Computing System," in *CCGRID '09 Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, Shanghai, China, 2009, pp. 124–131.

[135] M. D. de Assuncao, A. di Costanzo, and R. Buyya, "Evaluating the cost-benefit of using cloud computing to extend the capacity of clusters," in *HPDC '09 Proceedings of the 18th ACM international symposium on High performance distributed computing*, Munich, Germany, 2009, pp. 141–150.

[136] M. H. Hugos and D. Hulitzky, *Business in the Cloud: What Every Business Needs to Know About Cloud Computing*, 1st ed. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2010.

[137] H. Takabi, J. B. . Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments," *IEEE Educational Activities Department, Piscataway, NJ, USA*, vol. 8, no. 6, pp. 24–31, Dec. 2010.

[138] Platform Computing, "Platform ISF: End-to-end private cloud management software." White Paper, Platform Computing Corporation. Ontario, Canada, 2009.

[139] C. Gong, J. Liu, Q. Zhang, H. Chen, and Z. Gong, "The Characteristics of Cloud Computing," in *Proceedings of the 39th International Conference on Parallel Processing Workshops 2010*, San Diego, CA, USA, 2010, pp. 275–279.

[140] S. Ramgovind, M. M. Eloff, and E. Smith, "The Management of Security in Cloud Computing," in *Proceedings of the Information Security for South Africa (ISSA), 2010*, Johannesburg, Gauteng, South Africa, 2010, pp. 1–7.

[141] C. Babcock, *Management Strategies for the Cloud Revolution: How Cloud Computing Is Transforming Business and Why You Can't Afford to Be Left Behind*, 1st ed. Maidenhead, Berkshire, UK: McGraw-Hill, 2010.

[142] R. Craig, J. Frazier, N. Jacknis, S. Murphy, C. Purcell, and P. Spencer, "Cloud Computing in the Public Sector: Public Manager's Guide to Evaluating and Adopting Cloud Computing." White paper, Pearson Education, Cisco Press, Indianapolis, Indiana, USA, 2009.

[143] G. Fenu and S. Surcis, "A Cloud Computing Based Real Time Financial System," in *ICN '09 Proceedings of the 2009 Eighth International Conference on Networks*, Gosier, Guadeloupe, France, 2009, pp. 374–379.

[144] D. Lin and A. Squicciarini, "Data Protection Models for Service Provisioning in the Cloud," in *SACMAT '10 Proceedings of the 15th ACM symposium on Access control models and technologies*, Pittsburgh, Pennsylvania, USA, 2010, pp. 183–192.

[145] J. Rosenberg and A. Mateos, *The Cloud at Your Service: The when, how, and why of enterprise cloud computing*, 1st ed. Shelter Island, New York, USA: Manning Publications Co, 2010.

[146] B. Sosinsky, *Cloud Computing Bible*, 1st ed. Wiley Publishing, Inc., Indianapolis, Indiana, 2011.

[147] P. Simmonds, C. Rezek, and A. Reed, "Security Guidance for critical areas of focus in Cloud Computing V3.0," The Cloud Security Alliance (CSA), Ferndale, WA, USA, Technical Report V3.0, Nov. 2011.

[148] K. Schmotzer and B. J. Donovan, *IBM Redbooks | IBM LotusLive: A Social Networking and Collaboration Platform for the Midmarket*. North Castle, New York, USA: IBM Redbooks publication, IBM Press, 2011.

[149] A. Surana, D. S. Vellal, and R. Guru, "Introducing IBM LotusLive," *IBM*, 14-Oct-2009. [Online]. Available: http://www.ibm.com/developerworks/lotus/library/lotuslive-intro/. [Accessed: 22-Jul-2012].

[150] D. Amrhein, "IBM & Cloud Computing: Self-Service Clouds with Fine-Grained Control, WebSphere CloudBurst provides self-service access with controls," *SYS-CON MEDIA*, 2009. [Online]. Available: http://websphere.sys-con.com/node/1029500. [Accessed: 18-Jul-2012].

[151] D. Chappell, "Introducing Windows Azure." White paper, Chappell & Associates, San Francisco, California, USA, 2010.

[152] D. Chappell, "Introducing the Azure Services Platform, an Early Look At Windows Azure, .Net Services, SQL Services, and Live Services." White paper, Chappell & Associates, San Francisco, California, USA, 2008.

[153] K. Pijanowskih, "IaaS, PaaS, and the Windows Azure Platform." White paper, Microsoft Press, Redmond, Washington, USA, 2009.

[154] A. W. S. AWS, "Overview of Amazon Web Services." White Paper, Amazon Web Services LLC, Herndon, Virginia, USA, 2010.

[155] Amazon Web Services, "About AWS," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/what-is-aws/. [Accessed: 09-Jul-2012].

[156] J. Varia, "Amazon Web Services - Architecting for The Cloud: Best Practices." White Paper, Amazon Web Services LLC, 2010.

[157] Amazon Web Services LLC, "Amazon Web Services The Economics of the AWS Cloud vs. Owned IT Infrastructure." White Paper, Amazon Web Services LLC, Herndon, Virginia, USA, 2009.

[158] J. Varia, "Amazon Web Services: Architecting for The Cloud: Best Practices." White Paper, Amazon Web Services LLC, Herndon, Virginia, USA, 2011.

[159] J. Murty, *Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB*, 1st Ed. Sebastopol, California, USA: O'Reilly Media, 2008.

[160] Amazon Web Services LLC, "Amazon Elastic Compute Cloud (Amazon EC2)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/ec2/. [Accessed: 08-Jul-2012].

[161] Amazon Web Services LLC, "Amazon Elastic MapReduce (Amazon EMR)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/elasticmapreduce/. [Accessed: 08-Jul-2012].

[162] Amazon Web Services LLC, "Auto Scaling," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/autoscaling/. [Accessed: 08-Jul-2012].

[163] Amazon Web Services LLC, "Elastic Load Balancing," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/elasticloadbalancing/. [Accessed: 08-Jul-2012].

