Internet Tomography
Internet Tomography: 
An Introduction to Concepts, Techniques, 
Tools and Applications

By

Abia Moloisane, Ivan Ganchev 
and Máirtín O’Droma

CAMBRIDGE SCHOLARS
PUBLISHING
TABLE OF CONTENTS

Chapter One .................................................................................................................. 1
Introduction
  Introduction to Internet Tomography Measurement Systems (ITMS)
  Internet: Evolution, Operation, Usage Statistics, Problems and Future Services
  Internet Performance Demarcation Perspectives
  Goal and Objectives of the Book

Chapter Two .................................................................................................................. 30
Internet Tomography: IP QoS Metrics, Measurement Methods and Tools
  IP QoS Metrics in Internet Tomography Measurement
  Internet Performance Measurement Methods
  Internet Measurement Tools

Chapter Three ............................................................................................................... 52
Global Scale Internet Tomography Measurement Activities
  Introduction
  International Collaborative Projects
  ITMS Collaborative Projects Going Forward

Chapter Four .................................................................................................................. 67
Service Level Agreements, QoS and Traffic Engineering
  Internet Tomography Measurement in Relation to Service Level Agreements, Objectives and Management
  Internet Tomography Measurement in Relation to QoS and Traffic Engineering Mechanisms

Chapter Five .................................................................................................................. 83
Experimental Implementation and IP QoS Performance Data Analysis
  Collaborative Internet Tomography Measurement
  ITMS Configuration and Measurement Methodology
  Internet Performance Data Analysis - Examples
  Scenario-based Analysis Using Temporal Performance Variation on Critical Paths
  From ITMS to Performance Analysis to SLAs
<table>
<thead>
<tr>
<th>Chapter Six</th>
<th>Internet Tomography in Support of Internet and Network Simulation and Emulation Modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Internet Evolution – QoS Technological Challenges</td>
</tr>
<tr>
<td></td>
<td>Minimally-invasive ITMS</td>
</tr>
<tr>
<td></td>
<td>Network Simulation and Emulation, and Analysis of IP QoS Performance Data</td>
</tr>
<tr>
<td></td>
<td>Empirical Network Modelling and Integrated Network Experimentation</td>
</tr>
<tr>
<td></td>
<td>Use of Internet Tomography Measurements to Enhance Network Simulation and Emulation</td>
</tr>
<tr>
<td>Chapter Seven</td>
<td>Intranet Tomography in Next Generation Wireless Networks</td>
</tr>
<tr>
<td></td>
<td>Internet Tomography in 4G Wireless Networks</td>
</tr>
<tr>
<td></td>
<td>Advanced NPI-ITMS Developments</td>
</tr>
</tbody>
</table>

### Acronyms and Abbreviations

- Acronyms and Abbreviations

### References

- References

### Index

- Index
CHAPTER ONE

INTRODUCTION

Internet tomography is a concept used to describe the science of analysing the performance of the Internet in its various parts or as a whole by probing it through various means from the outside. An Internet Tomography Measurement System (ITMS) is the collection and operation of such means. Introduced from basic principles through to techniques, tools and applications, Internet tomography is the subject of this book. The design of an ITMS aimed at mapping the Internet performance profile spatially and temporally over paths between probing stations is a particular focus. ITMS design criteria addressed include: minimally-invasive or non-invasive, independent and autonomous, active or passive measurement, flexibility and scalability, capability of targeting local, regional and global Internet paths and underlying networks, compliance with the standardised performance methodologies and Quality of Service (QoS) metrics, e.g., those of the Internet Engineering Task Force’s (IETF) Internet Protocol Performance Metrics (IPPM) Working Group.

The book also features:

- The use of Internet tomography measurement in support of network simulation- and emulation modelling for network and service design and analysis, and service deployment.
- The exploration of spatial and temporal Internet performance variations by means of scenario-based analysis using real-time Internet performance data.
- Aspects of Internet tomography in next generation networking (NGN) wireless architectures.
- The role of ITMS in Service Level Agreement (SLA) design, implementation and compliance.

In this chapter the rationale behind Internet tomography and its importance in the task of assessing Internet QoS performance is considered. The chapter includes an introduction of the ITMS and an
overview of the evolution and history of the Internet from ARPANET to NSFNET through to the ever present ‘virtual world’ of today’s Internet, so totally integrated with the ‘real world’ that it now also is ‘real world,’ perceived essential to every walk of human life by those ‘connected’, be that in the personal, professional, business, commercial, economic, governmental, educational, religious, cultural and social domains.

This chapter includes treatment of relevant aspects of the Internet Protocol’s (IP) packets routing dynamics, and a brief review of Internet usage statistics and current problems, and some perspectives on the evolving Internet and future Internet services. A core part of this chapter is setting out the context, reasons and main approaches towards assessing the Internet performance together with a summary of their advantages and disadvantages. In this way, the initial argument for Internet tomography is first set out, along with the goals and objectives adopted for this book to illustrate by examples its strengths, benefits and limitations.

**Introduction to Internet Tomography Measurement Systems (ITMS)**

The Internet has had—and is continuing to have—a profound impact on society, e.g., [1]. As pointed in [2],

not only has its emergence sparked a rapid transformation of the business world but it has also enabled vastly different business models to evolve and thrive.

