Saurav Dhungana

Mobile Web Usage: A Network Perspective

Department of Communications and Networking

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, 10.06.2013

Thesis supervisor: Prof. Heikki Hämmäinen

Thesis instructor: Antti Riikonen M.Sc. (Tech.)

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Saurav Dhungana

Title: Mobile Web Usage: A Network Perspective

Date: 10.06.2013 Language: English Number of pages:11+78

School of Electrical Engineering

Department of Communications and Networking

Professorship: Radio Communications Code: S-72

Supervisor: Prof. Heikki Hämmäinen

Instructor: Antti Riikonen M.Sc. (Tech.)

With recent advances in mobile devices and network capabilities, Mobile Internet subscription has caught up to and in some markets even surpassed that of the traditional fixed-line Internet. Hence, in order to sustain future growth and improve their business model, there is a need for the stakeholders to understand the ever evolving Mobile Internet user behaviour.

This thesis analysed data collected from a mobile cellular network in Finland during a week in 2010 using a modified version of the Tstat traffic classifier tool to capture HTTP header and network flow data. Since this was the first time this tool was used for the network measurements, the main aim of the thesis was to test the reliability of the data and then to create an analysis process to build in-depth understanding of the traffic usage patterns. Another goal was also to identify mobile handset devices using the new dataset available.

First, a study of the traffic symmetry and diurnal pattern of the traffic flow was done, which showed downlink dominating the traffic with periods of high traffic during the evening hours. Comparison with the port-based classification showed that the Tstat traffic classifier was more capable in identifying modern Internet applications correctly. The results also found HTTP to be the dominant protocol in Mobile Internet. These information rich HTTP headers enabled detailed study of the HTTP traffic. The Operating System (OS) information available in the *User-Agent* (UA) header validated the fact that most traffic is indeed from PC based devices and thus enabled separate study for mobile handset based traffic. For identifying the handsets, the UA headers were mapped to the WURFL database. From this study, Nokia devices were found to have the highest traffic volume and flows followed by the iOS and Android OS platforms. However, there were lot of malformed and non-standard UAs, which means there is a need to further refine the handset identification methodology.

Keywords: Mobile Internet, Network Measurements, Tstat traffic classifier, WURFL, Device Identification

Acknowledgement

I would like to express my deepest gratitude to all the people who have helped me over the course of this thesis.

First and foremost, I am deeply indebted to my supervisor Prof. Heikki Hämmäinen and Antti Riikonen, my thesis instructor for providing me with the opportunity and resources to conduct this thesis. It is with Prof. Hämmäinen's guidance and feedback that I have been able to successfully carry out the research. Antti has been a great mentor and the guiding light behind this work. Always there to point me in the right direction whenever I found myself in confusion, I am hugely indebted to him.

Special thanks also goes to Takeshi Kitahara for helping me during the early phase of this thesis and contributing greatly to my work. Another important person in this work is Markus Peuhkuri who is responsible for conducting the network measurements and helping with the technical arrangements.

I would also like to thank my colleagues and friends from the Network Business group. They welcomed me into the team with warm hearts and made the working environment fun and motivating. And finally, I would like to thank my parents for supporting me in continuing my studies here in Finland. Its the excellent upbringing and education they gave me that has enabled me to be where I am today.

Otaniemi, 10.06.2013 Saurav Dhungana

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2G	2^{nd} Generation
3G	3^{rd} Generation
3GPP	3^{rd} Generation Partnership Project
GSM	Global System for Mobile Communications
EDGE	Enhanced Data rates for GSM Evolution
UMTS	Universal Mobile Telecommunication System
GPRS	General Packet Radio Service
AAA	Authentication, Authorization, and Accounting
RADIUS	Remote Authentication Dial In User Service
SPI	Shallow Packet Inspection
DPI	Deep Packet Inspection
GGSN	Gateway GPRS Support Node
SGSN	Serving GPRS Support Node
HLR	Home Location Register
VLR	Visitor Location Register
PDP	Packet Data Protocol
MHD	Mobile Hand-held Device
HTTP	Hyper-Text Transfer Protocol
FICORA	Finnish Communications Regulatory Authority
ITU	International Telecommunications Union
ARPU	Average Revenue Per User
TCP/IP	Transmission Control Protocol/Internet Protocol
UDP	User Datagram Protocol
Tstat	TCP STatistic and Analysis Tool
WLAN	Wireless Local Area Network
WiMax	Worldwide Interoperability for Microwave Access

USB	Universal Serial Bus
PC	Personal Computer
ICMP	Internet Control Message Protocol
P2P	Peer-to-Peer
DNS	Domain Name System
FTP	File Transfer Protocol
POP	Post Office Protocol
SMTP	Simple Mail Transfer Protocol
IMAP	Internet Message Access Protocol
IRC	Internet Relay Chat
RTP	Real-time Transport Protocol
RTSP	Real Time Streaming Protocol
RTMP	Real Time Messaging Protocol
RFC	Request For Comments
PCU	Packet Control Unit
WCDMA	Wideband Code Division Multiple Access
RAN	Random Access Networks
QoS	Quality of Service
UE	User Equipment
ME	Mobile Equipment
SIM	Subscriber Identity Module
CN	Core Network
GERAN	GSM EDGE RAN
UTRAN	UMTS Terrestrial RAN
RNC	Radio Network Controllers
APN	Access Point Name
GTP	GPRS Tuneling Protocol

VoIP	Voice over IP
IDS	Intrusion Detection System
IMSI	International Mobile Subscriber Identity
L2TP	Layer 2 Tuneling Protocol
GRE	Generic Routing Encapsulation
PPP	Point-to-Point Protocol
Gn	Interface between the SGSN and the GGSN
Gi	Interface between the GPRS core network GGSN node and the Internet
Gb	Interface between a BSS and a SGSN
IuPS	Interface between the RNC with a SGSN
IETF	Internet Engineering Task Force
URL	Uniform Resource Locator
MNO	Mobile Network Operator
DAG	Data Aquisition and Generation
CAIDA	The Cooperative Association for Internet Data Analysis
HTML	HyperText Markup Language
UA	User-Agent
SSH	Secure Shell
PA Ratio	Peak hour to Average hour Ratio
DSL	Digital subscriber line
ISP	Internet Service Provider
WURFL	Wireless Universal Resource FiLe

1 Introduction

This chapter introduces the basic premise of this thesis. It goes through what the motivation in carrying out this thesis was and what type of research questions it tackles. The objectives are then laid out along with its scope, the research methods used and finally the structure of the thesis is explained.

1.1 Motivation

Mobile telephony and the Internet have emerged as the pre-dominant technologies that drive modern day life in the last couple of decades. They have fundamentally transformed the way in which we communicate and access knowledge, thus becoming an essential part of our daily lives. However, the mobile networks were initially developed primarily to provide voice based services, while the Internet was developed as a wired communication medium. With time, attention has shifted from voice to other data related services in mobile devices. As a result, new technologies were standardised and the speed of transferring information wirelessly rapidly caught up with wired networks, especially in the developed world. Hence, there has been a period of coming together of these two technologies in the last few years.

This phenomenon along with significant developments in the device hardware side has seen a rapid rise in the number of Mobile Internet subscribers. Armed with their smartphones, tablets and laptops, most of these users use mobile broadband as their primary means of accessing the Internet (UMTS Forum 2011). Mobile data has already surpassed voice globally. Most of Internet subscribers already connect to the internet through mobile broadband as the number of Mobile Internet subscribers has outnumbered those with fixed line subscriptions(ITU 2012). According to the market review for the Finnish market (FICORA 2010), the number of Mobile Broadband subscriptions has already surpassed 50% in Finland in 2010.

Today Mobile Internet has become mainstream due to a confluence of several factors. There are salient differences in the ways in which mobile users connect and use the Internet as opposed to traditional Internet users. Mobility and the advancements in the device capabilities are the main drivers of this. The users of Mobile Internet tend to use location based applications which are most often coupled with social networking sites. Multimedia consumption is also a growing trend with online music and video services being increasingly popular. All of this is fuelled by the proliferation of applications (popularly called Apps) from the App Stores.

In developed markets like Finland where the number of mobile subscribers are saturated, companies are always looking at new ways to grow. Hence, it is essential for the different market players to analyse and anticipate the future demand and potential of these services as accurately as possible. They need to be confident that future traffic growth is manageable and their business models ensure that providing these services is profitable. New monetization models are therefore necessary for the Average Revenue Per User (ARPU) to keep pace with the bandwidth demand.

Some academic research has already been done in order to understand the Mobile Internet traffic. A few different approaches have been used in order to carry out these studies (Kivi 2009). Methods such as end-user surveys, panels and handset based monitoring are the most widely used ones and can provide information on individual users/devices. However, these methods only represent a small subset of the user population and so measurements of TCP/IP traffic from wireless network nodes have also been used. This method provides us with information about the actual usage on a highly granular level. The main drawback in using this method is the difficulty in separating individual users (or device) sessions accurately.

This thesis was conducted as part of the MOMI project [1] with the purpose of understanding Mobile Internet usage in Finland. It builds upon previous research done in this project by using new analysis techniques and new measurement tools. Thus, providing deeper insights on mobile Internet service usage. The study was carried out with empirical data collected from network measurements carried out in a Finnish cellular operator. A modified version of a widely used passive network sniffer with an advanced traffic classifier called Tstat (*Tstat homepage* 2010) was used for this. It also captures the Hypertext Transfer Protocol (HTTP) headers, which is the standard protocol for web browsing. Existing knowledge and expertise in network measurement was leveraged to assess Internet usage from this passively observed network traffic. Monitoring passively has the additional advantage of minimal bias as the subject under study usually does not know that he/she is currently being monitored. From these measurements the data was extracted and anonymised immediately. Hence, maintaining end user privacy and confidentiality. Also, since the data is from applications session, there is no directly available user information in the dataset used.

1.2 Problem Definition

The proliferation of Mobile Internet in recent years has created a growing need to understand the usage patterns of Mobile Internet. While there has been some work done in this regard, much is needed to be done in order to develop a more holistic view. The measurements carried out in the networks have been used to classify the

[1] Modelling of Mobile Internet Usage and Business (MOMI) is a national research project (2009-2010) funded by the Finnish Funding Agency for Technology and Innovation (TEKES), Nokia, Elisa, DNA Finland and a number of other companies. The project aims at understanding the ongoing transformation of the mobile industry by using empirical measurements that are implemented in collaboration with Finnish mobile operators and Nokia. The project, along with the empirical measurements, is a continuation of previous national LEAD (2004-2005) and COIN (2006-2007) research projects. The value of the longitudinal analysis conducted in the project increases as more datasets are collected.

traffic and show application usage. Furthering this to incorporate more advanced traffic classification methods promises to provide us with a better understanding of how users are using the Mobile Internet and could be of greater use to the stakeholders. This information helps them anticipate usage patterns and future traffic growths.

The main research questions that this thesis tries to answer are:

1. How can the measurements from Tstat traffic classifier setup be used to carry out Mobile Internet usage studies and how does it compare to previous methods?
2. How can HTTP headers be used to recognise Mobile Handset devices and what kind of insights does it provide into the Mobile Internet market?

1.3 Research Objectives and Scope

In order to tackle the foregoing questions, the following objectives have been set aside.

1. Test the reliability of the measurements obtained from the Tstat traffic classifier tool and verify if it can be used to carry out further analysis.
2. Explore the Tstat dataset and develop an analysis framework to study Mobile Internet usage characteristics from the information available.
3. Create a method to compare the results obtained with the port-based method.
4. Enhance the statistics on mobile Internet usage by taking into account recommendations from previous research.
5. Provide descriptive statistics on Mobile Internet usage in Finland.
6. Develop a method to identify the mobile handset devices and it's key features.

Since this thesis aims to measure and analyse how the "Mobile Internet" is used, it is important to properly define what this term means in our context. The measurement was conducted in a commercial mobile operator in Finland. This means the study is limited to the data traffic of these networks and does not consider other wireless access networks such as WLAN or WiMax. Thus, "Mobile Internet" in this context means all the TCP/IP traffic moving in and out of this network. This is different from the traditional fixed-line "residential" broadband which is not measured for the purpose of this research.

Since, the operator allows users to use their own mobile phones and has mobile broadband plans for laptops through USB/PC cards, this measurement represents quite a diverse device population. For the purpose of this analysis we separated these devices into Personal Computer (PC) devices and handheld devices. PC devices include desktops, laptops and netbooks while handheld devices include pocket devices like mobile phones, smartphones and also tablets like iPad that are gaining in popularity. The term "Mobile Internet" is used to denote traffic from both these types of devices.

Another important consideration is that the measurement was taken in a Finnish mobile operator over a period of several days. Hence, the analysis only represent a portion of the Finnish mobile market and its subscribers. All the generalisations that we make will be with regards to this user base.

It is important to emphasise here that the sessions in this analysis are referred to as *packet data sessions*. There is no distinction made between different *user sessions*.

1.4 Research Methods

The research methods used in the conduction of this thesis were *literature review* and *analysis of mobile TCP/IP network measurements*.

The *literature review* was carried out in order to create an overall understanding of Mobile Internet usage and its related terminology. First, a study of the mobile network and Internet protocols and traffic is done. This enables a review of previous research conducted related to Internet traffic characteristics and mobile Internet user behaviour.

Next, a thorough *analysis of the mobile TCP/IP network measurement* data was conducted. This process went through various stages where the data were preprocessed into a form suitable for analysis. This was then analysed using statistical and visualisation tools such as R/SPSS and Excel to provide tabular output and plots.

1.5 Structure of the Thesis

The remainder of this thesis is structured as follows.

Chapter 2 establishes some background on general Internet traffic and the Mobile Internet and its usage. It explains the different researches that have been used to carry out Internet traffic classification and what were the findings of those studies. Any literature relevant to this thesis is covered by this chapter. Thus, it lays the groundwork on which this thesis will be based.

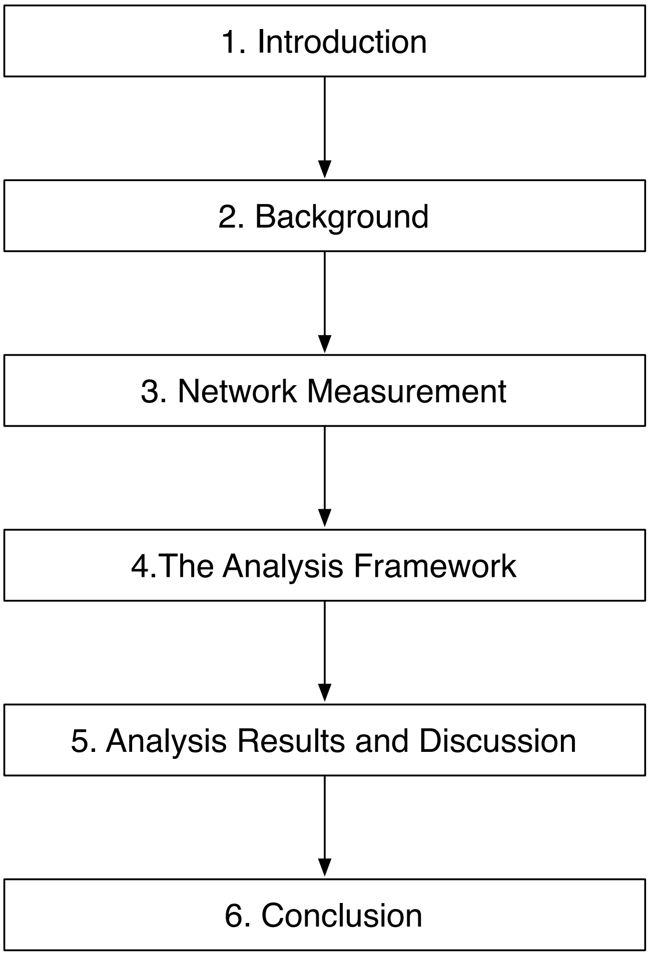


Figure 1: *Structure of the Thesis*

Chapter 3 gives the technical background on how the network measurements was conducted in the Finnish network operator, why these measurement points were used and what type of tools were used for mobile Internet usage study? The datasets used for this thesis and how they are relate to each other are explained. It also delves into the difficulties encountered in the measurement process and privacy issues are addressed.

Chapter 4 establishes the framework upon which the analysis is conducted. The main purpose is to go through the analysis process and explain the various steps carried out to obtain the results. It explains how the traffic classification was done, how web browsing traffic was identified and the data pre-processing methods used. It also looks into the handset identification process and the tools used to do that. Some points regarding the issues encountered in this process and how they were resolved is also made.