[164] Amazon Web Services LLC, "Amazon CloudFront," *Amazon CloudFront CDN - Content Delivery Network | Content Distribution Network*, 2012. [Online]. Available: http://aws.amazon.com/cloudfront/. [Accessed: 08-Jul-2012].

[165] Amazon Web Services LLC, "Amazon SimpleDB," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/simpledb/. [Accessed: 24-Jun-2012].

[166] H. Dewan and R. C. Hansdah, "A Survey of Cloud Storage Facilities," in *Proceedings of the The 7th IEEE 2011 World Congress on Services (SERVICES 2011)*, Washington DC, USA, 2011, pp. 224 –231.

[167] Amazon Web Services LLC, "Amazon Relational Database Service (Amazon RDS)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/rds/. [Accessed: 08-Jul-2012].

[168] Amazon Web Services LLC, "Amazon CloudWatch," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/cloudwatch/. [Accessed: 10-Jul-2012].

[169] Amazon Web Services LLC, "AWS CloudFormation," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/cloudformation/. [Accessed: 08-Jul-2012].

[170] Amazon Web Services LLC, "Amazon Simple Queue Service (Amazon SQS)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/sqs/. [Accessed: 08-Jul-2012].

[171] Amazon Web Services LLC, "Amazon Route 53," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/route53/. [Accessed: 08-Jul-2012].

[172] Amazon Web Services LLC, "Amazon Virtual Private Cloud (Amazon VPC)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/vpc/. [Accessed: 08-Jul-2012].

[173] Amazon Web Services LLC, "Extend Your IT Infrastructure with Amazon Virtual Private Cloud." White Paper, Amazon Web Services LLC, Herndon, Virginia, USA, 2010.

[174] J. Barr, "AWS Direct Connect," *Amazon Web Services Blog: AWS Direct Connect*, 2011. [Online]. Available: http://aws.typepad.com/aws/2011/08/aws-direct-connect.html. [Accessed: 08-Jul-2012].

[175] Amazon Web Services LLC, "AWS Direct Connect," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/directconnect/. [Accessed: 08-Jul-2012].

[176] Amazon Web Services LLC, "Amazon Flexible Payments Service (Amazon FPS)," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/fps/. [Accessed: 08-Jul-2012].

[177] W. Vogels, "The Amazon Flexible Payments Service (Amazon FPS)," *The Amazon Flexible Payments Service (Amazon FPS) - All Things Distributed*, 2007. [Online]. Available: http://www.allthingsdistributed.com/2007/08/the_amazon_flexible_payment_se.html. [Accessed: 08-Jul-2012].

[178] Amazon Web Services LLC, "Amazon DevPay," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/devpay/. [Accessed: 08-Jul-2012].

[179] Mike, "Amazon DevPay Graduates to General Availability," *Amazon Web Services Blog: Amazon DevPay Graduates to General Availability*, 2008. [Online]. Available: http://aws.typepad.com/aws/2008/12/amazon-devpay-graduates-to-general-availability.html. [Accessed: 08-Jul-2012].

[180] T. Eicken, "Amazon's Elastic Block Store explained," *Amazon's Elastic Block Store explained, RightScale Blog*, 2008. [Online]. Available: http://blog.rightscale.com/2008/08/20/amazon-ebs-explained/. [Accessed: 08-Jul-2012].

[181] Amazon Web Services LLC, "Amazon Elastic Block Store (EBS)," *Elastic Block Store*, 2012. [Online]. Available: http://aws.amazon.com/ebs/. [Accessed: 08-Jul-2012].

[182] Amazon Web Services LLC, "AWS Import/Export," *AWS Import/Export*, 2012. [Online]. Available: http://aws.amazon.com/importexport/. [Accessed: 08-Jul-2012].

[183] Amazon Web Services LLC, "Amazon Mechanical Turk," *Amazon Mechanical Turk*, 2012. [Online]. Available: http://aws.amazon.com/mturk/. [Accessed: 08-Jul-2012].

[184] Amazon Web Services, "Amazon Mechanical Turk Developer Guide." White Paper, Amazon Web Services LLC, Herndon, Virginia, USA, 2012.

[185] M. Kehrli and K. C. Vasconez, "Highway Evacuations in Selected Metropolitan Areas : Assessment of Impediments," FHWA, Washington, DC., USA, Techical Report FHWA-HOP-10-059, Apr. 2010.

[186] B. Schweiger, P. Ehnert, and J. Schlichter, "Simulative Evaluation of the Potential of Car2X-Communication in Terms of Efficiency," in *Communication Technologies for Vehicles*, vol. 6596, T. Strang, A. Festag, A. Vinel, R. Mehmood, C. Rico Garcia, and M. Röckl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 155–164.

[187] R. Mehmood and M. Nekovee, "Vehicular Ad hoc and Grid Networks: Discussion, Design and Evaluation," in *Proceedings of 14th World Congress on Intelligent Transport Systems (ITS)*, Beijing, China, 2007, p. 8.

[188] N. Owens, A. Armstrong, P. Sullivan, C. Mitchell, D. Newton, R. Brewster, and T. Trego, "Traffic Incident Management Handbook," FHWA, McLean, VA, USA, Handbook Report FHWA-HOP-10-013, Jan. 2010.

[189] RITA, "Emergency Transportation Operations: The Approach," *RITA | ITS | Emergency Transportation Operations*, 2012. [Online]. Available: http://www.its.dot.gov/eto/eto_approach.htm. [Accessed: 20-Sep-2011].

[190] Metropolitan Government of Nashville and Davidson County, Tennessee, "Emergency Preparedness Survey for Davidson County." 2008.

[191] Committee on the Role of Public Transportation in Emergency Evacuation, "The Role of Transit in Emergency Evacuation," transportation research Board of the National Academies, Washington, D.C., USA, Special Report 294, 2008.

[192] J. Buckland and M. Rahman, "Community-based disaster management during the 1997 Red River Flood in Canada," *Disasters*, vol. 23, no. 2, pp. 174–191, Jun. 1999.