However, as a complex communications network, the Internet has yet much evolving and developing to do especially to realise its true technological potential, for instance in respect of its operating efficiency, of its QoS performance and of its traffic engineering susceptibility. The exponential growth in traffic and services, characteristic of the life of the Internet, has created a continuous demand for new solutions to network delays, outages, network capacity, traffic overload, congestion, and bottlenecks that all make their presence felt through variable, limited and poor QoS to end-users. Especially the variability of these problems is complex and multi-factorial. Examples of factors, influences and dependencies of any assessment include time of the day, day of the week, and time of the year; geographic location, time-zones, connectivity and density of end-users; maintenance, quality and redundancy of composition networks, sub-networks, network back-haul and backbone elements; and so forth. Add to this an immense variability in user demands and unpredictability of
bandwidth demands, and success or otherwise, of new services.

For small and large end-users, the when, where and for what durations perceived QoS compromises occur already is a complex dynamic problem, and one that will of its nature continue to grow in complexity. And yet, in brief, for network operators, Internet Service Providers (ISPs), and other major integral service providers, such as Cloud Computing (CC) and Data Centre (DC) providers, as well as major commercial service providers such as online air travel ticket sales, all quite apart from end-users, it is of key importance both to know and predict network performance on a day-to-day basis. ISPs for instance need such knowledge in order to meet, or assess risks in meeting, obligations set out in service level agreements (SLAs). Among some initiatives to address the QoS issues are proposals [3-5] for establishing globally-acceptable standard QoS figures of merit (FOMs) and Internet tomography measurement means and approaches to capture performance data to enable QoS assessment and QoS FOM calculations. Dynamic and real-time capturing of performance data from an end-to-end (E2E) perspective provides a means for service providers or carriers to assess QoS on their networks, or on key parts of their networks, and adopt proactive modes of operation in addressing QoS issues/problems. Seeking to gather such information directly from ‘in house’ measurement statistics of network domain proprietors, for instance of the ISPs, raises security and privacy concerns, and understandably would not be welcomed or readily facilitated by the proprietors.

The motivation for having a non-invasive or minimally-invasive Internet tomography performance measurement technique derives from this consideration. If such were possible, the service it could provide in capturing Internet performance data independently of Internet infrastructure and services proprietors and providers can readily be appreciated. A non-invasive measurement system would have the goal of delivering long-term and short-term (even continuous), comprehensive QoS assessments and measurements of IP performance in the core, across the edge and even extending into customer sites. It would be an important consideration that it itself would not impact negatively on normal user traffic, or violate security and privacy concerns in the process [3, 6]. Systems may be implemented in a wide variety of organised and ‘unorganised’ ways. An example is to monitor what the end-user quality of experience (QoE) is at different locations geographically and, thereby, to construct network statistical performance pictures relevant to the interested parties.

An ITMS based on a minimally-invasive Internet QoS monitoring would typically be characterised by a constellation of automated probing
stations generating and receiving test traffic. These are strategically located and positioned in relation to the Internet areas being assessed and would include both wired and wireless networks sections of these areas. With enough probing stations a general Internet performance tomographic picture (or pictures) may be constructed of the Internet regions being probed, and the QoS measurement picture along certain geographic routes as a function of time, day, month, etc. may be constructed.

Internet: Evolution, Operation, Usage Statistics, Problems and Future Services

In creating an understanding of the problem, we will first briefly review how the Internet evolved and how it operates. We will treat some of its essential core protocols and examine some of the inherent operational inefficiencies. Typical relevant pertinent questions which arise are:

- Why the disorderly fashion of Internet evolution [2]?
- Could current performance shortfalls have been anticipated and averted with a different Internet design strategy [2]?
- Is the exponential growth so characteristic of the Internet set to continue?
- Will the Internet continue to scale stably and securely in the face of new traffic types, new Internet services, predictable and unpredictable traffic-service growths, and the trend towards bringing about an all-IP convergence of all networks and communication services?
- In respect of control maintenance, and improvement of QoS performance, QoS stability and monitoring, is there a role for a dedicated approach of Internet tomography measurement and traffic analysis?

A look at the history of the Internet and its open architecture should provide some insights and answers to these questions, e.g., [2].

Historical Context

The year 2011 saw approximately one third of the world’s 7-billion population ‘on line’ [1]. This revolutionary presence and impact of the Internet was far from the minds of its originators. It began and evolved in the USA as the 1960s Advanced Research Projects Agency Network (ARPANET), funded by the USA’s Department of Defence (DoD). It was here that the all important Transmission Control Protocol / Internet
Protocol (TCP/IP) suite was developed and realised for the first time. The priority given to survivability and robustness, e.g., that a route from source to destination will be found if it exists at all, is testament to the DoD influence. Initial pilot trials based on a few universities and ARPA were successful, and were expanded in to larger field trials, and these transitioned into live data-networking service. Seeing its benefit for fostering collaboration, effectiveness and efficiency for research work in all fields and disciplines, the project was adopted by the USA’s National Science Foundation Network (NSFNET) with a view to its deployment throughout the university and higher education sector. From there, it quickly became international and spread throughout the western world. Almost just as quickly and in parallel, it began to realise its commercial potential, rapidly developing into the commercial Internet that now exists and with which we are all familiar.