Chapter 5 presents the results of the analysis carried out. First, the general characteristics of the traffic is studied and the findings regarding traffic identification from the Tstat outputs shown and diurnal patterns identified. The effectives of this with the old port based analysis method is also given. The Web traffic is then analysed and the findings regarding that are shown and finally statistics on the share of the devices found along with feature analysis based on these identified devices is given. A brief discussion about what the results say about the methodology developed is given after each set of results.

Chapter 6 finally concludes the thesis by summarising the results obtained and going through their implications. It comments on how successful the thesis has been in answering the main research questions made at the beginning of the research phase. Some recommendations about value of carrying out the measurements and analysis to the MOMI project stakeholders are then made and any future improvements that can be done to continue the research conducted in this thesis is given.

2 Background

The main focus of this thesis is to analyse Mobile Internet traffic characteristics. Before delving into that it is important to understand how the Internet works in general and what are its underlying features. Fixed Internet has been around much longer than the Mobile Internet and a lot of literature is available on its traffic characteristics and usage behaviour. This knowledge forms the basis for understanding how Internet is used by mobile users as Mobile Internet is based on mostly the same set of protocols. This chapter first gives an overview of Internet traffic in general and then goes into understanding the state of the art in Mobile Internet.

2.1 The Internet Traffic

The Internet is a worldwide packet-switched data network that combines many different networks. It consists of a complex interconnection of very large number of nodes, each of which can communicate with any other node regardless of the type of network they belong to. Such functionality is managed by defining standardised protocols for various subtasks which are organised into layers. All the user generated data exchanged between applications in the Internet is carried by these protocols. Hence, understanding how these protocols define the architecture of Internet traffic is of great importance.

2.1.1 The Internet Protocols

Unlike traditional telephone networks, in the Internet the end terminals perform all the functions for reliable transfer of data. Hence, the transport network is greatly simplified and the end terminals can add different applications as the Internet evolves. There is one common protocol stack for all Internet related communication called the TCP/IP protocol suite. It was first implemented in the 1970s and is still used today. This implementation organises the Internet into four layers. Each layer has a particular responsibility and provides services to the layer above it. The upper layer makes use of these services and performs its tasks (Crovella & Krishnamurthy 2006). Figure 2 illustrates this architecture.

From Figure 2 it is to be noted that any kind of communication can only be performed among the protocols on the same layer. Thus, an HTTP client on one host can only communicate with an HTTP server on the other end and the TCP level information can only be understood by the TCP layer of the other host. This functionality is realised by adding control data to the packet received from the higher level as headers and conditioning it in a form suitable for transmission over the next

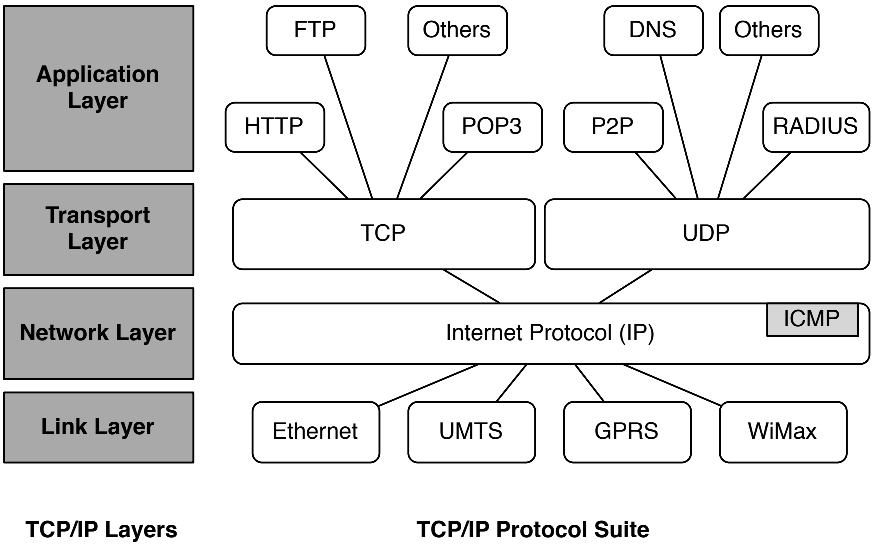
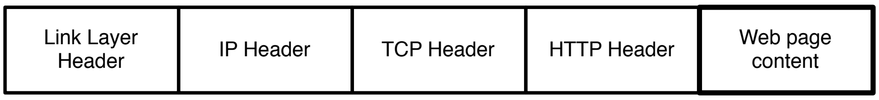


Figure 2: *TCP/IP Protocol Suite (Adapted from Kalden, 2004)*

lower protocol in the sender. This is reversed on the receiving end to get the information related to that particular layer. An illustration of how these packets would appear in a typical HTTP web session is given in Figure 3.



Link Layer Header	IP Header	TCP Header	HTTP Header	Web page content

Figure 3: *Packet structure of a typical web page.*

The **Link Layer** is the lowest layer and its responsibility is to perform physical interfacing with the underlying communication medium. Thus, basically it is responsible for moving the packets between the two hosts. There is no specific protocol defined for this layer in the TCP/IP stack and can be any wired (Ethernet) or radio (WLAN, GPRS, UMTS etc.) based technology.

The **Network Layer** has a single primary protocol defined which is the *Internet Protocol* (IP). It provides the functions necessary to deliver a package of bits from a source to a destination over an interconnected system of networks. There are no mechanisms to augment end-to-end data reliability, flow control, sequencing, or other services commonly found in host-to-host protocols (Postel 1981*a*). The identification of the sending and receiving hosts is done by addresses of fixed length called IP addresses. The most common version is IP version 4 (IPv4) which has 32 bits. A newer version called IPv6 which has 128 bits will eventually replace it.

The **Transport Layer** provides services to applications that allows a flow of data between hosts. It hides the underlying packets from these applications and enables an uniform communication channel to the applications. There are two main protocols in this layer that differ in the type of service they provide to the application layer.

The *Transmission Control Protocol* (TCP) provides means for reliable, connection oriented, end-to-end communication between two hosts. It is reliable in the sense that it ensures correct delivery at the receiver end. This is accomplished by mechanisms such as error recovery, packet reassembly, retransmission, flow and/or congestion control (Postel 1981*b*). It uses port numbers to differentiate each application in the hosts. Developed for wired networks it may have some problems in wireless transmissions. It does not distinguish packet loss caused by network congestion from those caused by transmission errors such as unstable channel, user mobility etc. in the wireless environment. Thus, data rates are reduced leaving radio resources unused. (Tian et al. 2005).

The *User Datagram Protocol* (UDP) is a much simpler protocol than TCP. It sends *datagrams*, which are small chunks of data that fit into a single packet. Unlike TCP it is a connectionless protocol and does not have any provisions for error recovery, flow control and/or congestion control functions (Postel 1980). However, UDP is a lot faster than TCP since it has a largely reduced control overhead and is widely used with applications where timely reception is more important than accurate transmission of data, such as video and audio streaming.

Finally, the **Application Layer** is concerned with implementing the particular applications. This is the protocol that is visible to the Internet users. There are a large number of protocols in this layer that perform different types of Internet Services with the help of the lower layers. The communication between the applications at the two ends is based upon *sockets*, which is a combination of the IP address and TCP ports that identify each application in a host. Some of the most common applications that are the subject of study in most Internet traffic related research and that are of interest to this thesis are.

Web browsing This includes all HTTP traffic excluding video and audio streaming over HTTP (Flash-Video, quicktime, mp3 etc).

Email This includes protocols for sending and receiving email like POP3, SMTP and IMAP.

Streaming This includes every type of streaming in the Internet including streaming over HTTP, RTP, RTSP, RTMP, ShoutCast, etc.

BitTorrent/P2P This includes all the P2P protocols such as gnutella, eDonkey and also the popular Bittorrent protocol.

Chat/IRC This includes messaging services like IRC, MSN, AOL etc.

2.1.2 The RADIUS Protocol

Remote Authentication Dial In User Service (RADIUS) is a client/server protocol used by ISPs and enterprises for authentication and IP address leasing. It acts as an centralised Authentication, Authorisation, and Accounting (AAA) management service that runs in the Application Layer over UDP and provides access to the Internet. This protocol is very important because it is used to identify the same subscriber using different IP addresses over time. Thus, allowing analysis on a per user basis (Varga et al. n.d.). RADIUS protocol is based on UDP destination port 1813.

The Authentication and Authorisation characteristics in RADIUS are described in RC 2865 (Rigney et al. 2000) while Accounting is described by RFC 2866 (Rigney 2000). Whenever a subscriber logs in, the Remote Access Server (RAS) sends a RADIUS Access-Request to the RADIUS server. If the user is successfully authenticated, the RADIUS server returns an Access-Accept message which contains an IP address (Rigney et al. 2000). Next, the RAS uses the RADIUS accounting protocol (RADA) for communicating events that involve data usage to the RADIUS server.

There are different types of accounting message possible within a data session, such as Start, Update and Stop. The Start messages indicate the beginning of a new accounting activity, e.g., when a new application is opened by the user. "Start" records usually consist of the user's identification, network address, point of attachment and a unique session identifier. The Update messages are generated periodically to indicate the status of currently active accounting session. The Stop message indicates the end of the session (Rigney 2000).

The end of a RADIUS session may be caused by either the user logging off or due to a timeout policy that the operators make use of. RADIUS supports two timeouts - *Session Timeout* and *Idle Timeout*. The Session Timeout sets the maximum number of seconds of service to be provided to the user before termination of the session or prompt. The Idle Timeout sets the maximum number of consecutive seconds of idle connection allowed to the user before termination of the session or prompt (Rigney et al. 2000).

2.2 The Mobile Internet

Having gone through the basics of the technology underlying the Internet an overview of data traffic in cellular networks is now given. Cellular mobile technology such as GSM evolved initially by providing voice-only services. But soon a demand for other services in addition to voice start to arise which led to the development of the General packet radio service (GPRS) standard by adding a Packet Control Unit (PCU). It was the first mobile technology focused on providing data-centric services to the end users. Later in early 2000s the Universal Mobile Telecommunications System

(UMTS) standard was introduced that allowed significantly faster data rates and increased the number of available services. Since the GSM/GPRS networks were already widely deployed, UMTS was developed to be compatible with the existing network. So a UMTS unit can also serve GPRS and GSM Radio Access Networks (RAN). The most common form of UMTS uses WCDMA as it radio interface. Let us take a closer look at this network.

In 3G GPRS/UMTS networks users experience high speed data transfer on their mobile devices. Thus, the core networks supporting the users are required to handle large amounts of traffic. Studying these applications helps us understand what network resources are consumed by a user and for what purpose. That would aid charging policies and clever Quality of Service (QoS) mechanisms that would be able to adjust to users' needs so everybody would experience the best from the network.

2.2.1 Mobile Network Architecture

Figure 4 shows the basic architecture and interfaces of a GPRS/UMTS Network. It can be divided into three parts.

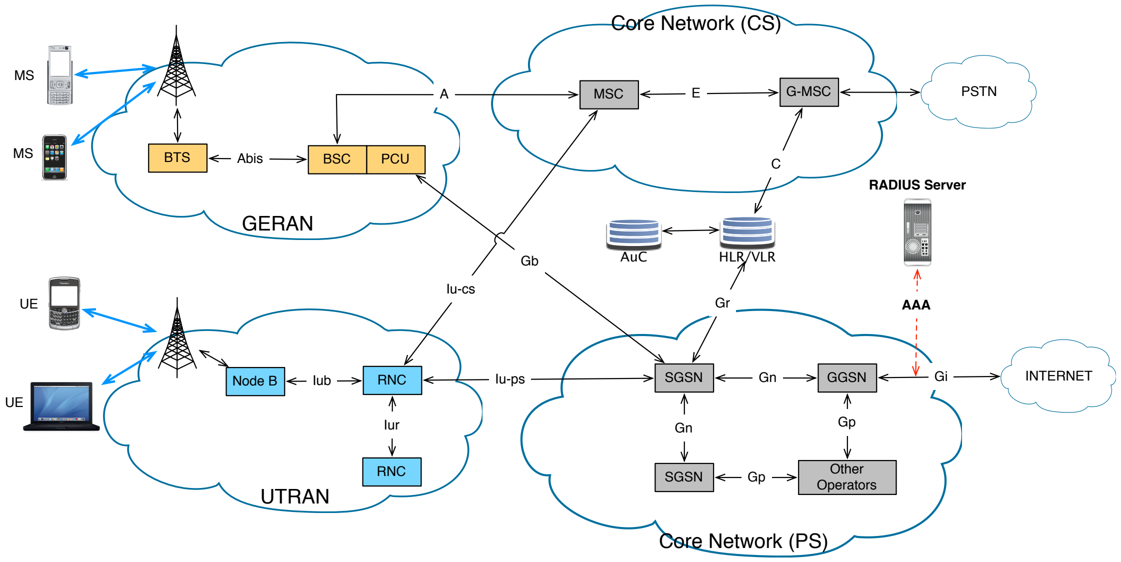


Figure 4: *Basic Architecture of GPRS/UMTS Network. (Adapted from Svoboda 2008)*

The first part is the **User Equipment (UE)**. It is that part of the network with which the user directly interacts. It usually consists of the Mobile Equipment (ME) and the Subscriber Identity Module (SIM), which is called UMTS SIM (USIM). The ME provides the necessary hardware and software to access the UMTS services. The UMTS Subscriber Identity Module (USIM) is a smartcard that holds the subscriber identity, performs authentication algorithms, and stores authentication and encryption keys and some subscription information that is needed at the terminal (Holma & Toskala 2000).

The second part is the **Access Network** which is responsible handling the radio-related functionality for both circuit-switched and IP based packet switched traffic. In GPRS its called GSM EDGE RAN (GERAN) while in UMTS its called the UMTS Terrestrial RAN (UTRAN). The UTRAN contains the base stations, which are called Node Bs, and Radio Network Controllers (RNC). The Node Bs handle channel coding, interleaving, rate adaption, adding scrambling codes and modulation while RNC handles all the radio resource management tasks.

The third part called the **Core Network(CN)** is responsible for switching and routing calls and data connections to external networks. It can be divided into the Circuit Switching (CS) and the Packet Switching (PS) part. Both GERAN and UTRAN can share the Core Network (CN). The main elements of CN related to the PS core are:

The **Home Location Register (HLR)** is a very important part of the CN. It is a database located in the user's home system that stores management data for each user of the mobile operator. The HLR holds all permanent user data such as Mobile Subscriber ISDN Number (MSISDN), available services, QoS, international ID, the IMSI and further temporal data like the location area the ME was last seen.

The **Visitor Location Register (VLR)** is a database that contains data about one particular UE located in a MSC unit and provides Circuit Switched services. It is mainly used to hold information about the visiting subscribers service profile.

The **Serving GPRS Support Node (SGSN)** is responsible for the delivering data to and from the UE within its geographical service area. Its main functions include switching and routing of data traffic, session management, location management, ciphering, cell updates, authentication and billing. SGSN stores location information such as current cell, current VLR and user profiles such as International Mobile Subscriber Identity (IMSI) and the address(es) used in the packet data network.

The **Gateway GPRS Support Node (GGSN)** is the link between the CN and the external packet data network such as Internet, X.25 etc. This communication is carried over the Gi interface. It is responsible for providing mobility to the UE. This is done by keeping a record of the SSGN that a active user originally belongs to and routing the data to this SSGN using the GPRS Tunnelling Protocol (GTP). When it receives GPRS data from the user via SSGN, it converts them into appropriate Packet Data Protocol (PDP) format depending on the PDP-context. It performs the functions of IP Address allocation, Authentication, charging etc. (3GPP 060). As shown in the figure, a RADIUS server can be used for AAA and assigning IP address to the users in each data session. Here, the GGSN can be thought of as the RADIUS client (Kaaranen 2005).

2.2.2 Mobile Internet Traffic

Having introduced the Internet protocols and the general GPRS/UMTS network, a simple model of how data transmission works over such networks is now given. Unlike wired Internet the process of establishing a packet data session for a user in the mobile network requires more steps. The network devices have to take care of tasks like session and mobility management to provide seamless experience to the users. An explanation of how these sessions are established is given first and a hierarchical model of the traffic is then formed based on previous work.

Packet Data Sessions

In general literature, a packet data session is defined as the duration between the time the user is authenticated by the authentication server to the time the user logs off. In GPRS/UMTS networks this is usually a RADIUS Server. There can be any number of application(HTTP, email etc.) running during this user session. The various steps required in this procedure in Figure 5.