[193] R. Drake, "The Hierarchy of Emergency Preparedness," in *Safeguarding Homeland Security*, 1st ed., S. Hakim and E. A. Blackstone, Eds. New York, NY, USA: Springer New York, 2009, pp. 31–40.

[194] A. Buchenscheit, F. Schaub, F. Kargl, and M. Weber, "A VANET-Based Emergency Vehicle Warning System," in *(VNC) 09' Proceedings of 1st IEEE Vehicular Networking Conference*, Tokyo, Japan, 2009, pp. 1–8.

[195] S. R. . Rizvi, S. Olariu, M. E. Rizvi, and M. C. Weigle, "A Traffic Chaos Reduction Approach for Emergency Scenarios," in *IPCCC 07' Proceedings of the 26th IEEE International Performance Computing and Communications Conference*, New Orleans, Louisiana, USA, 2007, pp. 576–578.

[196] S. R. . Rizvi, M. E. Rizvi, and M. C. Weigle, "InVANETs for First Responders," in *SIGCOMM 08' Proceedings of the Special Interest Group on Data Communication (SIGCOMM 2008)*, Seattle, Washington, USA, 2008, pp. 443–444.

[197] J.-S. Park, U. Lee, S. Y. Oh, M. Gerla, and D. S. Lun, "Emergency related video streaming in VANET using network coding," in *VANET '06 Proceedings of the 3rd international workshop on Vehicular ad hoc networks*, Los Angeles, California, USA, 2006, pp. 102–103.

[198] R. W. Pazzi, K. Abrougui, C. Rezende, and A. Boukerche, "Service Discovery Protocols for VANET Based Emergency Preparedness Class of Applications: A Necessity Public Safety and Security," in *Information Systems, Technology and Management*, vol. 54, S. K. Prasad, H. M. Vin, S. Sahni, M. P. Jaiswal, and B. Thipakorn, Eds. Berlin, Heidelberg, Germany: Springer, 2010, pp. 1–7.

[199] M. A. Serhani and Y. Gadallah, "A Service Discovery Protocol for Emergency Response Operations Using Mobile Ad Hoc Networks," in *AICT '10 Proceedings of the 2010 Sixth Advanced International Conference on Telecommunications*, Barcelona, Spain, 2010, pp. 280–285.

[200] R. Mehmood, J. Crowcroft, S. Hand, and S. Smith, "Grid-Level Computing Needs Pervasive Debugging," in *GRID '05 Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, Seattle, Washington, USA, 2005, pp. 186–193.

[201] S. V. R. K. Rao and V. Diwanji, "WiMax'ble Pervasive Cloud – Empowering Next Generation Intelligent Railway Infrastructure," in *Communication Technologies for Vehicles*, vol. 6596, T. Strang, A. Festag, A. Vinel, R. Mehmood, C. Rico Garcia, and M. Röckl, Eds. Berlin, Heidelberg, Germany: Springer, 2011, pp. 58–68.

[202] "'Master Plan of the City of Al-Ramadi, Second Stage Report: Analysis of Existing Situation, Regional Context and Major Development Issues'."Hydrosult Center for Engineering Planning (HCEP), Montreal, Canada,, Nov-2009.

[203] R. Mehmood, "'Disk-Based Techniques for Efficient Solution of Large Markov Chains'," Ph.D. Thesis, School of Computer Science, University of Birmingham, UK,, 2004.

[204] R. Mehmood, "Towards Understanding Intercity Traffic Interdependencies," in *Proceedings of 14th World Congress on Intelligent Transport Systems (ITS)*, Beijing, China, 2007, pp. 1–8.

[205] G. Ayres and R. Mehmood, "On Discovering Road Traffic Information Using Virtual Reality Simulations," in *UKSIM '09 Proceedings of the UKSim 2009: 11th International Conference on Computer Modelling and Simulation*, Los Alamitos, CA, USA, 2009, pp. 411–416.

[206] M. J. Lighthill and G. B. Whitham, "On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 229, no. 1178, pp. 317–345, May 1955.

[207] P. I. Richards, "Shock Waves on the Highway," *Operations Research*, vol. 4, no. 1, pp. 42–51, 1956.

[208] S. Akioka and Y. Muraoka, "HPC Benchmarks on Amazon EC2," in *WAINA '10 Proceedings of the 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, Perth, Australia, 2010, pp. 1029 –1034.

[209] C. Atkinson, T. Schulze, and S. Klingert, "Modelling as a Service (MaaS): Minimizing the Environmental Impact of Computing Services," in *SERVICES '11 Proceedings of the 2011 IEEE World Congress on Services*, Washington, DC USA, 2011, pp. 519–523.

[210] P. L. Bannerman, "Cloud Computing Adoption Risks : State of Play," in *Proceedings of 17th Asia Pacific Software Engineering Conference, (APSEC 2010) Cloud Workshop*, Sydney, Australia, 2010, pp. 1–7.

[211] M. Barnes, H. Leather, and D. K. Arvind, "Emergency Evacuation using Wireless Sensor Networks," in *LCN '07 Proceedings of the 32nd IEEE Conference on Local Computer Networks*, Dublin, Ireland, 2007, pp. 851–857.

[212] B. Biocic, D. Tomic, and D. Ogrizovic, "Economics of the cloud computing," in *2011 Proceedings of the 34th International Convention MIPRO*, Opatija, Croatia, 2011, pp. 1438–1442.

[213] B. Blau, D. Neumann, C. Weinhardt, and S. Lamparter, "Planning and Pricing of Service Mashups," in *CECANDEEE '08 Proceedings of the 2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*, Washington, DC, USA, 2008, pp. 19–26.

[214] M. A. Bochicchio and A. Longo, "Modelling Contract Management for Cloud Services," in *Proceedings of IEEE 4th International Conference on Cloud Computing (CLOUD 2011)*, Washington, D.C., USA, 2011, pp. 332–339.

[215] M. A. Bochicchio, A. Longo, and C. Mansueto, "Contract Management for Cloud Services: Information modelling aspects," in *Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Dublin, Ireland, 2011, pp. 1035–1042.