At the beginning, ARPA experimented with linking multiple computers to a single network for purpose of basic data transfer. Data communications, as the discipline became widely known by, was just data, i.e., text-based information, which was ASCII or EBCDIC encoded being carried in packets and routed between different nodes in the networks from source to destination. Transparency techniques were employed to allow transmission of non-text data, such as executable computer programme code. TCP/IP of course has evolved a great deal in terms of information payloads of the packets, but at the birth of TCP/IP other communication services (sometimes referred to as ‘media’) such as telephony, commercial radio and television, i.e., were all in existence but had their own dedicated telecommunication networks matched to specific telecommunications requirements of these services. Only in this present era of the first decades of this third millennium is there a real resemblance of true convergence with these diverse communication services being integrated into the Internet, i.e., into a ‘multimedia’ Internet where all services are carried in streams of IP packets within the network and by some transmission control protocol forms (TCP, UDP, etc.) external to the network. However, that first 1960s network consisted of three computers, one at ARPA, a second at the Massachusetts Institute of Technology (MIT) and the third at the University of California [7]. Hence, as such, ARPANET provided a limited number of researchers with shared, interactive communications between computing systems at different locations and developed key innovative technologies and technological concepts which characterise the Internet such as:
• ‘Packet switching’ in which a message is divided into packets that are transmitted to their destination and reassembled in a connectionless manner – rather than using a dedicated circuit as in the telephone system or ‘store and forward’ that is used in the telegraph system.

• A distributed rather than a centralised network – even when some intermediate nodes (routers, gateways) fail, the remaining network continues to function, with protocols and algorithms to enable updated information about functioning routes (and routers) to filter into the routing tables of all routers. As indicated, this high-level survivability and robustness design goal was a most vital component of the DoD’s design plan. In effect in a context of chaotic network disruption, if a path exists between the source and destination, it will be found by the Internet routing protocols, either by brute force routing, e.g., through flood-search routing, or through more refined dynamic updating of routing tables as in today’s Internet.

• Adaptivity – packets can be transmitted to the same destination over different parts of the network based on current network load and connectivity. However, this was more a rough congestion solution rather than anything nearing a network traffic engineering philosophy.

Early success inspired ARPA to connect even larger groups of computers using existing telephone lines and the public switched telephone network (PSTN) as the wide area networking (WAN) medium. ARPA’s Communications Group, an early think-tank, designed the basic architecture to construct such a system. Prospective suppliers were invited to develop a rudimentary computing network using this architecture and the first true network was established in 1970. By 1971 the network consisted of 15 sites and 23 host computers [2].

The success of the initial experimental system created its own demand for a stable network especially as the user community began to grow. In 1984, the Military Network (MILNET) was split off from the ARPANET for use by the USA DoD users. USA Federal agencies established networks to support their research and development (R&D) communities. In 1986, the USA’s National Science Foundation (NSF) created and launched NSFNET for the nationwide university research community, NASA built their own NASA Science Internet (NSI), and Department of Energy (DoE) created the Energy Sciences Network (ESnet). These networks and an ever-growing profusion of local area networks (LANs)
and regional networks form the High Performance Computing and Communications (HPCC) component of the Internet, a ‘network of networks’, and demonstrated the fundamental strengths of robustness and scaling attributes of new networking technologies created in the early ARPANET. Bolstered by advances in data transfer technology, NSFNET facilitated the distribution of growing amounts of data and stirred demand for a commercially-oriented network [2]. From about 1988 ISPs emerged and became inter-connected over an emerging international Internet backbone network constructed over the existing PSTN back-haul, trunking and leased-line networking resources. These interconnected LANs users and computing resources as well as the opening of access to everyone at the end of a telephone line through a computer equipped with a modem. Global computer networking and data communications between interconnected computer users on a grand scale thus came into existence. Today the Internet is seen as the biggest success story of the 1990s [2] and fundamental to about creation of the information age. As an economic driver and contributor in itself, it would not be amiss to put the annual investment into the technological infrastructure at billions of dollars, with the annual investment into the services offered over the Internet at several times this.

The phenomenal growth of the Internet beyond the HPCC community is the result of educational requirements and services, public and governmental service needs, the quickly growing realisation of potential benefits in industry, the commercial, advertisement and entertainment world—especially with drivers from productivity to business globalization—and in the private, personal and social networking sector. The HPCC program in itself directly stimulated the emergence of a vigorous and highly competitive private sector industry in Internet hardware, software, and connectivity in which the USA, the birthplace of the Internet, was and probably still is the world leader. The authors in [2] identify three historical technological innovations (introduced in the 1980s) that had a profound impact on the development and exploitation of the embryonic TCP/IP Internet in this start-up phase, namely: the IBM personal computer (PC), the LAN, and the UNIX operating system. Before the introduction of the PC, existing computer networks consisted of large specialised mainframe computer systems that functioned in protected environments. The PC immediately made computing more accessible [2]. LANs initially were mostly proprietary in topological design, physical layer protocol and infrastructural solutions until the open Ethernet solution began to dominate (driven largely by USA government policy, mainly through DARPA). Ethernet technology evolved through several generations and changed the
face of local area networks by ever-increasing speeds and quality of data transmission. Ultimately the introduction of UNIX, with its multiple hardware capabilities, allowed users to break away from vendor-specific technology environments, thus creating greater choice between a growing range of hardware and software applications [2].