The first step is that the UE needs to attach to the GPRS in order to make itself known to the network. Authentication and authorisation of the user is done, along with location update in the HLR and initiation of the mobility management context creation. Only now can the user start its data sessions. The first phase is activation of a PDP-context by the UE. A PDP-context is the set of parameters of all the information required for establishing an end-to-end connection. The SGSN then routes the activation request to the GGSN using the Access Point Name (APN) provided by the service provider. The GGSN then contacts the Radius server to authenticate the subscription and if successful receives a unique IP address corresponding to that context. The PDP-context activation request is the accepted upon which, the SGSN and GGSN communicate via the GTP protocol. The subscriber is now able to transmit user data to external packet data networks over the IP layer. Any number of application sessions can take place within this data session. To end the data session a PDP-context deactivate request is send by the UE. Following similar procedure as defined above the PDP-context is deactivated and the GPRS is detached.

The Traffic Structure

Based on the knowledge of how a data session is established for a user, traffic models of mobile Internet have been made. One of the earlier attempts to model Mobile Internet traffic was carried out by (Kalden 2004) in a GPRS network. Also, (Varga et al. n.d.) performed some structural analysis and modeling of WAP traffic in GPRS networks. Most recently, (Svoboda 2008) has done some work in modelling 3G UMTS traffic in addition to the GPRS traffic. In his work, Svoboda creates traffic models at the application layer of the network model. He emphasises how modelling traffic on the lower layers can be done by using lesser number of parameters than

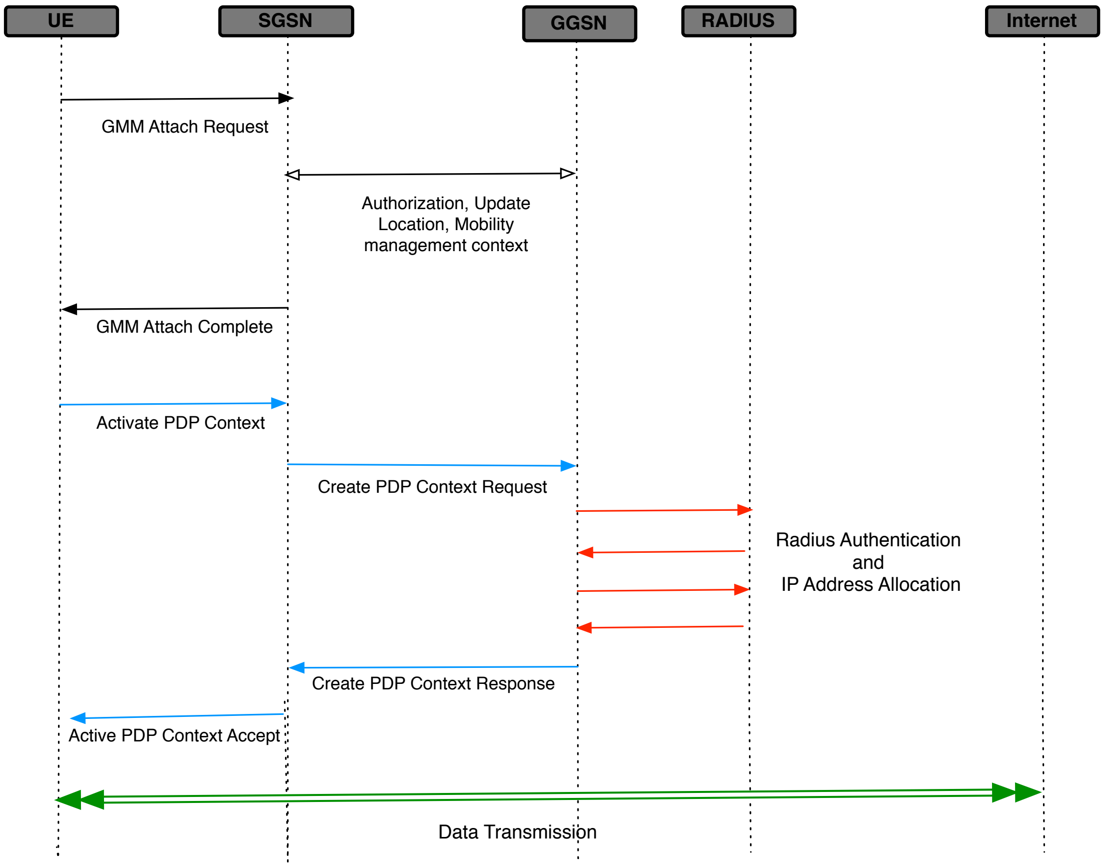

UE
SGSN
GGSN
RADIUS
Internet
GMM Attach Request
Authorization, Update Location, Mobility management context
GMM Attach Complete
Activate PDP Context
Create PDP Context Request
Radius Authentication and IP Address Allocation
Create PDP Context Response
Active PDP Context Accept
Data Transmission

Figure 5: *PDP Context Activation Mechanism.*

the higher ones but cannot really provide us with information on user behaviour. Table 6 provides that structural analysis, which is also used in this thesis.

As seen in the hierarchy the most basic data units consist of packets in the network layer. These are however very small objects and a very large number of packets are send between the hosts in either direction. Obtaining any useful information about application usage is a difficult task. Work done by (Jain & Routhier 2002) found that these packets can be aggregated into "packet trains" that occur together. A common way of defining such packet trains is by separating each by an inter-arrival time that is greater than a defined threshold. A popular way method is to define the 5-tuple, which are streams of packets that have the same source and destination IP addresses, source and destination ports and protocol number as a *flow* (Karagiannis et al. 2004). From this we can identify flows belong to a particular application and perform the analysis on an application level.

If analysis is needed to be performed on a per user basis then it can be further abstracted into PDP-contexts. From the 3G networks perspective, a PDP-context is the basic data unit. Within these contexts there can be several applications running.

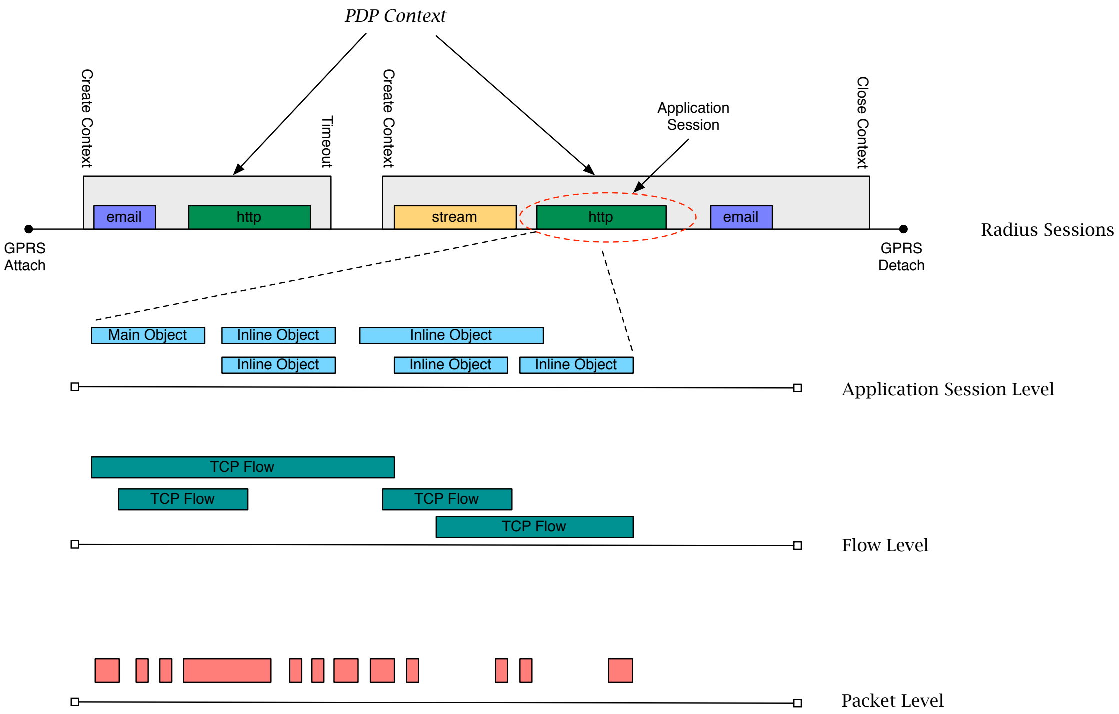


Figure 6: *Traffic structure of different abstraction levels. (Adapted from Svoboda 2008)*

Each PDP-context is over when a users logs off or there is a timeout. Associating the application flows with the PDP-context which they belong to provide a way to study how the users actually use the Mobile Internet.

2.3 Measuring the Internet traffic

In this section an overview of a typical GPRS/UMTS mobile network is given showing the most commonly used measurement points and interfaces and the type of information available from them.

The purpose of Internet traffic measurements is to study various properties of the network traffic like bit rate, packet rate and the protocols used. This enables us to recognise the protocols being used in the application layer. The information gathered and the statistics produced are of utmost importance from a network operator's point of view. It helps them make important decisions such as what kind of upgrade their networks require to what kind of business model and pricing schemes should they apply to improve profits. Accurate traffic classification is of great importance to these operators because it allows them to prioritise certain applications, such as Voice over IP (VoIP), detect and deal with bottlenecks etc. Such statistics can also be part of the security measures in the network by helping to identify traffic patterns which are used in Intrusion Detection Systems (IDS)(Zhu & Zheng 2008).

Another important result that can be given by such analysis is how the users are using these applications. Informed and accurate statistics on user behaviour enables the operators to make better business decisions and offer better services to their customers.

As already explained in the previous chapter, the data being exchanged between the various nodes in these networks are carried in packets. Each packet carries only a small portion of the data being transferred and along with the source and destination addresses. The process of traffic recognition is performed by filtering and capturing the packets that pass through various links in the network. Filtering is the process that selects for only those packets that satisfy the criteria of the measurement, such as only TCP/IP traffic and capturing saves a copy of each of these packets so that it can be analysed later with post-processing and statistical tools. This capturing is done by network analysis tools and the saved files are called packet traces. Certain network tools then analyse the contents of each packet and build flows of traffic that are the communication channels between a source and a destination. The most common application being HTTP, IMAP, POP3, P2P, FTP etc.

Internet traffic analysis in GPRS/UMTS networks is more challenging than traditional networks. The cellular core is different from the Internet in terms of the protocols being used and the requirements for traffic recognition. As an example, in residential broadband Internet the users can be usually identified via the unique IP address of the user. However, in cellular networks the primary user identification parameter is the subscriber identity, International Mobile Subscriber Identity (IMSI) (3GPP TS 23.003) while the IP address allocation is dynamic. The different protocols and interfaces in a mobile network and the information that can be extracted from these is explained now.

2.3.1 Network Traffic Measurements in a Mobile Network

As shown in Figure 7 the GPRS/UMTS core contains many different interfaces that connect the elements of the network. Some interfaces are used for control information while others carry control and user information (data). The interfaces that carry the packet switched data inside the cellular core are the Gn and Gi. The Gn interface is used to connect the network elements of the cellular core while the Gi is a gateway interface to external networks and the Internet.

The GTP on the Gn interface handles the encapsulation of user data in order to keep the core independent of the protocols being used in the endpoints. The data are transferred via tunnels set up by the control stack of GTP. The data are transferred to an external network along the Gi interface. Here, encapsulation using Generic Routing Encapsulation (GRE) or Layer 2 Tunneling Protocol (L2TP) / Point-to-Point Protocol (PPP) is possible or the data may be transferred as plain IP packets. It is operator dependent whether authentication and/or accounting should be applied in the form of RADIUS, AAA servers.

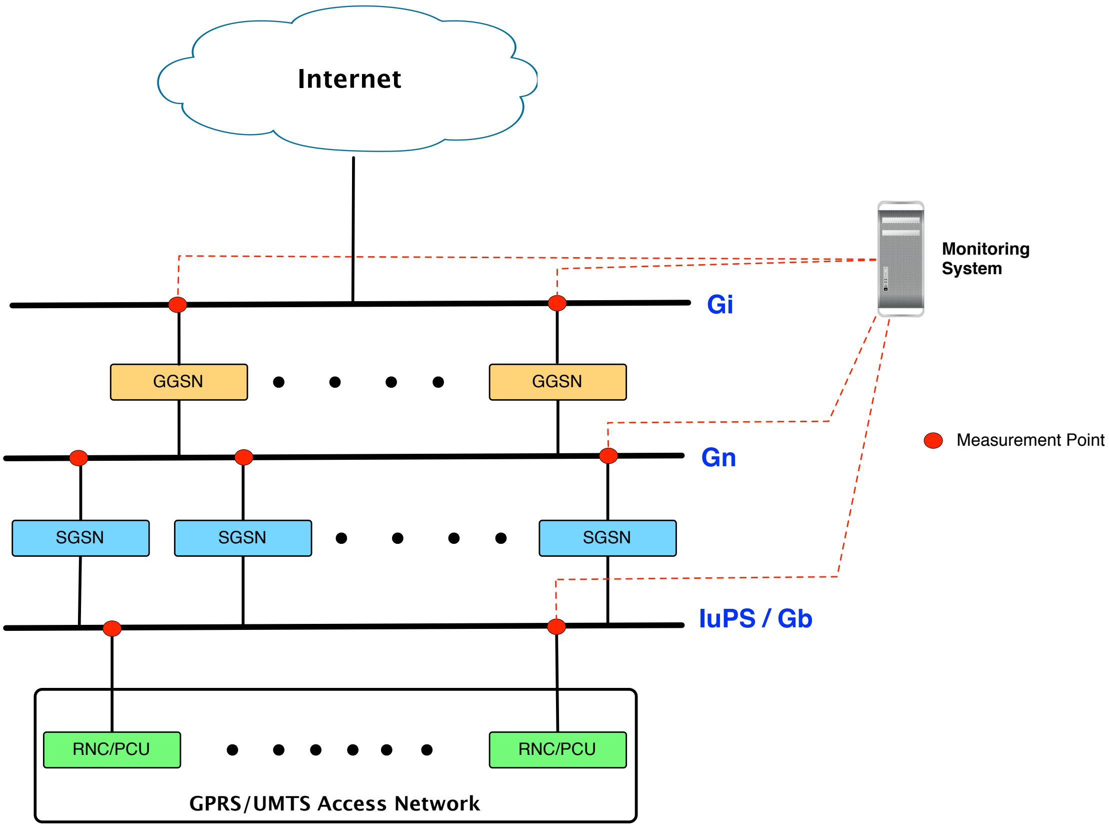


Figure 7: *Common measurement interfaces in a GPRS/UMTS Network.*

The Gn and Gi interfaces are of most importance regarding Internet traffic measurements and in terms of PDP-contexts and are explained here:

The **Gn interface** is located between the SGSN and the GGSN. The PDP-context is used to transfer user data on the Gn interface. It can be seen as two tunnels, one each for the control and data. The context is set up by the Create PDP Context Request message sent by a SGSN to the GGSN handling the APN that the user wishes to connect to. The GGSN should then assign tunnel identifiers for the context, perform authentication, accounting and optionally allocate a network address to the user. It then replies to the SGSN with a Create PDP Context Response message containing information if the request was accepted or not and the tunnel identifiers being set up in the GGSN. If the latter accepted the PDP context then it is now possible for the user to communicate with the APN she selected. Thus, capturing on the Gn interface allows PDP-context analysis. There are separate SGSNs used for UMTS and GPRS. Thus, it is possible to differentiate GPRS and UMTS traffic based on IP address of the SGSN found in the IP header below the GTP layer.

The **Gi interface** is located between the GPRS core network GGSN node and an external packet network such as the Internet. It carries user data but since it is dependent on the network how the user would access the Internet traffic recognition can be obscure. It may be required for authentication, authorization and accounting to be performed on the interface, typically via a RADIUS server with the GGSN acting as the client Network Access Server(NAS). The protocols used to communicate to the RADIUS servers are the ones that carry the control information on the Gi interface. If we ignore these protocols then the Gi interface resembles an Internet backbone connection link with the addition of tunneling. It is common for Radius Accounting to be applied on Gi for charging, statistical or network monitoring purposes. As a result the protocol is analyzed here in order to extract the necessary control information being the MSISDN and APN (3GPP TS 29.061).

Multiple research projects have concentrated on measuring Mobile Internet from these two interfaces. The association of control information to data traffic is also performed by other tools. A commercial tool called RADCOM (*Radcom Homepage* 2010), provides many tools for network protocol analysis and monitoring that include the Gn and Gi interfaces. An example is the Cellular Expert. Such commercial solutions however do not disclose detailed information and although all support the association of data traffic to a particular user using control information, none mentions if e.g. application recognition is performed.