[216] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," in *HPCS '09 Proceedings of the International Conference on High Performance Computing & Simulation, 2009.*, Leipzig, Germany, 2009, pp. 1–11.

[217] R. Buyya, S. Pandey, and C. Vecchiola, "Cloudbus Toolkit for Market-Oriented Cloud Computing," in *CloudCom '09 Proceedings of the 1st International Conference on Cloud Computing*, Beijing, China, 2009, pp. 24–44.

[218] K. Chard, S. Caton, O. Rana, and K. Bubendorfer, "Social Cloud: Cloud Computing in Social Networks," in *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD 2010)*, Miami, Florida, USA, 2010, pp. 99–106.

[219] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the cloud: the Montage example," in *SC '08 Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, Austin, Texas, USA, 2008, pp. 1–12.

[220] F. Donno and E. Ronchieri, "The Impact of Grid on Health Care Digital Repositories," in *Proceedings of the 42st Hawaii International International Conference on Systems Science (HICSS-42 2009)*, Big Island, Hawaii, USA, 2009, pp. 1–9.

[221] C. Evangelinos and C. Hill, "Cloud Computing for parallel Scientific HPC Applications: Feasibility of Running Coupled Atmosphere-Ocean Climate Models on Amazon's EC2," in *Proceedings of the Cloud Computing and Its Applications (CCA 2008)*, Chicago, Illinois, USA, 2008, pp. 1–6.

[222] Z. Fang, J. Chen, M. Yi, Z. Wu, and H. Qian, "Cloud Computing Business Model Based on Value Net Theory," in *Proceedings of the 2010 IEEE 7th International Conference on e-Business Engineering (ICEBE)*, Shanghai, China, 2010, pp. 462–469.

[223] I. Foster, "The globus toolkit for grid computing," in *Proceedings of the 1st IEEE International Symposium on Cluster Computing and the Grid (CCGrid'01)*, Brisbane, Australia, 2001, pp. 2–2.

[224] R. Gaofeng and C. Jing, "Online course development based on a public cloud computing infrastructure," in *In Proceedings of the 2nd International Conference on Networking and Digital Society (ICNDS), 2010*, Wenzhou, China, 2010, vol. 2, pp. 47 –50.

[225] P. Goyal, "Enterprise Usability of Cloud Computing Environments: Issues and Challenges," in *WETICE '10 Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Larissa, Greece, 2010, pp. 54–59.

[226] N. Guilbault and R. Guha, "Experiment setup for temporal distributed intrusion detection system on Amazon's elastic compute cloud," in *ISI'09 Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, Dallas, Texas, USA, 2009, pp. 300 –302.

[227] S. M. Habib, S. Ries, and M. Muhlhauser, "Cloud Computing Landscape and Research Challenges Regarding Trust and Reputation," in *Proceedings of the 2010 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC)*, Xian, Shaanxi, China, 2010, pp. 410–415.

[228] C. N. Hoefer and G. Karagiannis, "Taxonomy of cloud computing services," in *Proceedings of the IEEE GLOBECOM Workshops (GC Wkshps), 2010*, Miami, Florida, USA, 2010, pp. 1345–1350.

[229] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the Use of Cloud Computing for Scientific Workflows," in *Proceedings of the IEEE 4th International Conference on eScience, 2008.*, Indianapolis, Indiana, USA, 2008, pp. 640–645.

[230] K. R. Jackson, L. Ramakrishnan, K. Muriki, S. Canon, S. Cholia, J. Shalf, H. J. Wasserman, and N. J. Wright, "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference On Cloud Computing Technology And Science*, Indianapolis, USA, 2010, pp. 159 –168.

[231] K. H. Kim, A. Beloglazov, and R. Buyya, "Power-aware provisioning of Cloud resources for real-time services," in *MGC '09 Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science*, Urbana Champaign, Illinois, USA, 2009, pp. 1–6.

[232] S. G. Kim, H. Han, H. Eom, and H. Y. Yeom, "Toward a cost-effective cloud storage service," in *ICACT'10 Proceedings of the 12th international conference on Advanced communication technology*, Dublin, Ireland, 2010, vol. 1, pp. 99–102.

[233] E. Lagerspetz and S. Tarkoma, "Mobile search and the cloud: The benefits of offloading," in *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Seattle, Washington, USA, 2011, pp. 117–122.

[234] Z. Lu, J. Wu, and W. Fu, "A Novel Cloud-Oriented WS-Management-Based Resource Management Model," in *ICWS '10 Proceedings of the 2010 IEEE International Conference on Web Services*, Miami, Florida, USA, 2010, pp. 676–677.

[235] P. Murray, “Enterprise Grade Cloud Computing,” in *WDDM '09 Proceedings of the Third Workshop on Dependable Distributed Data Management*, Nuremberg, Germany, 2009, pp. 1–1.

[236] S. K. Nair, S. Porwal, T. Dimitrakos, A. J. Ferrer, J. Tordsson, T. Sharif, C. Sheridan, M. Rajarajan, and A. U. Khan, “Towards Secure Cloud Bursting, Brokerage and Aggregation,” in *ECOWS '10 Proceedings of the 2010 Eighth IEEE European Conference on Web Services*, Ayia Napa, Cyprus, 2010, pp. 189–196.

[237] X. Pu, L. Liu, Y. Mei, S. Sivathanu, Y. Koh, and C. Pu, “Understanding Performance Interference of I/O Workload in Virtualized Cloud Environments,” in *CLOUD '10 Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing*, Miami, Florida, USA, 2010, pp. 51–58.

[238] B. P. Rimal, E. Choi, and I. Lumb, “A Taxonomy and Survey of Cloud Computing Systems,” in *NCM '09 Proceedings of the 2009 Fifth International Joint Conference on INC, IMS and IDC*, Seoul, Korea, 2009, pp. 44–51.

[239] Y. Singh, F. Kandah, and W. Zhang, “A Secured Cost-Effective Multi-Cloud Storage in Cloud Computing,” in *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Shanghai, China, 2011, pp. 619–624.