At the technological heart of the success of the Internet is the connectionless nature of IP and the (then) largely delay insensitive nature of data communication services. While the digitisation of the existing telephony-based global telecommunications network played a very important role in that completely autonomous network and was done for the benefits and efficiencies it brought to telephony, data over the telephone network, i.e., WAN, was in many ways transparent to the network operators, being perceived as simply a telephone call, and commercially insignificant when compared to the volume of actual voice telephone calls. It is noteworthy that the Internet drive as such, i.e., the drive to interconnect LANs over WANs, was coming from the computer community wanting fully networked ‘data communication’ services, and it was engaged in an uphill battle against the commercial realities as perceived by the telephone network providers. In fact it is not unreasonable to say the needs, demands and service requirements of the ‘Internet-driving community’ not only took the traditional telecommunications engineering community by surprise but it took quite some time for them to respond in ways and on a scale which would truly correspond to the ‘what and how’ of the mushrooming Internet.

Telecommunications QoS and traffic engineering were strengths telephone network providers were rightly proud of in their global PSTN. However, achieving this as they were on connection-oriented real-time (i.e., telephony mainly) network services was a completely different engineering challenge to the form of the problem which was then posed by ‘data communications’ community with its connectionless IP layer. The idea that TCP/IP in itself was robust enough, in regard to its early stage QoS requirements, or that the QoS performance of the underlying network, in this case core resources from the PSTN, was not a matter of major concern to these new type of PSTN users, i.e., to the data communications’ community, was a paradigm shift in traditional telecommunications network engineering thinking and priorities. That the TCP/IP data communications community was not showing much interest in the PSTN engineers’ first ‘data communications’ response—of connection-oriented packet-switched public data network (PSPDN) based as it was then on X.25, with its strong data link layer (DLL) ‘error-free’ transmission control, except that it might be just another networking
option over which IP packets may pass—was also disconcerting.

In fact how IP packets were handled by the underlying network in getting those packets from router to router was not a matter the TCP/IP community considered directly essential to the creation of the TCP/IP Internet; those networks (or combinations of interworking networks), which ended up doing the packet transmission and delivering best, simply became the routes of choice in the router tables for all IP packets. This philosophy was simple and robust, but inherently was resistant to traffic engineering [2, 8]. The latter’s absence could well be considered the TCP/IP Internet’s greatest flaw, especially as it scaled globally and in traffic volumes. In this new environment, new approaches to both QoS performance and traffic engineering would need to be devised.

**Internet Operation**

Knowledge of the basic workings of the Internet helps appreciate the role and place of Internet tomography. The Internet today is organised as an interconnection of Autonomous Systems (AS). Any one of these consists of a collection of internet-worked routers managed and administered by a single authority or organisation. An abbreviated description on Internet traffic exchange follows. Messages greater than a certain size determined by the maximum transmission unit (MTU) allowed by the access network/link of the sending host, are broken into suitable TCP segment lengths and then encapsulated into IP packets for transmission over the Internet. In IP version 4 (IPv4), the maximum packet size limit is set to 64 kbytes, though the MTU is usually much smaller than this. The MTU chosen may be fixed or influenced by a standard (e.g., in Ethernet the MTU is about 1500 bytes) or decided at connection time (e.g., in point-to-point serial links). Larger MTUs result in higher bandwidth utilisation and efficiency. In IP version 6 (IPv6) the maximum-size limit allows transmission of what have been called ‘jumbograms’. As individual IP networks and Internet backbones, and the physical network infrastructure of the Internet in general evolved—an evolution which continues—in terms of bandwidth resources, all-IP infrastructure, and transmission QoS (especially error rates and associated dropped packet rates), the MTU will converge towards the maximum IP packet size for reasons of efficiency. Each IP packet consists of a header, containing control information, destination- and source addresses, and a payload—the service data unit (SDU) in ISO terminology—carrying the higher-layer data or ‘protocol data unit’. Examples are TCP segments or UDP datagrams. IP packets produced from a single long message, which has
been segmented into several TCP segments with one segment per IP packet, may travel different routes to the destination and thus may arrive out of order. At the destination, the TCP segments are extracted from the IP packets, their order is restored using their sequence numbers, and the reassembled original message is delivered up the protocol-stack higher layers (if such are present) to the relevant application. At the Internet edge, for insertion into, and delivery from, the Internet, the IP packets are sent via an access network to, and from, an ISP. The ISP, which provides routing capability, is connected to an Internet backbone and routes the packets out onto the Internet and receives packets intended for its own hosts and end systems. Once on the Internet, each packet travels from network to network, i.e., via gateway routers in and out of networks and via routers through each network. Routing in fact is the principle Internet service provided to ISPs. The Internet uses a hop-by-hop routing model, meaning that each router examines the destination address in the IP packet header, identifies the next hop (next router) along the path to the destination, and forwards the packet to it. For this to work, the routing tables of routers must have sufficient up-to-date information to send the packet onto a ‘next router’ on the way to the destination, or to the destination itself if it happens to be an end-point on this router’s network. These routing tables are updated, distributively and autonomously, based on the exchange of routing information with other (neighbouring) routers. As a consequence update information on the status of routers and links percolates or ripples through the network, leading to the appearance of an ever dynamic Internet topology. This phenomenon from another perspective may be considered as instability in routing table entries. Such instability could lead to packets circulating, having their ‘time-to-live’ (TLL) counters run down before the destination is reached and thus being dropped. A variety of distributive-control routing schemes and routing protocols are used to maintain routing information and compute paths, both within the confines of an AS, Intra-AS routing protocols, and between adjacent AS’s, Inter-AS routing protocols; c.f. [9-15]. In one way or another the goal is for routers to exchange routing information messages with one another so that they can keep their own routing tables up to date according to a best-route’ criterion for choosing routes to all destination networks. Examples of ‘best-route’ are cheapest, most efficient bandwidth-wise, shortest time or shortest distance. In this way routers are in a position to properly direct IP traffic on a hop-by-hop basis from one AS to another, and general ensure that each packet inserted reaches its final AS, final router and destination, and does this robustly if there is a route at all to that destination and the packet’s TLL value is sufficient. The
Inter-AS routing protocols are typically policy-based, meaning that the ‘best’ route selection is based on how IT management and network operators of each AS set their routing policies (e.g., whether AS agree/disagree to carry transit packets for specific AS’s), whereby routers choose their preferred routes to all destinations. The schematic in Figure 1-1 seeks to illustrate an example of a typical TCP/IP message processing concept encountered in the Internet, where a message is segmented at the source into IP packets, A and B, which then travel completely different routes to the destination.