In the academic circle, actual traffic recognition in GPRS/UMTS networks has been performed in the METAWIN project (Ricciato et al. 2006). The project is a study of an operational GPRS/UMTS network in Austria on its GPRS interfaces, including Gn and Gi. It covers the association of control information to user traffic by examining the signaling information carried on the interfaces, like the GTP control messages. It continues by detecting congestion and bottlenecks on the network, undesired traffic on the data like worm infections and the aspect. In Finland the MoMI project has been studying Mobile Internet usage utilizing measurements in the Gi interface.

2.4 Traffic Classification Techniques

Several different techniques have been in use, in order to study Internet traffic over the years. In this section two of the most common such techniques used to identify and classify Internet traffic are introduced. Both of them have different uses and can be used to measurement different levels of detail of Internet packets. These packet inspection methods can be categorised by the detail upto which they traverse the IP packet.

2.4.1 Shallow Packet Inspection

The oldest and most common method is **Shallow Packet Inspection (SPI)**. This method only reads the information contained in the header on the network layer. It relies on simple port mapping and is also called the *port-based* technique (Finamore et al. 2011). It provides information on the origination and destination IP addresses of a particular packet, and it can see what port the packet is directed towards. This is of limited use for modern day traffic because many modern applications like P2P employ randomly chosen ports, effectively making this method unable to fully classify the traffic.

2.4.2 Deep Packet Inspection

In order to overcome the shortcomings of the *port-based* techniques **Deep Packet Inspection (DPI)** was developed. This method goes beyond the network layer headers and inspects the application layer payload to get meaningful information about the contents. It is also called the *behavioural* technique (Finamore et al. 2011). It is most widely used as a network management tool, especially for P2P traffic and also other uses as measuring network performance, intrusion detection etc. One of the main functions of DPI is as a traffic classification tool for Internet traffic. The "deep" in deep packet inspection refers to the fact it moves beyond the IP and TCP header information to look at the payload of the packet to identify the application in use.

Various open and commercial tools are already available that can perform this task. Other more efficient techniques are available (Salgarelli et al. 2007), but they are still under research and don't have the scalability in large real networks (Cascarano et al. n.d.). The DPI classifier uses regular expressions (*signatures*) to identify a particular application taking into account each application uses some unique protocol headers to initiate communication. One of the most widely used tools to perform such classification is Tstat (Finamore et al. 2010). This tool is introduced later on in this thesis. It has already been used to study P2P traffic behaviour (Torres et al. 2009) and Skype traffic (Bonfiglio et al. 2007).

All of the mentioned studies have been conducted in traditional broadband and Wifi networks. However, these techniques can easily be leveraged to the Mobile IP networks to study user behaviour better and holds much promise. Some DPI methods are so sophisticated that when it cannot identify the application responsible for sending packets by examining the packets headers and/or payloads, it examines certain patterns in the flow of packets. Thereby, being able to identify application that would be unclassified otherwise. A good example of this is Skype traffic.

2.5 HTTP Traffic

HTTP is the most popular protocol used in the Internet today. It is the application layer protocol that sits on top of the TCP protocol. Beyond its traditional use in web browsing, where it was mainly used for user-driven content publishing, search and for delivering web advertisements, a number of web-based services and applications also run on top of HTTP. Examples of these are social networking, web mail, on-line gaming, audio and video streaming, news feeds etc. Infact, these types of services and applications account for more than 75% of residential broadband traffic (Maier et al. 2009).

Several studies have shown the same to be true for Mobile Internet. For example, (Heikkinen et al. 2009) studies P2P usage from passive UMTS header traces in Finland from 2005-2007. Here, HTTP is again the dominant protocol with 57-79% of bytes from mobile hand-held devices while email was 10-24%, and P2P was not noticeable. An interesting analysis regarding mobility and web usage in a 3G network from a metropolitan area was conducted by (Trestian et al. 2009). They find social networking, music and e-mail to be the most common application in the mobile web. This is done by characterising application usage by counting the number of HTTP requests.

Also these days most of the web services can be accessed through mobile apps in the mobile devices. HTTP (and it's secure version HTTPS) is also the de facto protocol used by these mobile applications. Most of these applications tend to send their data via HTTP, even though the payload is not HTTP (Falaki et al. 2010).

The reason for the increasing use of HTTP is mostly due to the request/response paradigm between a client and a server who communicate using HTTP. This makes it robust and flexible for many different use cases besides web browsing (Li et al. 2008). Such widespread use has made understanding HTTP traffic composition and usage patterns very important to network operators for many different applications like networking planning and optimisation, traffic prioritisation, marketing analysis etc.

2.5.1 Persistent and Pipelined HTTP

During the early implementations of HTTP, a new TCP connection was used for each HTTP request/response exchange. The overhead incurred in opening and closing of each TCP connection meant that each HTTP request has some latency for end users. In order to improve this the HTTP/1.1 standard incorporated **persistent connections** as the default behaviour.

Here, each TCP connection is kept open and reused to allow several HTTP request/response messages. This allows the client to fetch several resources from the server in a short amount of time and hence improving the latency issues. HTTP persistent connections are also called *HTTP keep-alive* or *HTTP connection reuse* (Cohen et al. 1999). Figure 8 shows how this works.

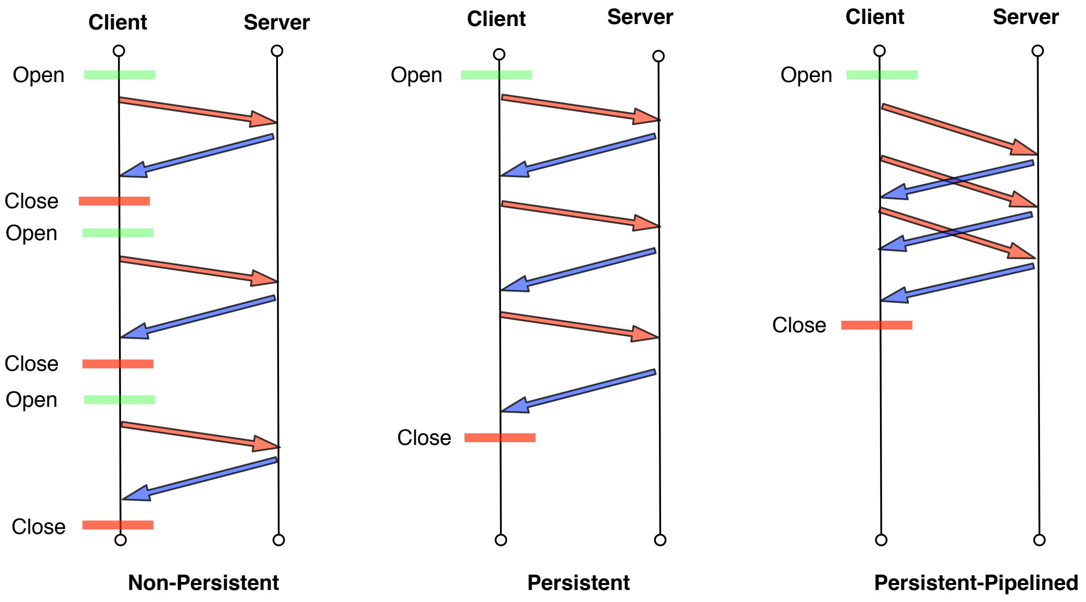


Figure 8: *Non-persistent, Persistent and Pipelined HTTP connection.*

For the existing TCP connection to be reused, there has to be a way to indicate the end of the previous HTTP response and the beginning of the next one. This is achieved by having a self-defined message length for each request/response pairs in a HTTP header.

However, the client does not need to wait for the completion of the current request before sending the next request over the persistent connection. It can send multiple requests without waiting for each response. This is knows as **pipelined connection** (Fielding et al. 1999). In a pipelined connection as shown in Figure 8, the server has to send the response to each request in the same order the requests were received.

2.5.2 HTTP Headers

HTTP header fields are one of the most important components of HTTP for studying and understanding Internet traffic as they define the operating parameters for HTTP. They are used by DPI tools, which look for signatures in the HTTP headers and the payload information over a series HTTP transactions for accurate Internet traffic classification.

Both the request and the reply HTTP messages transmit these headers. There can be several different types of HTTP requests which are called *Request methods*. The two most important such methods are GET and POST. The GET method is used to request data from a specified resource while the POST method is used to submit data to be processed to a specified resource.

The header fields are transmitted as colon-separated name-value pairs in clear-text string format and can be parsed easily. There is a standard core set of header fields defined by the IETF in (Fielding et al. 1999). Any application can however define it's own additional header names and permissible values.

A number of key HTTP header fields can reveal invaluable information about the client and activity carried out in a connection. These include the request method used, the name of the host, the URL of the requested object, the type of content requested, the size of the object returned, the user-agent of the client making the request etc.

2.6 Mobile Internet Usage Studies

After going through the underlying technologies and terms used in the study of Mobile Internet, a look at existing studies carried out in this field in discussed.

As Mobile Internet usage has become significant, it has been a topic of active research in the academic community. However, gaining a comprehensive understanding of mobile service usage presents a lot of challenges given the variety of the networks and devices accessing them. Also, these researches have mostly focused on the technical aspects like network performance and very little has been done in order to understand the service usage from a business point of view. A brief introduction is given on the data sources utilised for such studies and some discussion is then given on the various research conducted by different academic teams.

The most common method to collect data is *end-user surveys and interviews.* Surveys are normally conducted on a carefully selected panel of real users and experts from the industry. This method can give highly detailed usage information but can be subjective as the users will not give honest answers in many cases. Another way is to perform device monitoring of the handsets using certain monitoring software in the handsets. The most comprehensive way is to perform TCP/IP traffic measurements at various network nodes depending on the type of information we need to obtain. This can give us information on the whole subscriber base. Server logs also provide detailed data on usage of particular services.

The earliest attempts to understand Mobile Internet usage was conducted in Japan. Japan was the first market where Mobile Internet truly took off with NTT DoCoMo's i-mode service. (Ishii 2004) performed comparison of PC versus mobile Internet usage in Japan while (Habuchi et al. 2005) studied the use of Mobile Internet alone. Both of them used surveys as a data collection method. Another advances mobile Internet market has been Finland where the first GSM network was installed. Here, mobile service usage measurements have been taken annually in (Kivi 2007) and (Verkasalo & Hämmäinen 2007) [2]. In another works, (Verkasalo 2006) studies how

[2] This thesis has been conducted as part of the same research team.

Symbian phone features are used by instrumenting the handset. He finds that the camera feature and games are the most common multimedia applications.

One of the earliest researches to model IP traffic in a commercial 2G GPRS network was carried out by (Kalden 2004). He studies the usage of different applications and session properties, models WAP and MMS traffic and also mobility and self-similarity in these networks. Similar attempt to understand traffic in a CDMA2000 network was done by (Williamson et al. 2005). In Europe, the study was furthered by (Svoboda et al. 2006) to model and analyze 3G UMTS traffic in addition to GPRS traffic. They used anonymised header traces from 2004 and 2005. They study traffic volume per user and service mix utilising PDP-contexts. They find HTTP to be the dominant protocol which was 40-60% of the total traffic.

Similar work has been done by (Kivi 2007) utilising mobile Internet traffic in Finnish network operators while (Riikonen 2009) studies PC versus Mobile Internet usage and gives statistics on daily and weekly usage behaviour along with the most popular websites visited by the subscribers.

Though carried out in a residential broadband network (Maier et al. 2009) analyse Internet traffic on user session basis and provide some interesting statistics on web usage behaviour. Related work is done studying mobile handheld device (MHD) usage in the same network that use WiFi (Maier et al. 2010). They characterise this traffic to provide insights into how a typical device is used. Their focus in studying per user usage of applications.

More recently (Erman et al. 2010) have attempted to study the use of non-traditional sources of HTTP traffic such as smartphones, gaming consoles etc. This is also done in residential broadband networks with the focus solely on HTTP traffic. By utilising the *User-Agent* field in HTTP headers they were able to identify these devices and analyse their application usage.

Internet backbone hardware makers Cisco have been publishing their annual Visual Networking Index (VNI) reports (*Cisco VNI: Usage Study* 2010). In this report they publish forecasts and analysis on the growth and use of IP networks worldwide driven by a combination of video, social networking and advanced collaboration applications.

3 Network Measurement

Before delving into the analysis process, it is important to understand how the network traffic measurement was carried out in the mobile operator. The measurement conducted in this thesis is based on anonymized packet-level traces taken at previously chosen data points. In this chapter the measurement setup for the MOMI project is explained along with the various data capture tools used to extract the information we want. The challenges faced in the implementation of this setup is then explained along with the privacy issues. Finally the datasets obtained are explained.

3.1 The Measurement Setup

Network measurement in mobile operators have been conducted as part of the MOMI project and its predecessors by Aalto University, School of Electrical Engineering since 2005 in Finland. This measurement setup was an continuation of the previous methods taking into consideration some of the recommendations made by previous work (Riikonen 2009). In addition to the previous measurements, Tstat traffic classifier data was also made available to perform user and device level analysis.

The main objectives of this measurement setup can be summarised as follows.

1. Perform Passive Monitoring so as not to intrude regular network operations.
2. Continue providing the same flow/application flow level traces as previous years so as to maintain longitudinal data.
3. Enable DPI based analysis to get more detailed and accurate information regarding traffic characterisation.
4. Maintain privacy and confidentiality of individual users.

The structure of the measurement setup can be seen in Figure 9.

The network measurement was carried out in a national GPRS/UMTS network of a Finnish Mobile Network Operator (MNO). Since it is a major MNO, it represents a substantial subscriber base upon which the analysis can be made. At least a week's worth of data was decided to be necessary in order to analyse the user behaviour patterns. The measurement traces were recorded on the Gi interface of one of the GGSNs in the operator's network. Since, the Gi interface connects the GGSN with the Internet link of the operator it allows for high level analysis of the total traffic of the network, e.g., service mix, up and down data volume etc. One thing to remember is that since we record packet traces going through a single GGSN only it doesn't

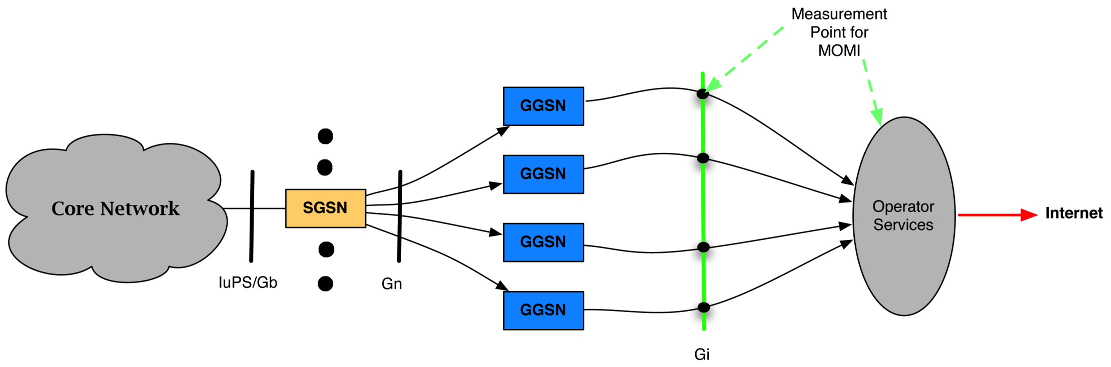


Figure 9: *MOMI Network Measurement Setup for 2010. (Adopted from Riikonen 2008)*

measure all the traffic for the operators. It can be however assumed that the traffic load is equally distributed among its GGSNs and the captured data can be taken to represent a sufficient subset of the total traffic (Riikonen 2009). Also, since every PDP-context goes via the same GGSN, all traffic related to one TCP connection and other connections related to the same "app usage session" will be included.

The measurement hardware consisted of Linux based computers with Data Acquisition and Generation (DAG) cards and multiple hard disks for storage. This equipment recorded the traces passively so did not obstruct the network in any way. There were two types of Linux based software tools utilised to capture these packets into a form that could be later pre-processed and analysed. Both of them were proprietary tools developed by modifying standard open-source packet capture tools to fit our requirements. The first was used the libpcap library which was also used in the 2008 measurements. The second capture tool was a modified version of Tstat which is a tool for collection and statistical analysis of TCP/IP traffic. It provides us with a lot of information through its DPI engine that makes analysis easier. In addition to TCP, UDP and IP packet headers it is modified to capture HTTP headers which provides accurate web traffic and device identification data.