[240] F. Teng and F. Magoules, “Resource Pricing and Equilibrium Allocation Policy in Cloud Computing,” in *CIT '10 Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology*, Bradford, West Yorkshire, United Kingdom, 2010, pp. 195–202.

[241] R. Vilaça and R. Oliveira, “Clouder: a flexible large scale decentralized object store: architecture overview,” in *WDDM '09 Proceedings of the Third Workshop on Dependable Distributed Data Management*, Nuremberg, Germany, 2009, pp. 25–28.

[242] M. A. Vouk, “Cloud Computing - Issues, Research and Implementations,” in *Proceedings of The 2008 30th International Conference on Information Technology Interfaces (ITI)*, Dubrovnik, Croatia, 2008, pp. 31–40.

[243] Z. Wan, “Cloud Computing Infrastructure for Latency Sensitive Applications,” presented at the ICCT 10’ Proceedings of the 2010 12th IEEE International Conference on Communication Technology (ICCT), Nanjing, Jiangsu, China, 2010, pp. 1399–1402.

[244] D. Wentzlaff, C. Gruenwald, N. Beckmann, K. Modzelewski, A. Belay, L. Youseff, J. Miller, and A. Agarwal, “An operating system for multicore and clouds,” in *SoCC '10 Proceedings of the 1st ACM symposium on Cloud computing*, Indianapolis, Indiana, USA, 2010, p. 3.

[245] D. Wentzlaff, C. Gruenwald, N. Beckmann, K. Modzelewski, A. Belay, L. Youseff, J. Miller, and A. Agarwal, “An Operating System for Multicore and Clouds: Mechanisms and Implementation,” in *SoCC '10 Proceedings of the 1st ACM symposium on Cloud computing*, Indianapolis, Indiana, USA, 2010, pp. 3–14.

[246] T. Wood, E. Cecchet, K. K. Ramakrishnan, P. Shenoy, J. Van Der Merwe, and A. Venkataramani, “Disaster Recovery as a Cloud Service: Economic Benefits & Deployment Challenges,” in *HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, Boston, Massachusetts, USA, 2010, pp. 1–7.

[247] L. Wu, S. K. Garg, and R. Buyya, “SLA-Based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments,” in

CCGRID '11 Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Newport Beach, California, USA, 2011, pp. 195–204.

[248] P. Xiong, Y. Chi, S. Zhu, H. J. Moon, C. Pu, and H. Hacigumus, "Intelligent Management of Virtualized Resources for Database Systems in Cloud Environment," in *ICDE '11 Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, Hanover, Germany, 2011, pp. 87–98.

[249] Y. Zhang, G. Huang, X. Liu, and H. Mei, "Integrating Resource Consumption and Allocation for Infrastructure Resources on-Demand," presented at the Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), Miami, Florida, USA, 2010, pp. 75–82.

[250] R. Accorsi, L. Lowis, and Y. Sato, "Automated Certification for Compliant Cloud-based Business Processes," *Bus Inf Syst Eng*, vol. 3, no. 3, pp. 145–154, Apr. 2011.

[251] O. Ardaiz and L. Navarro, "Grid-based dynamic service overlays," *Future Generation Computer Systems*, vol. 24, no. 8, pp. 813–823, Oct. 2008.

[252] S. Bourbonnais, V. M. Gogate, L. M. Haas, R. W. Horman, S. Malaika, I. Narang, and V. Raman, "Towards an information infrastructure for the grid," *IBM Systems Journal, Grid Computing*, vol. 43, no. 4, pp. 665–688, Oct. 2004.

[253] J. Broberg, S. Venugopal, and R. Buyya, "Market-oriented Grids and Utility Computing: The State-of-the-art and Future Directions," *Journal of Grid Computing*, vol. 6, no. 3, pp. 255–276, Sep. 2008.

[254] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems, The International Journal of Grid Computing and eScience*, vol. 25, no. 6, pp. 599–616, Jun. 2009.

[255] C.-H. Chiu, H.-T. Lin, and S.-M. Yuan, "CloudEdge: a content delivery system for storage service in cloud environment," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 6, no. 4, pp. 252 – 262, Sep. 2010.

[256] J. R. Corney, C. Torres-Sanchez, A. P. Jagadeesan, and W. C. Regli, "Outsourcing labour to the cloud," *International Journal of Innovation and Sustainable Development, Inderscience Publishers*, vol. 4, no. 4, pp. 294 – 313, 2009.

[257] M. Cusumano, "Cloud computing and SaaS as new computing platforms," *Commun. ACM*, vol. 53, no. 4, pp. 27–29, Apr. 2010.

[258] A. di Costanzo, M. D. de Assuncao, and R. Buyya, "Harnessing Cloud Technologies for a Virtualized Distributed Computing Infrastructure," *IEEE Internet Computing*, vol. 13, no. 5, pp. 24–33, Oct. 2009.

[259] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, "Cloud Computing: Distributed Internet Computing for IT and Scientific Research," *IEEE Internet Computing*, vol. 13, no. 5, pp. 10–13, Oct. 2009.

[260] N. G. Duffield, P. Goyal, A. Greenberg, P. Mishra, K. K. Ramakrishnan, and J. E. van der Merwe, "Resource management with hoses: point-to-cloud services for virtual private networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 5, pp. 679– 692, Oct. 2002.

[261] P. T. Endo, A. V. de Almeida Palhares, N. N. Pereira, G. E. Goncalves, D. Sadok, J. Kelner, B. Melander, and J. Mangs, "Resource allocation for distributed cloud: concepts and research challenges," *IEEE Network*, vol. 25, no. 4, pp. 42–46, Aug. 2011.

[262] J. S. Erickson, S. Spence, M. Rhodes, D. Banks, J. Rutherford, E. Simpson, G. Belrose, and R. Perry, "Content-Centered Collaboration Spaces in the Cloud," *IEEE Internet Computing*, vol. 13, no. 5, pp. 34–42, Oct. 2009.

[263] I. Foster, "The Grid: Computing Without Bounds.," *Scientific American*, vol. 288, no. 4, pp. 78–85, 2003.