![Fig. 1-1. Schematic of typical message TCP/IP processing over the Internet.](image)

**Internet Usage Statistics and Economic Impact**

High-speed networks offering broadband connections to users are today core to the basic and strategic infrastructure of most countries. Recognised as a key technological component in the foundations of the knowledge economy, they offer a unique, cost-effective opportunity to enhance competitiveness and deepen, and make pervasive, the involvement of increased numbers of the world’s population in the global economy. For many under-developed and isolated countries, the Internet mediated through such connections provides an opportunity to rise above
physical and geographic constraints [16]. By 2011 one third of the world’s population were connected to the Internet. As an indication of growth some 45% of those connected are below the age of 25. Some 10% have active fixed broadband access and 20% have active mobile broadband access, [1, 17]. Mobile telephony, and its network infrastructure through which wireless connection to the Internet is accomplished (i.e., going way beyond telephony services), is greatly accelerating the pace of growth and pervasiveness of Internet usage. In itself alone, mobile telephony is seen as having significant impact on economic growth and poverty reduction [16]. Today the number of mobile phones far outweighs the number of fixed lines. With the onset of smartphones and wireless broadband access to the Internet, millions of new and existing Internet users worldwide now also connect to the Internet using these mobile broadband technologies [17]. With greater openness and portability in mobile telecommunications, ubiquitous Internet access offers connectivity that follows users seamlessly as they move from place to place and from device to device [16].

The Internet usage statistics and users growth presented in Table 1-1 and Figure 1-2 provide a picture of how the Internet has grown and continues to do so overall and in various world regions. Statistics presented for fixed broadband market in Figures 1-3 and 1-4 include world regions and top 20 countries in terms of broadband subscription and usage. Figures 1-4 and 1-5 present statistics for mobile broadband subscriptions in the world and compared to other information and communication technology (ICT) developments.

**Current Problems**

Internet attributes of open architecture, network of multiple independent autonomous networks and sub-networks infrastructure, the adoption of new technology and the creation of multiple independent agreements among network service providers in relation to commercial collaborations on local, national, regional and international bases, do not lend themselves to central planning. Rather Internet growth and development are driven by the initiative, entrepreneurial spirit and commercial/economic opportunity, user demands—especially networks of users such as ‘social networks’—and the technological ingenuity of users, network providers, service providers and content providers.
Table 1-1. World regions’ Internet usage penetration (adapted from [18]).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>347.4</td>
<td>108.1</td>
<td>273.1</td>
<td>78.6</td>
<td>152.6</td>
</tr>
<tr>
<td>Oceania / Australia</td>
<td>35.4</td>
<td>7.6</td>
<td>23.9</td>
<td>67.5</td>
<td>214.0</td>
</tr>
<tr>
<td>Europe</td>
<td>816.4</td>
<td>105.1</td>
<td>500.7</td>
<td>61.3</td>
<td>376.4</td>
</tr>
<tr>
<td>Latin America / Caribbean</td>
<td>597.3</td>
<td>18.1</td>
<td>235.8</td>
<td>39.5</td>
<td>1205.1</td>
</tr>
<tr>
<td>Middle East</td>
<td>216.3</td>
<td>3.3</td>
<td>77.0</td>
<td>35.6</td>
<td>2244.8</td>
</tr>
<tr>
<td>Asia</td>
<td>3879.7</td>
<td>114.3</td>
<td>1016.8</td>
<td>26.2</td>
<td>789.6</td>
</tr>
<tr>
<td>Africa</td>
<td>1037.5</td>
<td>4.5</td>
<td>139.9</td>
<td>13.5</td>
<td>2988.4</td>
</tr>
<tr>
<td>WORLD TOTAL</td>
<td>6930.1</td>
<td>361.0</td>
<td>2267.2</td>
<td>32.7</td>
<td>528.1</td>
</tr>
</tbody>
</table>