3.1.1 Implementation Challenges

The task of network traffic capture is not an easy task because of various reasons. Due to the immense scale of today's networks and the sensitivity of the user related data there are not only technical limitations to the traffic analysis process but also legal and socio-political issues (Clegg et al. 2009). The added complexity of mobile data networks is another major issue.

As data networks become faster and more complex, the traffic recognition has to be done as fast as possible. But there are certain limitations to current monitoring hardware with regards to memory and system bus throughput (Risso et al. 2008).

The increase of the amount of traffic being transferred in a link per second by a factor of two, doubles the size of data for processing by the traffic recognition software. The traffic recognition process may be split into two phases. Capture the data on storage and process the data. However, the storage required for Gbps speeds may be very large, particularly in traces of long duration. Furthermore, the analysis phase requires a considerable amount of time to process the data. The optimal solution would be to perform traffic recognition on the fly, by executing both phases at once and saving storage. In this case the performance of the analysis phase is of great importance. If it cannot keep up on the traffic speed the result would be that some packets would not be included in the analysis.

3.2 The Dataset

There were different types of traces collected in the operator network measurements. As mentioned earlier, there are different data sources that provide us with different kinds of details. A brief explanation of where these data are collected, what information they carry and how they are used is now given.

It is to be noted that the datasets obtained from the live networks have to be post-processed into an acceptable level of abstraction such that useful statistics could be derived from it. The pcap packet traces need to be later processed into flows so as to get a manageable level of abstraction. The Tstat tool performs some advanced processing using its DPI capabilities but it still requires post-processing to get the type of granularity we need.

3.2.1 The Flow Data

The flow data was obtained from the pcap packet headers. The packet headers were aggregated into flows using the CoralReef tool (*CoralReef homepage* 2010) from CAIDA [3]. This tool takes the pcap format packets and gives out the basic 5 tuple flows using a timeout of 60 seconds (Riikonen 2009).

The flow data summarises the network traffic on a connection level. All packets that belong to a “flow” are aggregated into a record that yields information such as, the number of packets, volume of the flow, or the TCP flags that were present in that flow. Flow data is often used for traffic engineering, routing optimisation, demand analysis, or network dimensioning. Although flow data is an important data source to understand how much data is flowing from where to where, it usually does not

[3] The Cooperative Association for Internet Data Analysis (CAIDA) is a collaborative undertaking among organisations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure. CAIDA provides a neutral framework to support cooperative technical endeavors.

expose enough information to easily attribute a flow with the Internet application that caused the data transfer (Moore & Papagiannaki 2005). This data was not used for the analysis conducted in this thesis.

3.2.2 Tstat Logs

Tstat is a passive IP network sniffer that can both capture the packet traces and produce a wide set of statistics. Its basic entities are flows. It can identify client-to-server and server-to-client flows. The core features of Tstat are implemented in the trace analyser. It is shown in Figure 10.

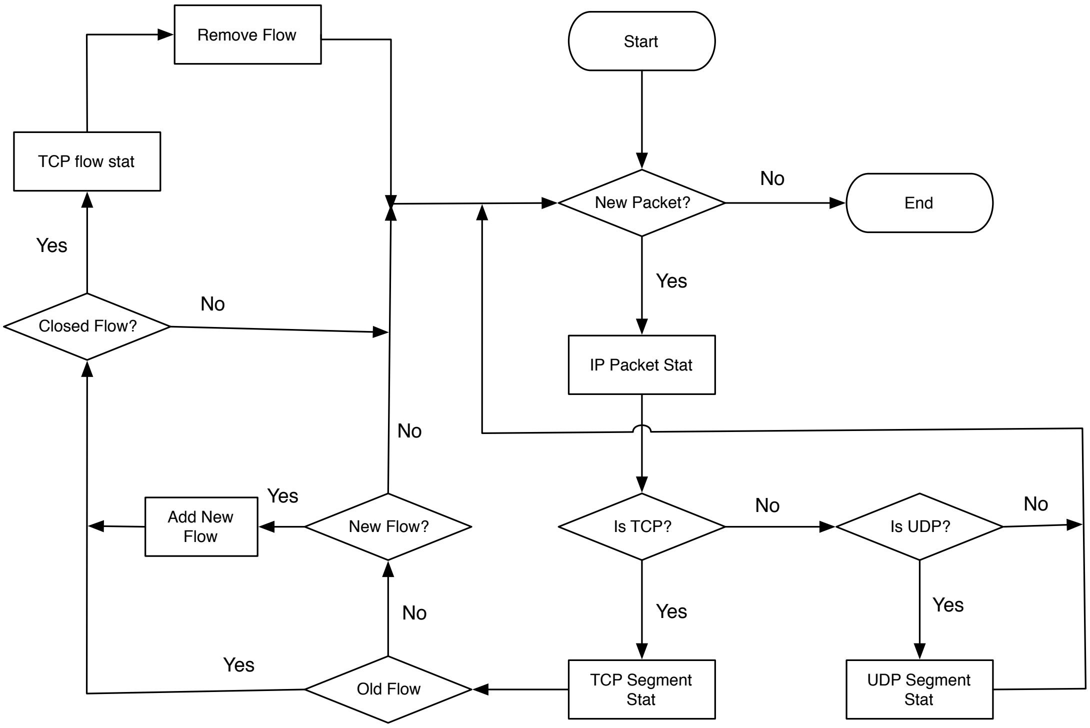


Figure 10: *The Tstat Analyser. (Source: Mellia et al.)*

At the lowest level it processes the IP layer packets to produce some per-datagram statistics like bitrate, packet length etc. The control then goes to the transport layer analysis. It maintains common statistics like traffic, volume, number of packets and flows for both TCP and UDP etc (Finamore et al. 2010).

On the application layer, traffic classification is done. There are a set of classification engines that are triggered to identify which application has generated the flow under analysis. This process is carried out even without a complete payload. This is one of the biggest strengths of Tstat and one that is of interest in this thesis.

The different formats of output statistics produced by Tstat are:

Connection Logs: These are set of text files that have details for each monitored connection. They are arranged as simple tables where each column is associated to a specific information while each line contains statistics related to the two flows of a connection. It generates several of these log files for TCP, UDP, multimedia (RTP/RTCP) etc.

Histograms: These are collections of empirical frequency distributions of parameters such as IP packet length, per protocol bit rates, no. of flows active in a time slot etc. These are produced periodically as 5 min bins by default.

Round Robin Database: These are databases that have been designed as a scalable mechanism to store historical data by aggregating them with different granularity. The newer data are stored with higher frequencies while oldest data are averaged in coarser time scales. Thus, reducing space. These are not used in this thesis.

Only the TCP and UDP connection logs are used in this thesis.

3.2.3 HTTP Headers

The Tstat tool was modified for this particular implementation purpose to provide HTTP header traces in the logs. A script was then created to get the HTTP headers from these logs. The headers include information IP addresses, port numbers, and all HTTP commands and HTTP header information, e. g., a GET /index.html command including the URL and all HTTP headers (i. e., Host:, Cookie:, etc.) that follow the command but not the HTML page that is requested. This information is sufficient for reverse-engineering the actions and interests of the users.

The most important information however are the browser User-Agent (UA) strings. Using UA string parsing tools we can recognise the browser and the OS that generated the request and hence enables device identification. This was one of the main goals of the thesis and extensive analysis was carried out to study the device feature usage patterns.

3.3 Issues with the Measurement Setup

As mentioned above in this chapter, this was the first implementation of the measurement setup with the Tstat tool. This brings with it the possibility of several errors in the measurement process. One of the key goals of this thesis is to examine the integrity of this new dataset before performing any further analysis. To make sure that the data collected is complete without anything missing or corrupted, the data underwent some preliminary due diligence checks. It is important to mention some of the issues discovered and the adjustments made to ensure the dataset is unbiased and it's limitations known.

3.3.1 Dataset Limitations

The data check done to the raw data obtained after the measurement period revealed few issues in the captured data. These issues are listed below.

1. There are some issues in measurements for a few hours in the whole data, about 1% of trace is not present.
2. Only few Tstat histogram files were seen in the dataset.
3. Comparing the TCP log data and HTTP headers. Around 30% of the HTTP headers are missing when compared to TCP log data.

Although these issues showed that some part of the dataset was indeed missing, they didn't provide a significant impact of the analysis process. The few hours where 1% of the data were missing represent a very small portion of the total data and were random to skew the results. The Tstat histograms were not used for the purpose of this thesis at all and can be ignored. The 30% missing HTTP data however needed some further analysis to examine if it introduces any bias in the analysis process.

Study of the missing HTTP header data showed that the missing data was quite random and not specific to any hour of the day, content type or HTTP application. Hence, this missing data can be ignored and the analysis can be performed on the remaining 70% of the data which still represents a good enough sample size for the study to be conducted on HTTP. Further explanation detailing how the data is prepared for the final analysis is given in the next chapter.

3.3.2 Privacy Issues and Anonymisation

Since we are monitoring the actual traffic flowing through the operator's network without the user's knowledge, great care must be taken to ensure the privacy of the end users. This means maintaining the anonymity of the users so that there is no means of identifying an individual and by not storing any confidential data in our university servers.

The legal system in most countries have legislations to regulate the access and analysis of such private data. Finland also has such legislation in place to protect individual privacy. These laws are not technology specific but are interpreted in generic terms. Basically, it considers all information that is not intended for the general public as confidential (Riikonen 2009). Two laws that are of interest in this case are the Act on the protection of Privacy in Electronic Communications (516/2004) and the Personal Data Act (523/1999). The first one which is directly related to our research states that user identification can be done by the operator/institute to develop its product and services, billing, marketing research etc with the user's consent. However, it forbids the identification of an unique individual for statistical research. The second law states that even though personal data can be used for study, research or statistical purposes without explicitly identifying the individual, he/she always has the right to prohibit this usage.

Such issues have been taken into account thoroughly during all our measurements. To ensure this, anonymisation scripts were run in real time during the measurement. We obtained the RADIUS logs from the operator in already anonymised form into our servers. Due to dynamic IP address allocation in mobile networks, IP addresses can no longer be tracked back to individual users and hence doesn't need such comprehensive anonymization.

The HTTP headers is another dataset that needs to be anonymised properly because it provides detailed information on web usage. With regards to this, certain techniques have been applied to avoid logging of potentially private websites. This was done by assuming that a user usually visits public and popular websites several times a week while only visiting private websites a few times. The details of this are outside the scope of this thesis but the general rule used for logging HTTP headers information in our measurements was that, *only frequently observed header values were recorded.* This rule was seen to be working successfully in our measurements.

Another factor to take note of is that DPI has been a contentious issue in the Internet industry. It has been subject to controversial debates about network neutrality and online privacy (Riley & Scott 2009) for a long time now. However, the DPI tool in Tstat here does not store any such private data, neither does it look at complete packet payload during packet inspection (Finamore et al. 2010). Hence, our DPI provides only application identification results and not any sensitive information mentioned above.

Besides the confidentiality of the users, there is also the issue of maintaining the business sensitiveness of the operator. In order to preserve this all the researchers concerned with the research have signed strict Non-disclosure agreements (NDAs). The access to the servers is given to only those select researchers. For the purpose of analysis when the data needs to stored in the researcher's computers, it is done by using encrypted partitions. Also, the results are presented in such a way as to avoid any such inferences about the operator by proper normalisation and not made public without the prior approval of the stakeholders.

3.4 Post-processing tools

There were several tools used in order to perform the necessary analysis on the datasets. Since, these datasets were stored in a Linux based server most of these are command line tools.

The tools that were used for the measurements are proprietary modifications to the open-source tools like pcap and Tstat. Before the different datasets can be used to generate the various aggregated results certain amount of pre-processing is necessary in order to clean up the data and make it suitable for further statistical analysis. This may be using certain tools to get the data into a format usable, changing certain fields to make the format more readable, perform mapping between several datasets and using performance boosting tricks to make the processing of the enormous volume of data quicker and more efficient. After getting the data into a form that was aggregated enough to perform statistical analysis, they were imported into statistical and analytics tools. From this the desired plots and results were obtained and discussed upon.

So basically there are two categories of applications used for the two steps. The main tools for each step is shown below:

Data Aggregation and cleaning tools
Perl, awk, bash scripting for text processing, cleaning, mapping and aggregation, other common command line tools for integrity check of data and pywurfl API and python language for device identification.

Statistical analysis and plotting tools
Statistical analysis tools like R, SPSS and Microsoft Excel for analysis and plots, gnuplot tool for some plots.

4 The Analysis Framework

Having gone through the research objectives, the technical literature and the data collection methodology, this chapter now introduces the analysis and pre-processing phase of the research. These stages occur naturally in any research project and have been followed in the MOMI project (Riikonen 2009).

The research framework thus established forms the basis for the data visualisation and reporting phase which is a deliverable of this project. The different datasets and the tools chosen to analyse them have already been explained in the previous chapter. The thesis now goes into explaining how these are related to each other and how the various metrics are derived from them in order to answer the two main research questions of this thesis.

4.1 The Analysis Process

A general overview of the different datasets and their relation to the different results produced is given in Figure 11. A brief explanation of the process is given below with more details presented in the coming sections.

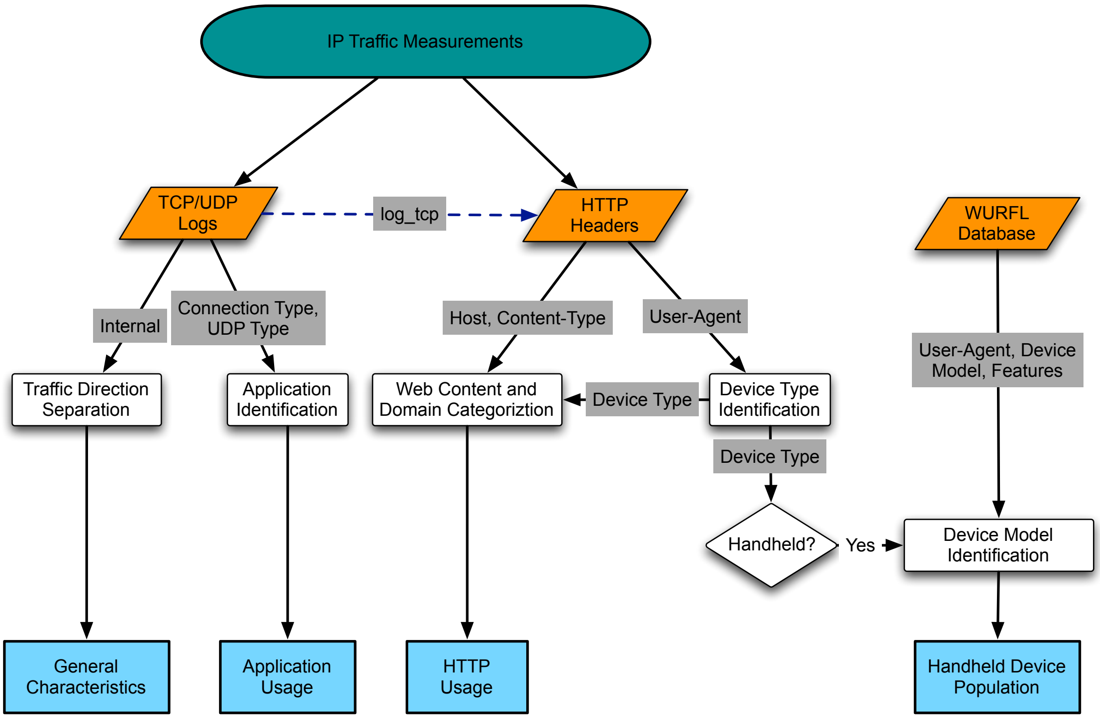

IP Traffic Measurements
TCP/UDP Logs
log_tcp
HTTP Headers
WURFL Database
Internal
Connection Type, UDP Type
Host, Content-Type
User-Agent
User-Agent, Device Model, Features
Traffic Direction Separation
Application Identification
Web Content and Domain Categoriztion
Device Type
Device Type Identification
Device Type
Handheld?
Yes
Device Model Identification
General Characteristics
Application Usage
HTTP Usage
Handheld Device Population

Figure 11: *The Analysis Process*

In the measurements both pcap traces as well as Tstat traces were captured. The pcap traces were not used in this particular analysis and not shown here. The Tstat traces have TCP and UDP flows logs with application identification done for each flow by it's DPI engine heuristics. This implementation also produces the HTTP headers from this traffic separately for the HTTP traffic analysis.