[264] I. Foster, "Service-Oriented Science," *The American Association for the Advancement of Science*, vol. 308, no. 5723, pp. 814–817, May 2005.

[265] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding Cloud Computing Vulnerabilities," *IEEE Security & Privacy*, vol. 9, no. 2, pp. 50–57, Apr. 2011.

[266] A. Haque, S. M. Alhashmi, and R. Parthiban, "A survey of economic models in grid computing," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1056–1069, Oct. 2011.

[267] U. Helmbrecht, "Data protection and legal compliance in cloud computing," *DuD*, vol. 34, no. 8, pp. 554–556, Oct. 2010.

[268] C. N. Höfer and G. Karagiannis, "Cloud computing services: taxonomy and comparison," *J Internet Serv Appl*, Jun. 2011.

[269] P. T. Jaeger, J. Lin, J. M. Grimes, and S. N. Simmons, "Where is the cloud? Geography, economics, environment, and jurisdiction in cloud computing," *First Monday, Bridgman, MI, USA*, vol. 14, no. 5, pp. 1–11, 2009.

[270] J. M. Kaplan, M. Löffler, and R. P. Roberts, "Managing next generation IT infrastructure," *McKinsey & Company*, no. 3, pp. 1–9, 2004.

[271] K. Keahey, M. Tsugawa, A. Matsunaga, and J. Fortes, "Sky Computing," *IEEE Internet Computing*, vol. 13, no. 5, pp. 43–51, Oct. 2009.

[272] A. Khajeh-hosseini, D. Greenwood, J. W. Smith, and I. Sommerville, "The Cloud Adoption Toolkit: Supporting Cloud Adoption Decisions," *IN THE ENTERPRISE," IN PRESS, SOFTWARE: PRACTICE AND EXPERIENCE*, 2011.

[273] K. Kifayat, M. Merabti, and Q. Shi, "Future security challenges in cloud computing," *International Journal of Multimedia Intelligence and Security*, vol. 1, no. 4, pp. 428 – 442, 2010.

[274] H. Kreger, "Fulfilling the Web services promise," *Communications of the ACM*, vol. 46, no. 6, pp. 29–34, 2003.

[275] K. L. Kroeker, "Grid computing's future," *Communications of the ACM*, vol. 54, no. 3, pp. 15–17, Mar. 2011.

[276] B. Leiba, "Having One's Head in the Cloud," *IEEE Internet Computing*, vol. 13, no. 5, pp. 4–6, Oct. 2009.

[277] K. W. Lin and D.-J. Deng, "A novel parallel algorithm for frequent pattern mining with privacy preserved in cloud computing environments," *International Journal of Ad Hoc and Ubiquitous Computing, Inderscience Publishers*, vol. 6, no. 4, pp. 205 – 215, 2010.

[278] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development, NY, USA*, vol. 54, no. 1, pp. 4:1–4:12, Feb. 2010.

[279] M. A. Lindner, L. M. Vaquero, L. Rodero-Merino, and J. Caceres, "Cloud economics: dynamic business models for business on demand," *International Journal of Business Information Systems*, vol. 5, no. 4, pp. 373 – 392, 2010.

[280] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing: The business perspective," *Decision Support Systems*, vol. 51, no. 1, pp. 176–189, Apr. 2011.

[281] H. Motahari-Nezhad, B. Stephenson, and S. Singha, "Outsourcing Business to Cloud Computing Services: Opportunities and Challenges," *IEEE IT Professional, Special Issue on Cloud Computing*, vol. 11, no. 2, pp. 1–17, 2009.

[282] K. Mukherjee and G. Sahoo, "Performance Analysis of Cloud Computing using Multistage Ant System," *International Journal of Computer Applications, Foundation of Computer Science*, vol. 1, pp. 75–80, Feb. 2010.

[283] M. Olive, H. Rahmouni, and T. Solomonides, "From HealthGrid to SHARE: a selective review of projects," *Studies in Health Technology and Informatics*, vol. 126, pp. 306–313, 2007.

[284] B. P. Rimal, A. Jukan, D. Katsaros, and Y. Goeleven, "Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach," *J Grid Computing*, vol. 9, no. 1, pp. 3–26, Dec. 2010.

[285] T. Rings, G. Caryer, J. Gallop, J. Grabowski, T. Kovacikova, S. Schulz, and I. Stokes-Rees, "Grid and Cloud Computing: Opportunities for Integration with the Next Generation Network," *J Grid Computing*, vol. 7, no. 3, pp. 375–393, Aug. 2009.

[286] M. Sahinoglu and L. Cueva-Parra, "CLOUD computing," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 1, pp. 47–68, Jan. 2011.

[287] Shicong Meng, Ling Liu, and Ting Wang, "State Monitoring in Cloud Datacenters," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1328–1344, Sep. 2011.

[288] T. Truong Huu, G. Koslovski, F. Anhalt, J. Montagnat, and P. Vicat-Blanc Primet, "Joint Elastic Cloud and Virtual Network Framework for Application Performance-cost Optimization," *J Grid Computing*, vol. 9, no. 1, pp. 27–47, Nov. 2010.

[289] H.-W. Wang, "Finance e-learning and simulation toward the cloud service environment," *International Journal of Internet Protocol Technology*, vol. 5, no. 4, pp. 210–218, Mar. 2010.

[290] M. M. Yunis, "A 'cloud-free' security model for cloud computing," *International Journal of Services and Standards*, vol. 5, no. 4, pp. 354 – 375, 2009.

[291] M. Yuriyama and T. Kushida, "Integrated cloud computing environment with IT resources and sensor devices," *International Journal of Space-Based and Situated Computing*, vol. 1, no. 2/3, pp. 163 – 173, 2011.

[292] P. Hofmann and D. Woods, "Cloud Computing: The Limits of Public Clouds for Business Applications," *IEEE Internet Computing, Information Overload*, vol. 14, no. 6, pp. 90–93, Nov-2010.