The development in recent years of cloud computing, made feasible through the growth of broadband, is set to drive the definition of broadband to ever higher bandwidths with significant consequences for network infrastructure provision. Hence the unpredictable and dynamic nature, both in time and Internet location, of congestion and such problems will continue, but perhaps its order of magnitude and its rate of variable will grow. QoS performance, which will naturally grow in importance in the minds of all stakeholders with the increase pervasiveness and penetration of the Internet from population to the range of services being provided, including ever more key and vital services, will be affected with variable local, regional and global significance. It’s not unusual to find specialist commentators indicating that latency attributable to congestion and other forms of sub-optimal performance are the major obstacles to the
Internet realising its full commercial and economic potential at any given time. Such uncertainties can hinder the growth of e-commerce. The variability of the QoE for ISP’s corporate and residential users can lead to disenchantment, frustrations and disagreements all of which can slow large enterprises from extending benefits of high-speed IP services on the one hand or entering the Internet service market on the other.

Fig. 1-2. Internet users growth graph [18].

Time-of-day and day-of-week problems with response times are often the first indication to users, network providers and/or service providers of a performance problem in the infrastructure [20, 21]. User perspective studies on fixed access [2, 180] and mobile access [181] of the web indicate that user attention spans diminish when response time goes beyond two seconds, as shown in Fig 1-6. The tolerance latency tipping point of web browsers waiting for a web page to load of 7 to 8 seconds on fixed access in 2000, [2], has transformed over the decade, to 2009, to rapid abandonment already happening between 2 to 4 seconds for both fixed- and mobile access. The commercial perspective here, such as lost business transactions due to this latency, would clearly be for a web page response time to be less than 4 seconds. Or even better--as, for instance, an
extrapolation of the more recent data would indicate—it should be less than 3, and if possible less than 2 seconds. This sets a clear upper bound requirement on Internet bandwidth and performance. It is not difficult to appreciate a desire to see such performance requirement reflected in service level agreements. The value of the response time generally depends on the performance of the IP networks along the route between client and server, or the many servers which could be involved in serving a web page request. Especially of importance, here will be the performances of the various routers/gateways along each path, and the link-by-link bandwidths, and server processing times. However, other issues could play major roles at different points of the networks such as performance of routing algorithms, update periods for router tables, local- and wide area stability of network routes especially in periods of temporary disruption of major routers/gateways, Internet backbone links or sub-networks, cost formulae used in distributed routing algorithms, absence or presence of alternate back-up routes in router tables, and so forth.

Fig. 1-3. Fixed broadband subscriptions, 2000-2010 [19].

As a transaction abandonment in the business and commercial world could mean a business transaction lost with consequent loss of income, the Internet QoS requirements of business and commercial users—such as response time—will be of great importance. While there is of course a multi-factor variability in the user-abandonment and user-tolerance figures related to individual users, user expectations, user experience, the
importance of the transaction to the user, and so forth, nonetheless it is clear that those whose businesses rely on these e-transactions will want well-defined SLAs from network providers, together with compliance auditing through agreed objective performance measurement extraction in relation to the QoS aspects of these agreements.

Table 1-2. Broadband usage of top 20 countries (adapted from [18]).

<table>
<thead>
<tr>
<th>Country</th>
<th>Broadband Subscribers (million)</th>
<th>Broadband Usage (%)</th>
<th>Population in 2007 (million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>5.4</td>
<td>32.8</td>
<td>16.4</td>
</tr>
<tr>
<td>Korea, South</td>
<td>14.0</td>
<td>27.4</td>
<td>51.3</td>
</tr>
<tr>
<td>Sweden</td>
<td>2.5</td>
<td>27.2</td>
<td>9.1</td>
</tr>
<tr>
<td>Canada</td>
<td>7.7</td>
<td>23.7</td>
<td>32.4</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>14.0</td>
<td>23.1</td>
<td>60.4</td>
</tr>
<tr>
<td>France</td>
<td>13.7</td>
<td>22.3</td>
<td>61.4</td>
</tr>
<tr>
<td>United States</td>
<td>66.2</td>
<td>21.9</td>
<td>302.0</td>
</tr>
<tr>
<td>Germany</td>
<td>17.5</td>
<td>21.2</td>
<td>82.5</td>
</tr>
<tr>
<td>Japan</td>
<td>27.2</td>
<td>21.1</td>
<td>128.6</td>
</tr>
<tr>
<td>Taiwan</td>
<td>4.5</td>
<td>19.6</td>
<td>23.0</td>
</tr>
<tr>
<td>Australia</td>
<td>3.9</td>
<td>18.8</td>
<td>21.0</td>
</tr>
<tr>
<td>Spain</td>
<td>7.5</td>
<td>16.7</td>
<td>45.0</td>
</tr>
<tr>
<td>Italy</td>
<td>9.4</td>
<td>15.8</td>
<td>59.5</td>
</tr>
<tr>
<td>Poland</td>
<td>2.6</td>
<td>6.9</td>
<td>38.1</td>
</tr>
<tr>
<td>Turkey</td>
<td>3.6</td>
<td>4.8</td>
<td>75.9</td>
</tr>
<tr>
<td>China</td>
<td>48.5</td>
<td>3.7</td>
<td>1317.4</td>
</tr>
<tr>
<td>Mexico</td>
<td>3.7</td>
<td>3.5</td>
<td>106.5</td>
</tr>
<tr>
<td>Brazil</td>
<td>6.4</td>
<td>3.4</td>
<td>186.8</td>
</tr>
<tr>
<td>Russia</td>
<td>2.9</td>
<td>2.0</td>
<td>143.4</td>
</tr>
<tr>
<td>India</td>
<td>2.5</td>
<td>0.2</td>
<td>1129.7</td>
</tr>
<tr>
<td><strong>TOP 20 Countries</strong></td>
<td><strong>268.2</strong></td>
<td><strong>6.9</strong></td>
<td><strong>3890.4</strong></td>
</tr>
<tr>
<td><strong>Rest of the World</strong></td>
<td><strong>36.3</strong></td>
<td><strong>1.4</strong></td>
<td><strong>2684.3</strong></td>
</tr>
<tr>
<td><strong>Total World Subscribers</strong></td>
<td><strong>304.5</strong></td>
<td><strong>4.6</strong></td>
<td><strong>6574.7</strong></td>
</tr>
</tbody>
</table>
Fig. 1-4. Mobile broadband subscriptions, 2000-2010 [19].