In order to answer the first research question, the TCP and UDP logs were first used to perform the general traffic analysis. The *Internal* field can be used to get the direction of the traffic flow to separate the uplink and downlink. The logs contained fields including the identification of the applications based on DPI heuristics. Hence, we could use this dataset to observe the traffic symmetry, share of TCP and UDP, various application protocols and their diurnal patterns etc.

The HTTP headers contained information about the web traffic and consist of several fields which can be used to get a variety of information about the web traffic. The *UA* fields can contain enough information to identify the browser, OS and the application that generated the HTTP requests. Using the OS identification from the UA string the PC devices and the handheld devices can be separated.

There are also other HTTP header fields such as *Content-Type*, *Host* etc that provide information about the type of content present in the HTTP traffic and the domains that are requested. These contents and domains were categorised into some general categories for the purpose of studying the web usage.

For the other main question that the thesis tries to answer, the *UA* field was used. For handheld device this can be used to identify the particular device as well as its properties. A freely available database of UA to device mapping called the Wireless Universal Resource File (WURFL) was used in order to perform this identification.

4.2 Internet Traffic Classification

Identifying the applications that make up the traffic flows has always been an important way of understanding their usage. This was previously performed using port based methods. But it is seen to be unable to classify P2P applications that did not use the traditional ports. They tend to use ports dynamically which makes port mapping ineffective in identifying these applications. Studies such as (Karagiannis et al. 2005) have shown that this is indeed the case. Hence, to get a more accurate insight into the composition of the network traffic and identify the P2P protocols correctly, a DPI based approach is taken. The process is explained below and the results presented in the next chapter.

4.2.1 Tstat Traffic Classifier

The Tstat tool has a robust DPI traffic classifier that is able to identify applications based on several advanced heuristics. These methods are based on a combination of payload signature identification and behavioural classifiers (Finamore et al. 2010). Hence, not only is it capable of identifying usual traffic such as HTTP, email, messaging etc but also P2P traffic such as Bittorrent, eMule, kaaza, gnutella etc and online streaming protocols like RTP, RTCP, RTMP, Skype and also online messenger protocols such as MSN and Yahoo (*Tstat homepage* 2010).

Since, there are around 60 different applications identified in the measurements, these applications were grouped into more general categories. The different protocols identified and their categories are given in Table 1 below.

Table 1: *Categorization of the Applications identified*

Category	Application
HTTP	HTTP
Bittorrent	Bittorrent, Bittorrent MSE/PE
Other P2P	DirectConnect++, eMule, Gnutella, Kaaza, eDonkey etc.
Partial P2P	P2P over HTTP, eDonkey/Bittorrent etc.
SSL	SSL TLS
Email	POP3, SMTP, IMAP
Streaming	RTP, RTCP, RTSP, Sopcast
Skype	Skype, Skype SIG, Skype E2E, Skype E2O
Messaging	MSN, Yahoo Messenger, XMPP
Others	SSH, DNS, Other

It is to be noted that since the share of Bittorrent was seen to be very large among the P2P traffic identified and it has been categorised separately from rest of the P2P traffic. All other P2P traffic identified have been collectively categorised as *Other P2P*. There is a category for *Partial P2P* traffic which are those cases that the DPI couldn't conclusively classify as being P2P. These were seen to be mostly P2P protocols operating over the HTTP protocol and those cases where the classifier couldn't distinguish which P2P protocol it is (e.g, eDonkey or eMule etc.). Also, the SSL traffic was classified separately from HTTP here even though they are both web traffic using the HTTP protocol.

4.2.2 Comparing DPI to Port based Analysis

In order to see if the results of the DPI traffic classification is indeed better than that of the older port based method, a comparison method was used. The classification done was based on the standard ports used by these protocols and simply mapping the ports in the flows.

A way is needed to establish how effective has the DPI classifier been in comparison to the port based method. A comparison method based on similar analysis done by (Maier et al. 2009) was chosen for this purpose. For each protocol, we choose three parameters:

$\mathrm{V}_{\mathrm{DPI}}$ is the volume identified by DPI tool for a given protocol.
$\mathrm{V}_{\mathrm{Port}}$ port is the volume identified in the default ports of the protocol.
V_{PD} is the volume of the intersection of the two previous methods, i.e on the default port of that protocol and also identified as that protocol.

For each protocol, V_{PD} / $\mathrm{V}_{\mathrm{DPI}}$ is the fraction of the traffic volume that is observed on P's default port that the DPI tool identifies as P.

V_{PD} / $\mathrm{V}_{\mathrm{Port}}$ gives the proportion of the traffic that would be correctly identified by the port mapping method.

Calculating these ratios for each protocol provide us with a means to make inferences about these two methods of application classification.

4.2.3 Measuring Diurnal Application Usage Patterns

To provide a different perspective into the usage of the applications identified two metrics are used which are explained here briefly. Both of these metrics are calculated by first aggregating the traffic volumes of each protocol into a single 24 hour period. Hence, we have traffic volumes of each protocol for each hour of this aggregated 24 hour day.

The first of these metrics is the *Volatility* of the application protocol. It gives us an insight into how the traffic volume is spread during the course of the day for each protocol. This metric is quite common in financial markets in order to study how the stock prices fluctuates throughout the day. In studies related to Internet usage there is only the Cisco Visual Networking Index (VNI) Usage research (*Cisco VNI: Usage Study* 2010) which is found to be using this metric. Mathematically, it is de fined as:

$$Volatility = \frac{\sqrt{\frac{1}{23}\sum_{i=1}^{24}(V_{hour,i} - \overline{V_{hourly}})^2}}{\overline{V_{hourly}}}$$

$V_{hour,i}$ = volume for each hour of that application
$\overline{V_{hourly}}$ = average hourly volume of that application

The other metric is the *Peak hour to Average Hour Ratio (PA_{Ratio})* of the application protocol. It gives us an indication of how much more volume is flowing in the busy hour as compared to the average for each protocol. Mathematically, it is defined as:

$$PA_{\text{Ratio}} = \frac{V_{hour,max}}{\overline{V_{hourly}}}$$

$$V_{hour,max} = \text{Maximum hourly volume of that application}$$
$$\overline{V_{hourly}} = \text{average hourly volume of that application}$$

One important observation was made while performing this sort of hourly aggregation from the data files. Although the data files were separated on a hourly basis, i.e. there was one data file for each hour of each day of the measurements, there was a possibility that the TCP or UDP connection that started in one hour ended in the another. This is understandable because these connection durations can vary quite a lot from a few seconds to several hours.

Because of this instead of considering the hours aggregated from the data files, the timestamps we read from the data itself by taking into consideration the entire week of measurement. The hourly aggregation thus obtained is the true representation of the traffic characteristics.

4.3 HTTP Request/Reply Pairing

In section 3.3.1 it was mentioned that despite 30% of the HTTP data missing from the HTTP header, analysis is still conducted on the data. The results so obtained will still be unbiased as the remaining data is a representative sample. Before carrying on with the HTTP data an important pre-processing step is done to pair the request/reply for a TCP connection.

As mentioned in section 2.5.1, more than one subsequent HTTP messages can be sent over one persistent TCP connection. Therefore, there are a number of HTTP request and reply messages associated with each TCP flow. Studies have shown that most connections have more than two requests. Infact, (Schneider et al. 2012) found that 60% of all HTTP requests are not the first in a TCP connection.

Performing the same kind of checks on the dataset also gave similar results. Of the HTTP headers, 41.94% of the data under study had only a single HTTP request/reply pair (non-persistent HTTP connections). Also, 0.35% didn't have any corresponding request or reply pair. This means there is either only a single HTTP request (GET/POST) or reply (REPLY) header, not both. These are not significant enough to impact the result greatly and can be ignored. The remaining data had more than one HTTP GET(or POST) and REPLY pairs.

Despite this knowledge, only the first request/reply pair will be considered in this thesis. The reason for doing this is that knowing only the first HTTP request/reply pair will give us all the information we need about the Device and the OS generating those requests. The results presented in this thesis for HTTP headers will consider the number of TCP flows and bytes transferred during the TCP connection. Although, the byte volume can be obtained by considering all the HTTP requests in that TCP connections and adding up the *Content-Length* header, Schneider et al. (2012) have found that this field can have incorrect values for cancelled transfers or erroneous Web servers. Hence, the byte volumes for the TCP connections are considered instead as being more accurate.

4.3.1 The Pairing Algorithm

The steps involved in making the corresponding HTTP request/reply pairs from multiple persistent connections is shown in the next section. This was required because not all the HTTP messages have all the desired header information and we need to choose the best pair if this information is missing in the first HTTP pair.

Before, the pairing was performed the HTTP requests and replies were mapped to the corresponding TCP flow via the ID value present in both the tcp logs and HTTP headers. Along with the ID, the IP addresses, ports, bytes and timestamps corresponding to the TCP flow were also mapped. As mentioned previously, some part of the HTTP data were missing.

1. Find all the cases with lone GET/POST or REPLY and exclude those.
2. For every case where there is exactly one GET/POST and REPLY pair, mark them as a pair.
3. For cases where there are more than one GET/POST and REPLY, choose the first one that satisfies the criteria and mark them as a pair.
    - For GET/POST requests, choose the first one that has both *Host* and *User-Agent* header fields.
    - Choose the first REPLY after the GET/POST which has both *Content-Length* and *Content-Type* header fields.
4. Combine the corresponding GET/POST request and REPLY pair into one single line by adding the corresponding volume information to represent each TCP connection.

4.4 Understanding HTTP Headers

To study mobile web usage characteristics, HTTP headers were used. These headers consist of individual HTTP Request and Response Messages. Each HTTP Request-Response pair provides information regarding the usage of the particular web resource. These can be used to identify the website that is being requested, the browser that made the request, the web resources that link to it as well as some information about the type of content sent back.

For the purpose of this thesis the main fields used were the *Host* and *User-Agent* fields in the request header and the *Content-Type* field in the response header. It is to be noted that there are several cases when some of these fields are missing as many web servers and security softwares tend to hide them for security reasons.

The *User-Agent* field is a very important field in these headers and has been used extensively in handheld device identification and analysis in the industry. This process is explained in detail section 4.5. This field is also used to identify the OS. However, this is only done for the HTTP traffic and cannot be said representative of the entire traffic flow. The identification is performed by modifying the Perl HTTP::BrowserDetect UA parsing library. This method identifies the OS in the PCs and laptops while the identification of the handheld OS is carried out in the device identification analysis explained later.

Content-Type Distribution

The first observation was performed by analysing the content mix of the HTTP traffic using the *Content-Type* field. This field specifies the mime-type of the requests object. More than 1000 different content-types are found in the dataset. So, these were classified into a few general types. This classification was done by manually inspecting the different content-types and creating some basic keywords to categorise the traffic into multimedia content like video, audio and normal web content like html, images, flash, binary applications etc. Here flash video (flv) was categorised with other web video and other flash content like banner ads are categorised as flash.

Here, same category of mime-types were taken together and grouped into one Content-Type by giving a “web/” prefix . For example, all image objects with mime-type like image/jpeg, image/png were grouped into *web/image*.

Website Categorization

Identifying what websites and web services that the users in the networks are accessing is of great importance in studying the behaviour of mobile Internet users. The next analysis done was performed to compare the website categories visited by the PCs and the handheld devices.

The *Host* field was used for the identification of the sites visited. Since, the total HTTP data consists of the main webpage objects along with images, stylesheets, videos etc that are embedded in the webpage, only html objects with Content-Type web/html were considered.

Since, there were well over 100,000 different domains in the dataset, these websites were then categorised into more generic categories by inspecting the URLs. Some of these categories were Social Networks, Video Sharing, News Portals, Online Shopping, Online Gaming etc. These categories are based on the online URL categorisation tool from Mcafee (McAfee TrustedSource 2010). For the purpose of this thesis, only the top 100 domains are considered and manually categorised.

From this categorisation we can look at the most popular website categories for both PCs and handhelds in terms of both traffic volume and number of hits. The categorisation done for the websites is shown in Table 2 below.

Table 2: *Categorization of the Web Domains*

Web Category	Example Sites
Social Networking	facebook.com, twitter.com, vkontakte.ru
Video Sharing	youtube.com, vimeo.com
Search Engine	google.com, google.fi, bing.com
Online Advertising	doubleclick.com, google.fi, bing.com
Online Gaming	zynga.com, miniclip.com
Online Entertainment	mtv3.fi, yle.fi, tv7.fi
Software, Antivirus, Windows Update	armdl.adobe.com, symantecliveupdate.com, windowsupdate.com
News Portal	iltalehti.fi, hs.fi, kauppalehti.fi
Adult Site	pornhub.com, redtube.com
Online Shopping	nettiauto.com, nettix.com
Online App Store	store.ovi.com, market.android.com
Financial News	kauppalehti.fi, sm-liiga.fi
Online Gambling	paf.com
Online Repository	repository.maemo.org
Online Portal	htc, apple, elisa, hsl.fi

Any domains related to the operator website, any third party services like surveys and Content Delivery Networks (CDNs) have been filtered out.

4.5 Handset Device Identification

One of the main objectives of this thesis is to create a mechanism for identifying the handheld devices that are generating the HTTP traffic. As mentioned previously identifying the end devices is not easy with network traffic measurements. However, being able to do so is increasingly important with the explosion of Internet capable mobile devices besides desktops and laptops. Answering questions such as: *Which are the most popular devices/platforms in use?*, *What device generates the most web traffic?*, *Is the web browser the most popular way of accessing the internet in these devices or are native applications more popular?*, is a compelling topic for researchers as well as the operators, policy makers, content providers and device manufacturers. This gives a deeper insight into understanding user behaviour and also helps the operators plan for future growth.

This section looks into how this is performed in this thesis and what sort of information is available about these devices.

4.5.1 Identifying Handset Devices

As mentioned in the earlier chapter we have the *User-Agent (UA)* field available from the HTTP Request header. This information has been used to make device identifications in other researches. For instance, (Erman et al. 2010) have used the *User-Agent* field to identify non-traditional sources of web traffic like game consoles, phones and TV sets among DSL broadband users in the US. A similar study is done among European DSL broadband users by (Maier et al. 2010) to identify mobile handheld devices (MHDs) in their home Wifi networks.

Hence, a similar approach was taken in this thesis, where the *User-Agent* field is extracted from the HTTP headers and the device identification performed. However, it is necessary to first separate the HTTP traffic generated by the handheld devices from that by traditional PC devices. To do this a *Perl UA parser* was utilised that can look at the UA string and separate the handheld devices. The parser can also get the OS from the string for PC based devices, which was used to identify the shares of OSes in the HTTP traffic.

The UA strings potentially carry a lot of information regarding the devices and applications used to generate this traffic. For example, a typical Firefox based UA would look like: *Mozilla/5.0 (Windows; U; Windows NT 6.1; ru; rv:1.9.2.3) Gecko/20100401 Firefox/4.0 (.NET CLR 3.5.30729)*. This gives us enough information to obtain the OS as Windows and the browser as Firefox 4.0. Likewise a typical iPhone user-agent would look like: *Mozilla/5.0 (iPhone; U; CPU like Mac OS X; en) AppleWebKit/420+ (KHTML, like Gecko) Version/3.0 Mobile/1A543a Safari/419.3*. From this we can infer that this is generated by a iPhone device. Also, many applications use custom UAs like *Evernote Android/110327 (fi); 2.1-update1/XWJM2; GT-I9000/7;*. This allows us to infer that this request is generated by the Evernote for Android Application. Hence, we can utilise this information to differentiate the type of device (PC or handheld), identify the device (if handheld) and also potentially identify the application.

After separating the UAs belonging to handheld devices, the identification mapping was carried out in them. Upon researching and testing different tools available to do this, the WURFL database was chosen. This choice is driven mainly by its widespread use, comprehensive and up-to-date database and availability of APIs to extract the information from the database (*WURFL homepage* 2010). The method used to perform device identification from WURFL is now explained.

4.5.2 The WURFL Database

WURFL is a freely available XML-based configuration file that contains information on around 500 device capabilities and features for a large variety of mobile devices. It is quite extensive in the number of devices it covers as the information is contributed from around the world and is updated on a regular basis. Hence, containing information on even the newest devices.

Now, in addition to the XML configuration file there are also several APIs available so that you can interact with the configuration file through your code. In this analysis a python based library called pywurfl (*pywurfl homepage* 2010) is used. The main reason for choosing this is because it is maintained and updated currently by its author whereas the Perl version is outdated. There are also Java and PHP versions available.