[293] I. Blanquer, S. Lloyd, R. McClatchey, J. Montagnat, H. Bilofsky, T. Solomonides, I. Oliveira, B. Claerhout, and J. Herveg, "HealthGrid initiative,HealthGrid White Pape." The HealthGrid initiative, 22-Sep-2004.

[294] P. Chaganti, "Cloud computing with Amazon Web Services, Part 1: Introduction." IBM, 2008.

[295] R. Dargha, "Cloud Computing: Key Considerations for Adoption." Whitepaper, Infosys Technologies, Bangalore, India, 2009.

[296] W. Forrest and C. Barthold, "Clearing the Air on Cloud Computing." McKinsey and Company, 2009.

[297] J. Hagel and J. S. Brown, "Cloud computing Storms on the horizon." Deloitte Development LLC, 2010.

[298] R. Lovell, "Introduction to cloud computing." White Paper, ThinkGrid, 2012.

[299] G. Morton and T. Alford, “The economics of cloud computing: Addressing the benefits of infrastructure in the cloud.” Booz Allen Hamilton, 2009.

[300] D. Paessler, “Network Visibility: The Key to Risk Management -- Enterprise Systems.” the network monitoring company, 2008.

[301] J. Varia, “Cloud Architecturs.” White Paper, Amazon Web Services LLC, 2008.

[302] T. Winans and J. Brown, “Cloud Computing: A Collection Of Working Papers.” Deloitte Consulting LLP, 2009.

[303] “Emergency Preparedness Survey for Davidson County.” Metropolitan Government of Nashville and Davidson County, Tennessee, 2008.

[304] VmWare Inc., “VMware vSphere, the First Cloud Operating System, Provides an Evolutionary, Non-disruptive Path to Cloud Computing.” VMWare Inc. White Paper, 2009.

[305] K. Broderick, “Worldwide Enterprise Server Cloud Computing 2011–2015 Forecast,” International Data Corporation (IDC), Framingham, MA, USA, Market Analysis 228916, Jun. 2011.

[306] D. Catteddu and G. Hogben, “Cloud Computing. Benefits, risks and recommendations for information security. European Network and Information Security Agency,” ENISA, Heraklion, Crete, Greece, Educational and Information Report 460/2004, Nov. 2009.

[307] J. Cole, “Emergency Response and Civil Defence,” RUSI, London, United Kingdom, Information Report, Dec. 2007.

[308] E. T. Curtiss and S. Eustis, “Worldwide Cloud Computing Market Opportunities and Segment Forecasts 2009 to 2015,” WinterGreen Research, Inc., Lexington, MA, USA, Market Analysis and Statistics Report SH24131315, 2009.

[309] E. T. Curtiss and S. Eustis, “Cloud Computing Systems Provide IT Efficiency - Capacity on Demand,” WinterGreen Research, Inc., Lexington, MA, USA, Market Analysis and Statistics Report SH24381513, 2010.

[310] E. T. Curtiss and S. Eustis, “Cloud Middleware Market Shares, Strategies, and Forecasts, Worldwide, 2011 to 2017,” WinterGreen Research, Inc., Lexington, MA, USA, Market Analysis and Statistics Report SH24721513, 2011.

[311] L. Lachal and S. Mann, “Planning for Cloud Computing: Understanding the organizational, governance, and cost implications,” Ovum Plc, London, United Kingdom, IT Management and Strategy Report OI00005-006, Nov. 2010.

[312] R. P. Maccubbin, B. L. Staples, F. Kabir, C. F. Lowrance, M. R. Mercer, B. H. Philips, and S. R. Gordon, “Intelligent Transportation Systems Benefits, Costs, Deployment, and Lessons Learned,” RITA, Washington, D.C, USA, Decision Making Report FHWA-JPO-08-032, Sep. 2008.

[313] J. Mathews, “The Future of Virtualization, Cloud Computing & Green IT - Global Technologies & Markets Outlook – 2011-2016,” Market Intel Group LLC (MIG), Colorado Springs, CO, USA, Market Analysis and Statistics Report MIGL2835681, Oct. 2010.

[314] H. Matthew, K. Wunderlich, and J. Bunch, “Structuring, Modeling, and Simulation Analyses for Evacuation Planning and Operations,” RITA, Washington, DC, USA, Decision Making Report FHWA—HOP-08-029, Jun. 2009.

[315] H. Nancy, W. John, M. Robin, K. Ram, K. John, B. Craig, S. Jeff, K. Susan, G. Kevin, and E. Andrea Vann, “Information Sharing Guidebook for Transportation Management Centers, Emergency Operations Centers, and

Fusion Centers," FHWA, McLean, VA, USA, Service Guidance Report FHWA-HOP-09-003, Jun. 2010.

[316] H. R. M. Nezhad, B. Stephenson, and S. Singhal, "Outsourcing Business to Cloud Computing Services: Opportunities and Challenges. HP Laboratories. HPL-2009-23," Hewlett-Packard Development Company, HP Labs, Palo Alto, CA, USA, Technical Report HPL-2009-23, Feb. 2009.

[317] D. C. Plummer, D. Mitchell Smith, T. J. Bittman, D. W. Cearley, D. J. Cappuccio, D. Scott, R. Kumar, and B. Robertson, "Five Refining Attributes of Public and Private Cloud Computing," Gartner Inc., Stamford, CT, USA, Market Analysis and Statistics Report G00167182, May 2009.

[318] L. W. Roeder Jr, "Harnessing Information and Technology for Disaster Management," Global Disaster Information Network (GDIN), South Riding, VA, USA, Disaster Information Task Force Report, Nov. 1997.

[319] J. R. Rymer, M. Gualtieri, M. Gilpin, and A. Anderson, "Cloud Computing Brings Demand For Elastic Application Platforms," Forrester Research, Inc, Cambridge, MA, USA, Technical Report 58569, Apr. 2011.

[320] L. Schubert, "The Future of Cloud Computing: Opportunities for European Cloud Computing Beyond 2010," European Commission CORDIS, Brussels, Belgium, Expert Group Report 6862917, Jan. 2010.