Fig. 1-5. Global ICT developments, 2000-2010 [19].
Fig. 1-6. Web page abandonment (percentage) due to latency (seconds) in web page response times. Data mined from [2], [180] and [181].

The business and commercial world’s perspective on managing Internet QoS cannot remain on the surface, external to the Internet itself and through SLAs. Appreciating and exploiting some Internet architectural aspects will often be part of the solution. Examples relate to where and how data is stored, mirrored and transferred in bulk or in response to individual client requests. Client requests/replies and transaction communications—realised by an exchange of IP packets travelling long distances—naturally take longer than short distances, and likely will have to pass through a greater number of router hops. IP packets en-route, on encountering any router overwhelmed at that moment by traffic, may be dropped. This in turn will likely lead to retransmission requests from higher-layer protocols, for the dropped packets or the full messages, leading to further delay and increased network traffic and congestion in the same locations. The nearer physically to the client user that the server transaction processing is performed, normally the smaller the number of router hops, and thus the better the response time. However, this intuitive rule of thumb can have large implications for how global businesses, dealing directly with users in a global market, will organise and distribute their servers, databases and cloud computing facilities. Specific service-oriented technological infrastructural design for a business’s Internet service is important, and is increasingly important to network performance as the Internet expands, and grid computing, cloud computing and such
like Internet-based infrastructural services take off. Internet tomography data and statistics can be very helpful when it comes to this kind of service delivery architectural design.

**Future Services**

The Internet conceived in the era of time-sharing and virtual connection-based networks, with basic service goals such as file sharing, remote login and resource sharing, both blossomed in and made the era of personal computers, client-server and peer-to-peer (P2P) computing, and today is driving the era of the network computer [22], grid computing, cloud computing and virtualisation. By 2012 collated statistics, reported in the IET’s Spring 2012 Members News (www.theiet.org/membernews) and in press releases for Facebook’s Initial Public Offering (IPO) in May 2012, on well known consumer services indicated that:

- 2 billion YouTube videos are watched every 24 hours;
- 300 billion e-mails are sent in every 24 hours;
- E-mails represent less than 10% of Internet traffic, whereas web browsing and P2P traffic represent c.45% and c.25%, respectively;
- 845 million users are active in Facebook every month;
- 100+ petabytes Facebook photos and video were stored;
- 90 billion, approx., Google searches were made in 2011;
- 155 million tweets are sent every day.

Such statistics are immediately indicative of size of Internet traffic flow, of the data centre infrastructure supporting this, which is in transition to becoming a key backbone component of cloud computing. It is notable that none of the services above are major commercial services in themselves, i.e., they are more social, and informational rather than business, such as financial institutions or financial markets, where real time demands will also dictate the location of data centres; whether these be sub-millisecond, millisecond, out-of-town, or remote.

In the very short time since e-mail and World Wide Web (WWW) became household names, the Internet has wrought an immense sociological, industrial, educational, business and governmental revolution. It became one of the most visible symbols of a truly globalised world in every domain of life. With its core design completed long before most of our modern telecommunications network architectures and infrastructures, such as the various generations of LANs, metropolitan area networks (MANs) and WANs, including switching and transport technologies such
as Frame Relay, Asynchronous Transfer Mode (ATM), Synchronous Digital Hierarchy (SDH) and many others, and today’s ever growing array of wireless communications technologies including 2G, 2.5G, 3G, 3.5G, 4G cellular systems, WiFi hotspots systems, WiMAX local loops, Femtocells, High Altitude Platforms (HAPs) and satellite systems—the list will continue to grow—it has demonstrated the robustness of its initial ARPANET design principles by not just accommodating this vast array of networking technologies with reasonable ease, but thriving on the addition of this ever expanding base of heterogeneous networks, network services and new users. The Internet really and amazingly has shown its capacity to thrive on such network scaling and multiplying effect, and deliver scaling benefits with the addition of ever greater transport and switching/routing facilities. Today then the Internet is perceived ever more as a service infrastructure, an integrated and essential part of all societal services, such as educational, commercial, governmental, security, and health services, with the global technological infrastructure itself becoming ever more transparent to end-users and teleservices providers, the intrusion of QoS performance problems aside. Notwithstanding the latter, this is an excellent sign of a maturing engineering discipline, where the sophistication of the engineering is hidden, especially from the service consumer.