WURFL works on the concept of family of devices where all devices are descended from a generic device. The algorithm then puts it into more specialized groups like Nokia or Android or iOS. The search is then done on the devices in that specialised group to get a match. If however, a match is not found then it “falls back” to its previous group so that we can at least identify it as belonging to a certain family of devices. For example, if an UA is found to be in the Nokia family of devices but the exact model numbers cannot be matched then it is identified as just a Nokia device and will have the generic capabilities.

Now, it is to be remembered that the UA strings may not match exactly with the one in the WURFL database, even for that of the same device. Two UAs may essentially be of the same device but aren’t exact matches and hence simple matching of given UA to WURFL won’t yield a positive match. For example both these UAs are from HTC Hero but the latter is not present in the WURFL database.

Mozilla/5.0 (Linux; U; Android 2.1-update1; en-fr; HTC Hero Build/ERE27) AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Mobile Safari/530.17 , and

Mozilla/5.0 (Linux; U; Android 2.1-update1; en-us; HTC Hero Build/ERE27) AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Mobile Safari/530.17 CSOUTH-6200

Also, there are different families of user agents that have different structures for different families of devices like iphones, androids, windows phones, blackberries etc. For example:

BlackBerry8800/4.2.1 Profile/MIDP-2.0 Configuration/CLDC-1.1 VendorID/134,

Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0) BlackBerry8800/4.2.1 Profile/MIDP-2.0 Configuration/CLDC-1.1 VendorID/134

Both of these are UAs of the same device but the former is standard "well-behaved" UA for Blackberry devices. But Blackberry allows changing the UA string in the browser setting to a form as shown in the latter case in which it is disguising as Microsoft Mobile Explorer. This will inherently cause the match to be wrong as the non-standard form is usually not present in the WURFL database. So, instead of having a single matching algorithm for all different types of devices, a two step process is employed.

1. In the first step a keyword search was done in the UA to identify which category it belongs to. This differentiates it as a Mozilla based, iPhone, Nokia, Windows etc. type of UA.
2. The appropriate *UA handler* was then chosen from a pool of handlers, each of which is specialised to handle the UA string belonging to that particular family of UA strings.

This way the accuracy of device identification was vastly improved as it takes into consideration the subtle differences between two UAs.

4.5.3 Choosing Handset Features for Analysis

For any device identified by WURFL, over 500 feature information may be available about the device. Although in most of the cases these features are not given. For the purpose of this thesis the feature set extracted was chosen as a smaller subset among them. The features extracted were chosen based on their usefulness to the analysis to be made as well as their presence in most of the device feature list.

These features can provide rather interesting analysis possibility of the type of devices in use and also answering what are the main features of the most popular devices. Appendix A provides a list of the features that were extracted. Here, results are shown for the manufacturer brand, device OS and model.

4.5.4 Identifying Browsers and Applications

With the advent of the App stores for various mobile platforms the use of applications for various online services is also popular these days. Many of those applications use HTTP as their protocol to communicate with the online service.

Although the WURFL tool cannot identify these Apps, the *UA string* of many devices also contain some information about the application that generates the traffic. Most of the browser generated UAs can be identified because almost all of them contain some common patterns. For example, both the UAs shown below are browser based.

Mozilla/5.0 (Linux; U; Android 2.1-update1; en-fr; HTC Hero Build/ERE27) AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Mobile Safari/530.17 , and

BlackBerry8800/4.2.1 Profile/MIDP-2.0 Configuration/CLDC-1.1 VendorID/134,

This first UA string starts with the keyword *Mozilla*, which is used by most of the popular browsers for historical reasons. This was indeed observed by testing the top browsers from the most popular mobile and desktop OSes. Only the Opera browser is seen to have a different keyword *Opera*. The second UA is specific to Blackberry and can be said to be from a browser.

On the other hand application have a slightly different pattern, in that the name of the application mostly occurs in the beginning of the UA string.

For example, the TweetDeck application generates an UA string that looks like *TweetDeck 0.9(GT-I9000; us; Android 2.2)*.

Hence, this knowledge can be used to show how HTTP traffic for mobile browsers and applications may look like.

4.5.5 Issues and Limitations

Although the WURFL database is quite comprehensive in listing the vast majority of handheld devices and the API used handled most of the UA matches accurately, there are still certain cases when the device identification is seen to be inaccurate. There were around 2300 unique user-agents identified in the measurements. Most of them conformed to the standard UA string for their respective device types. But there were also quite a few non-standard UAs that are generated by the various applications used by these devices that make it difficult to match. Another issue was with Android specific UAs. Given the diversity of the devices and manufacturers that use the Android operating system, there are so many different UAs that are similar in structure but different only in the device names, for example. Few of these UA strings are:

Mozilla/5.0 (Linux; U; Android 2.1-update1; en-gb; GT-I9000 Build/ECLAIR) AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Mobile Safari/530.17

This is a UA string for the Samsung Galaxy S that is identified correctly by the pywurfl library. However, another variation of the same device is wrongly identified as being HTC Hero which is shown below.

Mozilla/5.0 (Linux; U; Android 2.1-update1; en-us; GT-I9000 Build/ECLAIR) AppleWebKit/525.10+ (KHTML, like Gecko) Version/3.0.4 Mobile Safari/523.12.2 (AdMob-ANDROID-20101109)

The corresponding UA in the WURFL database that the second UA matched is given below.

Mozilla/5.0 (Linux; U; Android 1.5; en-us; HERO200 Build/CUPCAKE) AppleWebKit/525.10+ (KHTML, like Gecko) Version/3.0.4 Mobile Safari/523.12.2 (AdMob-ANDROID-20090728)

This wrong match is due to the fact that the second UA is closer to the third UA shown here than it is with any UA for Samsung Galaxy S in the WURFL database in terms of distance. By distance we mean how many characters are different between two strings. The algorithm implemented is based on taking into account this distance between the UAs where different handlers (iPhone, Nokia etc.) use different thresholds to get a match. This is a good way to match the UAs but as seen it can sometimes lead to wrong identification.

So, in order to address such issues certain patches had to be run into the result of the device identifications done by WURFL. These were done by custom perl scripts and have resolved almost all of the major issues regarding the false identifications. Also, the application identification from the UA strings was carried out wherever possible to get the share of the applications being run in the devices.

5 Analysis Results and Discussion

This chapter reports the results of the analysis process and thereby provides the basis to answer the main research questions raised in this thesis. The results are organised such that it begins with a network layer level view of the data and subsequently to the application layer protocols and deeper into the HTTP headers. At the end of each section, a discussion on the results observed is provided from a methodological point of view.

First, a high level view of the basic characteristics of the network traffic is provided. Here the study of the general characteristics of the traffic flows is shown. For this, the share of TCP and UDP protocols, symmetry of the traffic and basic diurnal patterns are presented. This is used to test if the results from the Tstat measurement agree with other previous research done.

The result of the traffic classification done by the Tstat DPI is then analysed. This provides the basis for understanding the type of applications that are popular as well as how they are being used. These two results are then compared to make an assessment of how much of an improvement was the Tstat DPI over the port-based method previously used.

The shares of PC and handheld devices as well as the identified operating systems (OS) based on mapping the HTTP headers with the flow data if given next. The share of different types of HTTP traffic content is then shown followed by classification of mobile web traffic for PC and handheld devices.

Finally, the results of the handset identification based on the UA mapping is presented with a brief look at the insights this provides about the device population.

One important thing to keep in mind is that the results presented are based on measurements carried out in the year 2010 and reflect the Mobile Internet traffic during that time. There has been a huge growth in the handheld device market since then. The methodology used here will however allow getting similar results for more recent measurements and form a basis to observe the trend in the traffic usage.

5.1 General Traffic Characteristics

The first step in understanding the network traffic is to inspect its basic characteristics so as to get an early indication of how to proceed with the analysis and if the data makes sense. A good way of doing this is by looking at the symmetry of the traffic.

5.1.1 Traffic Symmetry

The downlink traffic constituted roughly 83% of all the bytes, which shows it is highly asymmetric. This is an increase from previous years which were seen to be 75% (2008), 63% (2007), 73% (2006) and near the level of 85% in the 2005 measurements (Riikonen 2009). This dominance of downlink traffic is inherent in the way the modern 3G networks are designed, which offer more bandwidth in the downlink direction. One of the main reasons for an increase in the share from previous years may be attributed to the large share of P2P traffic (esp. Bittorrent, to be shown in the next section) that was seen. Also, increase in the consumption of online media such as videos, music and photos can be said to have contributed to this increase.

Table 3 shows the same distribution on the two transport layers' protocols detected. The TCP protocol, which was the main protocol in the flows captured (83% by volume) also dominated the downlink traffic. In the uplink side it was more balanced but TCP was still around 2/3rd of the bytes. This may be due to the fact that the UDP protocol is used by online telephony services such as Skype which are inherently more symmetric than normal web surfing and other applications. Also, it is to be noted that the share of total UDP traffic is increased to 17% from previous year's 5% (2008) . This may be due to the increase in popularity of online telephony services as well as online video and audio streaming services. Another factor may be the large share of P2P traffic which use UDP for their signalling and controlling operations.

Table 3: *Share of Flows*

	Downlink	Uplink
TCP	85.74%	67.08%
UDP	14.26%	32.92%

5.1.2 Diurnal Pattern of Traffic Flow

Understanding the daily usage pattern of the network traffic is important from the point of view of the network operators. This provides them with the knowledge of how the traffic flow changes throughout the day and helps in provisioning the resources accordingly and avoid bottlenecks during the busy hours. The daily downlink and uplink traffic pattern is shown in Figure 12. Here, the entire traffic flow was aggregated into a single 24-hour period. This can be seen to be representative of the average diurnal traffic flow.

Previous studies have found that the traffic activity is fairly small during the late night and morning hours. There is then a steadily increasing as the day progresses reaching the period of relatively high activity from early evening till midnight. The traffic then decreases gradually during the late hours. This was indeed the case in this study as well.

The downlink traffic showed great variation in activity as explained above while the uplink traffic was fairly constant throughout the day. The period of high traffic, which can be called *"The Internet Primetime"* was between 6 P.M till 10 P.M. Also, the *"Busy Hour"* for the traffic was seen to be around 9 P.M in the evening.

From this overall pattern a steady increase in traffic was seen around the morning hours between 8 A.M to 10 A.M. This may be because many people are reading the news and emails to get started for the day and most are also commuting to work or school. There was relatively slower growth during lunch hours and it went up again. The peak was reached around 9 P.M in the evening when most people are home and most actively using the Internet. It then went down steadily as people usually go to sleep at night.

The uplink traffic was generally the same throughout the day. This may be because the normal users are normally consumers of Internet services and there are only few scenarios like photo and video sharing etc. when they are themselves uploading any data. P2P applications like Bittorrent use the user's computers to share the files they're downloading with other users in the P2P network constitute the bulk of this uplink traffic. As shown in section 5.2.3 their traffic volume was relatively steady throughout the day which is in turn reflected in the uplink traffic pattern.

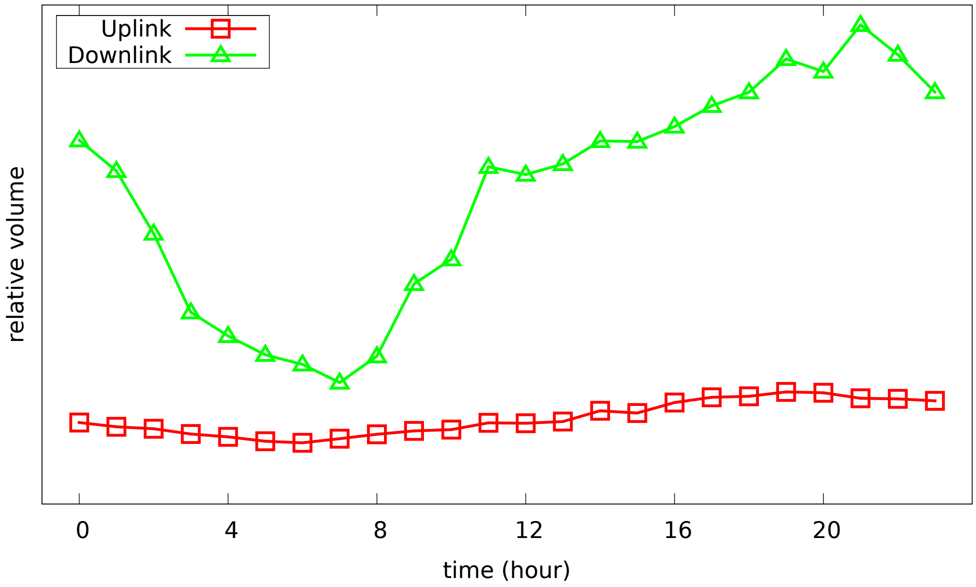


Figure 12: *Distribution of Traffic by Time of Day*

5.1.3 Discussion

The initial results showed that the Tstat TCP and UDP logs conform to the expected characteristics of the uplink and downlink traffic. Both the logs have clearly defined fields which enable us to determine the traffic direction accurately. It is generally expected that the mobile traffic is dominated by the downlink flows. This is indeed what was observed from the results.

The diurnal characteristics of the traffic can also be obtained by utilising the timestamp field in the dataset along with the traffic direction. Thus, the analysis methodology established was quite successful in showing the general traffic characteristics from the dataset available.

5.2 Application Usage Study

In this section the Tstat DPI results were utilised to observe the most popular applications in the Mobile Internet traffic. After going through the application share of different protocols, a comparison was made with the traditional port mapping method. The application usage pattern was then analysed to study how these applications are used.

5.2.1 Share of Application Categories

Table 4 shows the share of the application types identified both in terms of number of bytes as well as the flow count. The Tstat DPI engine was seen to be successful in identifying most of these applications as only around 19% of the flows could not be identified.

Various studies around the world for traditional fixed-line “residential” Internet traffic have shown that after a period where P2P file sharing was the dominant protocol, HTTP is making a comeback in recent years. This is due to the popularity of online video and streaming services. For Mobile Internet, the results obtained from previous years in this project show HTTP traffic to be the dominating with about 40% of the traffic in 2008, while traffic identified as P2P had only 1% of the traffic volume share. However, a large portion of the traffic was unidentified which can be attributed to P2P protocols that cannot be detected by simple port mapping methods employed in those studies.

The current result also showed HTTP to be the most popular application layer protocol with 46% of the total traffic volume. The SSL protocol can also be considered HTTP as it is used for secure authentication purposes (HTTPS). Among the P2P protocols Bittorrent was seen to be the dominant protocol with 35% of the traffic

Table 4: *Application Distribution (By Volume and Flow count)*

Applcation	Bytes	Application	Flows
HTTP	45.99%	Bittorrent	54.02%
Bittorrent	35.02%	Unknown	18.56%
Unknown	10.52%	HTTP	13.72%
Streaming	3.36%	Other	10.79%
Other P2P	2.19%	SSL	1.04%
SSL	1.55%	Other P2P	1.01%
Skype	1.03%	Streaming	0.49%
Other	0.17%	Partial P2P	0.24%
Partial P2P	0.10%	Skype	0.09%
Email	0.05%	Email	0.03%
Messaging	0.01%	Messaging	0.01%

volume. It had over 50% of the total flows with HTTP constituting only 14% of the flows. This was an early indication that most of the traffic was from PC devices using 3G cards. Other P2P protocols like eMule, Gnutella, Ares etc. made up only just over 2% of the traffic by volume and 1% of the flows. P2P applications are also known to use the HTTP protocol to hide its usage. Such cases were found in the Tstat results as well which makes bulk of the flows identified as partial P2P. But its share was too small to change the overall shares significantly. The unidentified traffic might as well be P2P but no conclusions can be made regarding that.

The streaming protocols like RTP, RTCP, Sopcast etc. were the other significant contributors to the total traffic with over 3% of the total traffic volume. The popularity of such services is increasing rapidly and this share can increase significantly in the near future. The Tstat tool was able to detect Skype traffic as well which is a very popular Voice-over-IP (VoIP) applications with 1% of the traffic volume. Many people have been using this application in place of traditional voice calls already because of it cheap rates and ever improving quality. The presence of this application was not seen in the handheld traffic of the current measurements but now with the availability of the Skype application for most popular mobile platforms its share is also bound to change. Messaging and Email protocols constitute very small shares as most of them are transmitted on top of HTTP these days.

5.2.2 Comparison of DPI and Port-based Results

A major objective of this thesis was to study how does the use of DPI based method improve the overall analysis. However, there is an overhead in using such methods which needs to be justified. An indication that this new approach is indeed beneficial has already been given by the results of the application category shares even though it still cannot fully classify all the traffic. To further corroborate this a method described in section 4.2.2 was used in this thesis. The results of this analysis and the conclusions derived from it is now shown in Table 5.