[321] D. Washburn, L. E. Nelson, J. Staten, C. Mines, and E. Chi, "Cloud Computing Helps Accelerate Green IT - Take Advantage Of Public And Private Cloud To Deliver Energy, Carbon, And E-Waste Efficiencies," Forrester Research, Inc, Cambridge, MA, USA, Market Analysis and Statistics Report, Jun. 2011.

[322] ABI Research, "Enterprise Mobile Cloud Computing: Cloud Services, Mobile Devices, and the IT Supply Chain Analysis," Allied Business Intelligence, Inc., Oyster Bay, NY, USA, Research Report RR-ECC-09, 2009.

[323] ABI Research, "Consumer Cloud Computing: Web-based Applications for E-Mail, Document Storage/Sharing, PC Protection, and Backup," Allied Business Intelligence, Inc., Oyster Bay, NY, USA, Research Report RR-CCC-10, 2010.

[324] Renub Research, "Cloud Computing – SaaS, PaaS, IaaS Market, Mobile Cloud Computing, M&A, Investments, and Future Forecast, Worldwide," Renub Research, Roswell, GA, USA, Market Analysis and Statistics Report RNBR2807915, Sep. 2010.

[325] Markets and Markets, "Cloud Computing Market - Global Forecast (2010 - 2015)," M&M, Wilmington, DE, USA, Market Research Report 1228, Oct. 2010.

[326] Global Industry Analysts, "Cloud Computing Services - A Global Market Report," Global Industry Analysts, Inc, San Jose, CA, USA, Market Analysis and Statistics Report MCP-6223, Oct. 2011.

[327] Markets and Markets, "Healthcare Cloud Computing (Clinical, EMR, SaaS, Private, Public, Hybrid) Market - Global Trends, Challenges, Opportunities & Forecasts (2012 – 2017)," M&M, Wilmington, DE, USA, Market Research Report MD 1562, Jul. 2012.

[328] D. Alger, *Build the Best Data Center Facility for Your Business*, 1st ed. Indianapolis, Indiana, USA: Pearson Education, Cisco Press, 2005.

[329] V. Atluri and P. Samarati, *Security of Data and Transaction Processing*, 1st ed., vol. 8. Berlin Heidelberg, Germany: Springer, 2000.

[330] N. Carr, *The Big Switch: Rewiring the World, from Edison to Google*. New York, USA: W. W. Norton & Company, Inc., 2008.

[331] M. Crouhy, D. Galai, and R. Mark, *The Essentials of Risk Management: The Definitive Guide for the Non-risk Professional.* New York, NY, USA: McGraw-Hill Professional, 2006.

[332] W. H. Dutton and M. Peltu, *Information and Communication Technologies: Visions and Realities.* Oxford, United Kingdom: Oxford University Press, 1996.

[333] I. Faulconbridge and M. Ryan, *Managing Complex Technical Projects: A Systems Engineering Approach*, 1st ed. Norwood, Massachusetts, USA: Artech House, 2002.

[334] B. Goyal and S. Lawande, *Grid Revolution: An Introduction to Enterprise Grid Computing*, 1st ed. Berkeley, CA, USA: McGraw-Hill Osborne Media, 2005.

[335] Y. Y. Haimes, *Risk Modeling, Assessment, and Management*, 2nd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.,, 2005.

[336] B. Halpert, *Auditing Cloud Computing: A Security and Privacy Guide*, 1st ed. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2011.

[337] R. Jennings, *Cloud Computing with the Windows Azure Platform*, 1st ed. Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 2009.

[338] F. Magoules, Ed., *Fundamentals of Grid Computing: Theory, Algorithms and Technologies*, 1st ed. London, United Kingdom: Chapman and Hall/CRC, 2009.

[339] C. Marrison, *The Fundamentals of Risk Measurement*, 1st ed. Maidenhead, Berkshire, UK: McGraw-Hill, 2002.

[340] J. N. Martin, *Systems Engineering Guidebook: A Process for Developing Systems and Products*, 1st ed. Boca Raton, Florida, USA: CRC Press, Inc., 1996.

[341] K. T. McDonald, *Above the Clouds: Managing Risk in the World of Cloud Computing*, 1st ed. Cambridgeshire, United Kingdom: IT Governance Ltd, 2010.

[342] D. A. McEntire, *Comparative Emergency Management: Understanding Disaster Policies, Organizations, and Initiatives from Around the World.* Emmitsburg, Maryland, USA: Emergency Management Institute, 2009.

[343] M. Nair, *Activity-Based Information Systems: An Executive's Guide to Implementation.* Indianapolis, Indiana, USA: John Wiley & Sons, Inc., 1999.

[344] K. Sadgrove, *Complete Guide to Business Risk Management*, 2nd Revised edition. Gower Publishing Ltd, 2005.

[345] J. P. Shedden, *Quantitative Risk- The Way Forward.* Saarbrücken, Germany: VDM Verlag Dr. Mueller, 2007.

[346] J. Van Bon and A. Van Der Veen, *Foundations of IT Service Management Based on ITIL*, New. Van Haren Publishing, 2007.

[347] L. D. Xu, A. M. Tjoa, and S. S. Chaudhry, Eds., *Research and Practical Issues of Enterprise Information Systems II*, vol. 255. Boston, MA: Springer US, 2008.

[348] Office of Government Commerce, *Introduction to the ITIL Service Lifecycle*, 1st Edition. Norwich, United Kingdom: The Stationery Office, 2007.

[349] Amazon Web Services LLC, "Amazon Web Services," *Amazon Web Services*, 2012. [Online]. Available: http://aws.amazon.com/. [Accessed: 30-Oct-2012].

[350] Tagxedo, "Tagxedo - Word Cloud with Styles," *Tagxedo - Word Cloud with Styles*, 2012. [Online]. Available: http://www.tagxedo.com/. [Accessed: 30-Oct-2012].

[351] Wordle, "Wordle - Beautiful Word Clouds," *Wordle - Beautiful Word Clouds.* [Online]. Available: http://www.wordle.net/. [Accessed: 30-Oct-2012].