Engineering challenges remain before the Internet achieves the standing of a fully integrated telecoms service platform. One service in particular, broadcast TV, is still a bit outside the IP embrace. Internet Protocol Television (IPTV), peer-to-peer (P2P) video streaming and downloads are important current steps on the way towards this. Voice services, especially telephony services are well and truly on the road to IP convergence by employing the Voice over Internet Protocol (VoIP) technology together with underlying supportive protocols such as the semi connection-oriented multiprotocol label switching (MPLS) protocol. In fact, in these first decades of the third millennium, VoIP solutions [130, 131] are having a transformative impact on the traditional telephony networks and business. Legacy telephony operator companies have adopted large-scale transition programmes to migrate the fixed telephony network to IP.

With progress being made in IPv6 take-up, an immense range of machine-to-machine (M2M) Internet services, evolving into the ‘Internet of Things’ (IoT), is being envisaged. The network edge, local-loop, infrastructure for these is largely wireless. Hence the ground for optimism in the convergence of all information technology, communications and broadcasting services into an ‘all-IP’ ubiquitous network rightly persists.

This immense success of the Internet does not ignore the technically
demanding challenge of achieving interoperability from hardware infrastructure to software applications, service platforms and service convergence. The inherent lag involved in, and cumbersome nature of, crystallising and agreeing global, regional and national infrastructural, protocol and interoperability standards in such dynamic and rapidly evolving technology and the supporting technological and service standards for this global communications ‘machine’ can result in costly and economically depressing delays. Nonetheless this time lag can open opportunities for innovative and disruptive technologies. TCP/IP itself in its time was such a technology with respect to the slow development, standardisation and take-up of the Open Systems Interconnection (OSI) protocol stack. While it predated the OSI project, its perceived flaws—especially its inherent absence of network traffic engineering (TE) capabilities as traditionally understood—did not enamour it to the legacy telecommunication engineering communities especially Telcos who prided themselves on their network and connection reliability and service-outage statistics. The latter, the development of OSI protocols, which filled many academic and commercial journal papers and texts books of the eighties and early nineties is now rarely mentioned, except for its architectural design template, the OSI 7-layer protocol reference model. While more sophisticated and arguably better designed than TCP/IP under a number of headings, it was nonetheless ‘defeated’ so to speak in the market place. This is an important observation along the lines that for integrated global networking, really only one standard will exist. The observation has important relevance in the context of a successful transition or not from IPv4-based Internet to IPv6-based Internet; a topic which is beyond the scope of this book.

**ITMS and Future Evolution of the Wireless Communications Environment**

It is probably only a matter of time before globally pervasive wireless access to the Internet will be pervasive. As of 2012, the estimate for global mobile connections is in the region of 6.1 billion, approximately 87% penetration, with global smartphone penetration in the region of 10% [177]. The existing wireless access business models will also come under pressure. From the perusal of leading telecommunication journals, it’s clear that further transformations of the wireless communication environment are being conceived and promoted. An example is the IP-based Ubiquitous Consumer Wireless World (UCWW) [23, 24]. Gaining traction, this may change the present network-centric, subscriber-based
business landscape of wireless communications, to a consumer-centric and
network-independent one causing the advent of a whole plethora of new
infrastructural and other IP services and businesses such as consumer-
oriented incoming call connection (ICC) service provision \[174\], third-
party authentication, authorisation and accounting (3P-AAA) \[175\], full
wireless network mobility and number-portability based on new consumer
identity modules (CIM) rather than subscriber identity modules (SIM)
cards, user-driven Always Best Connected and best Served (ABC&S)
\[172\], user-driven integrated heterogeneous networking and commercial
consumer-based Ad Hoc networking. This new kind of IP-based wireless
environment for communications is also supported by a new narrowband
broadcast infrastructure of wireless billboard channels (WBC) \[176\] and
wireless billboard advertising (WBA), which once in place will open the
doors to host other IP-based narrowband broadcast services quite
unconnected with UCWW. What we are indicating here is the clear
potential in the present day for quite significant shifts and changes in the
fixed and wireless Internet infrastructure, operations, services, techno-
business models and marketplace. Thus to build any Internet QoS
performance measurement system based on a particular snapshot of the
Internet’s infrastructural organisation, and especially based on access to
intra-Internet data—from ISP and network provider databases for
instance—may provide good service in the short term but in the long term
will likely always be cumbersome, outmoded, in need of upgrading or
replacement. Internet QoS performance measurement systems based on
probing of the Internet from the outside—Internet tomography concept—
will not suffer from the same temporal degradation in their capacity to
capture performance data, but in fact, the opposite; they will have the
advantage, with more widespread use and improved technology and data
analysis tools, of a generally monotonically improving capacity to capture
Internet performance.

**Real-Time Services QoS Challenge**

Quite seriously demanding in respect of QoS issues are services requiring
real-time transport, e.g., audio and video streaming, and especially so for
interactive real-time (e.g., IP telephony, video conferencing) and this on a
global scale. On the one hand, the ever improving portability,
‘intelligence’ and sophistication of end systems such as netbooks, tablets,
smart mobile phones and other high-end mobile wireless devices create a
natural expectation of high grade QoS on all communications. On the
other, the existence already of QoS benchmarks in users’ experience of the