Table 5: *Comparison between Tstat and Port-based Application Identification*

Application	V_{PD}/V_{DPI}	V_{PD}/ V_{Port}
HTTP	98.44%	96.91%
Bittorrent	0.32%	90.99%
Streaming	51.54%	30.50%
Other P2P	1.00%	3.16%
SSL	85.95%	87.91%
Email	90.64%	27.53%
Messaging	80.99%	76.43%

From this we saw that for traffic classified as HTTP, a significant proportion (98.44%) of the traffic was indeed on the default port 80. Also, 96.91% of the traffic in port 80 was HTTP. The remaining are P2P and other protocols running over HTTP. This means that for HTTP the port based method would also have classified it mostly correctly. The results are however very different for Bittorrent and Other P2P applications. Only a small fraction of the identified P2P traffic was seen to use the default ports. But Bittorrent was different from Other P2P as around 90% of the traffic in the default ports of Bittorrent was indeed Bittorrent while only around 3% of the traffic in the default ports of P2P were what they are supposed to be. The ineffectiveness of port mapping for P2P was clearly evident from this. A surprising observation was that for email only 27% of the traffic in the default ports were indeed email. This can also be attributed to the P2P protocols using dynamic ports and also the popularity of web-based email services like gmail and hotmail.

5.2.3 Daily Usage Characteristics

Having been through the share of the application categories some analysis on their nature of usage was carried out. This provides a deeper understanding into the significant trends in mobile Internet usage by representing the entire dataset into a single aggregated 24-hour period.

Diurnal Pattern of the Application Shares

The first plot shows the results obtained from the study of time-of-day effect on application usage. This was done by getting the diurnal pattern of the application shares. Figure 13 shows this trend.

It can be seen that the share of the different protocols remained steady throughout most of the day with HTTP occupying around 50% of the total traffic volume and Bittorrent around 35%. However, it was seen that Bittorrent dominated in the late night to early morning hours. This happened between 1 A.M to 9 A.M when the share of Bittorrent was seen to be higher than that of HTTP web traffic. This share gradually increased from 1 A.M onwards reaching its maximum of 61% at 5 A.M in the morning during which time HTTP traffic is at its lowest of 23.5%. As the day broke the share of HTTP started going back to normal and again exceeded Bittorrent after 9 A.M until midnight.

This behaviour of the application protocols can be attributed to the fact that Bittorrent and Other P2P traffic do not require active user participation as in web browsing. It is normally the case that the user is downloading files that require few minutes to several hours to download and leaves the download open throughout the night when he/she is not using the Internet for other purposes.

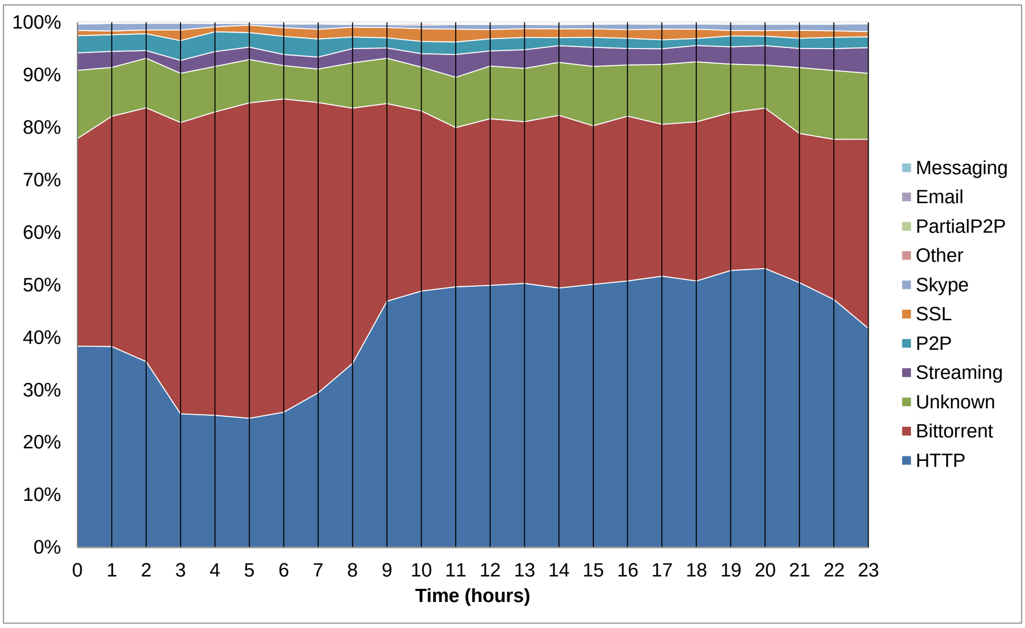


Figure 13: *Application Traffic Volume Share Percentage by Time of day*

Volatility and PA Ratio

As explained in Chapter 4 another way to understand the traffic behaviour of applications is provided by *Volatility* and *Peak-Hour-to-Average-Hour (PA) Ratio*. Table 6 provides a results of this analysis carried out in the current measurements.

It can be inferred from the table that use of communication protocols like messaging, email, Skype and streaming services were the most spread out during the day. That means their usage varied greatly depending upon the time of the day as compared to their average usage throughout the whole day. This is because unlike P2P file sharing which consume around the same amount of bandwidth whenever they are used the communication protocols show rather big fluctuations in traffic volume within a short time period. For example, when someone is chatting or sending/receiving emails the amount of bandwidth consumed is markedly higher than when the application is not in active use. Since, P2P applications mostly run without active participation of the user they have the least fluctuation in their traffic volumes. The same goes for $\mathrm{PA}_{\mathrm{Ratio}}$ which represents the difference in bandwidth consumption between the busy hour and the average hour. It is also seen that streaming, HTTP and Skype traffic had fairly similar spread of traffic throughout the day.

Table 6: *Volatility and Peak to Average Hour (PA) Ratio by Application Category*

Application	Volatility	$\mathrm{PA}_{\mathrm{Ratio}}$
Messaging	79.66%	2.70
Email	70.67%	2.23
Streaming	55.12%	1.85
Skype	52.93%	1.93
HTTP	50.64%	1.64
SSL	46.32%	1.84
Partial P2P	36.11%	1.55
Other P2P	25.48%	1.44
Bittorrent	20.00%	1.29

5.2.4 Discussion

The results of application usage study showed that the Tstat DPI tool indeed provide improvements in classifying Mobile Internet traffic. The fact that the Tstat tool cannot only identify a protocol to be P2P but can also show which P2P protocol is also a major benefit.

There were no major surprises in the share of the identified applications. Traditional web browsing was still the dominant application and lately more and more services are now being run over the HTTP protocol. File sharing over Bittorrent and Other P2P protocols also seem to be very popular among the population as their use is not strictly prohibited like in some other countries. The analysis done to compare these results directly with the port-based method also confirms that the Tstat traffic classifier does a much better job of identifying the applications in most cases.

In addition, the results obtained from the daily usage analysis proved highly effective in giving us a good overall view on application patterns. It was interesting to see that even though in terms of volume share the P2P traffic showed most variation throughout the day, when looking at the spread of bandwidth usage it had the least variation. Also, the communication protocols (online chat and email) that have the lowest share of traffic volume showed bursty nature of usage during the day. Introduction of the two new metrics (volatility and PA ratio) have enabled us to have this kind of useful observations.

Hence, it is seen that the Tstat measurements can be relied upon for traffic classification as it is based on highly advance and proven algorithms.

5.3 HTTP Protocol Usage

The focus of the analysis now shifts on the most dominant application protocol seen in our dataset, which was the HTTP protocol. First the share of the PC and handheld devices is given. It then goes deeper into understanding this by trying to answer questions such as: *What are the major types of content being delivered over the web?, How does this differ in traffic from handheld devices?, What are the most popular website categories and how is this different between desktop computers and handheld devices?, Where is this traffic coming from?*

5.3.1 The OS Mix

The separation of PC and handheld based devices was carried out using the *UA* string and in doing so different OS types identified for HTTP Traffic. This is shown in Figure 14.

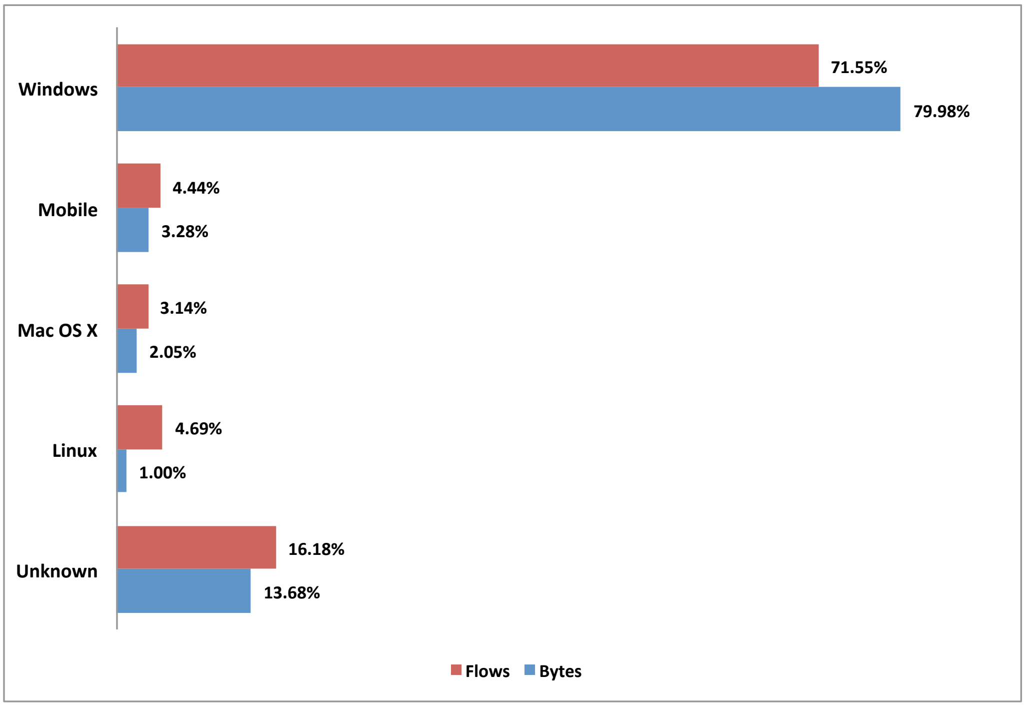


Figure 14: *Share of OS in HTTP Traffic*

Traditional desktop based OS types dominated the HTTP (Web) traffic with around 80% of the traffic flows and 85% of the total volume. The Windows family of operating systems alone constituted around 80% of the total HTTP traffic volume. In terms of flows, around 4% of the traffic generated by handheld devices consumed 3% of the traffic volume. Here about 16% of the HTTP traffic flows remained unidentified.

It is also worth noting that this is only the HTTP traffic (46% share by volume of total traffic) and not the entire traffic flow measured. The 2008 study by (Riikonen 2009) which took into account all the TCP and UDP flows for OS identification showed traditional computer generated traffic to be 98.5% with only 0.6% belonging to handheld traffic. In the current analysis this can indeed be the case if we consider that other protocols like the dominant Bittorrent and other P2P file sharing traffic which are traditionally used extensively in traditional desktop OS but not very prevalent in handheld traffic (Maier et al. 2010). The handheld traffic is usually heavily dominated by HTTP traffic. Thus, it is reasonable to say that the share of the desktop OS will be more closer to the previous studies when we consider the entire traffic.

However, this thesis is limited to OS identification on HTTP traffic and so only provides results for this protocol. The remaining analysis carried out in this section and the next will utilise this OS information thus obtained.

5.3.2 Content Type Distribution

The first step in understanding the nature of the HTTP traffic was to analyze its content mix. This was obtained from the *Content-Type* field of the HTTP headers. Table 7 shows the distribution of the content type by volume for both the PC device traffic as well as for the handheld traffic. The shares are shown by volume and not by the number of requests for these contents because of the nature of webpages. For example, a webpage can have a very large number of smaller image objects but only few video objects. Hence, counting the number of these objects doesn't provide a true indication of what type of objects consume the highest volumes in the webpage.

The PC device HTTP content was dominated by video which accounts for almost 36.5% of the volume. Most of this is from video sharing sites like youtube.com and other news sites that provide videos of the news stories. Most of this video content is flash-based. Given the popularity of online video and thats it is bandwidth intensive the result is as expected. This is then followed by images and HTML, which are the basic content that make up any website. The next significant content type are applications that represent different binary files, documents, online forms etc. that are requested or embedded with the websites. Flash content that is not video was about 3% of this traffic. These may be flash advertisements and flash based games that are highly prevalent in the web. Around 0.3% of the objects did not contain any Content-Type.

Table 7: *Content Type Mix of HTTP (PC Devices)*

Content Type	Bytes
web/video	36.47%
web/image	19.71%
web/html	17.56%
web/application	12.85%
web/javascript	4.86%
web/flash	3.17%
web/css	2.45%
web/audio	1.47%
web/others	0.52%
web/xml	0.44%
NA	0.27%
web/json	0.16%

The content types of handheld traffic were also studied to get a better understanding of what the users of such devices are consuming. This is shown in Table 8. As shown the results were not very different than the PC devices. Video was again the largest contributor to overall HTTP traffic with almost 28% share, followed closely by images. Another interesting observation is that the share of flash content was almost half as compared to the share in the overall traffic. This is because flash is not supported by many mobile devices, especially Apple iOS devices. Around 0.16% of the objects did not contain any Content-Type headers.

Table 8: *Content Type Mix of HTTP (Handheld Traffic)*

Content Type	Bytes
web/video	27.67%
web/image	26.88%
web/application	14.14%
web/html	9.39%
web/javascript	7.22%
web/plainttext	4.44%
web/css	3.62%
web/xml	1.92%
web/audio	1.92%
web/flash	1.43%
web/json	0.89%
web/others	0.28%
NA	0.16%

5.3.3 Web Browsing

After analysing the Content-Type, the distribution of traffic in various websites is now presented. For this the websites were differentiated into categories of websites and their popularity studied. The websites visited by PC based HTTP traffic and handheld traffic were separated to give a better understanding of the difference in usage of the two different types of platforms. Also, as seen from the previous section, HTTP traffic contains all type of traffic such as images, videos, advertising etc embedded in the web page along with the main HTML page. Only the objects of type *web/html* can be the websites and only those are considered in this section.In both PC and mobile based traffic, only the top 100 websites were considered for categorisation. Thus, all the shares are for the websites among the top 100 websites.

For PC Devices

Table 9: *Category Breakdown of Top 100 Web Hosts - PC devices*

Site Category	Bytes	Site Category	Flows
Video Sharing	33.08%	Online Advertising	34.11%
Online Entertainment	17.76%	Social Networking	21.31%
News Portal	6.87%	Online Entertainment	15.13%
Social Networking	6.84%	Search Engine	12.10%
Windows Update	6.53%	News Portal	5.12%
Online Gaming	6.24%	Online Shopping	4.11%
Online Advertising	6.08%	Video Sharing	2.88%
Software Update	3.81%	Adult Site	1.41%
Antivirus Update	3.17%	Online Gaming	1.15%
Adult Site	2.51%	Online Portal	1.12%
Online Portal	2.25%	Windows Update	0.82%
Online App Store	2.21%	Antivirus Update	0.64%
Online Shopping	1.40%	Software Update	0.06%
Search Engine	1.25%	Online App Store	0.04%

Figure 9 shows the most popular website categories for PC device based HTTP traffic. The share of the website categories is given both in terms of byte volume and the number of flows for that particular category. Video sharing sites consumed most bytes as expected, even though it does not constitute a lot of flows. Of these sites youtube was by far the most popular video sharing website. Other video rich sites such as the online entertainment portals of popular TV networks like mtv3 and online news portals like hs.fi also had a significant share of the traffic volume. The use of search engine like google was also pretty high though they consume very less bandwidth.

It is interesting to see that Online Advertisement had most number of domain requests. This was probably because of the fact that most websites carry these advertisements these days. Social networking as expected had very high number of requests as Facebook is very popular. In terms of bytes consumed it was not as high as Facebook is primarily a photo and status sharing site and video content is mostly through links to external sites.

Antivirus and other software updates also denoted a significant share of the volume though they have very few domain requests. Another interesting fact seen was that adult websites made up a low share of domain requests among the top 100 domains. Online gaming was also seen to be gaining popularity and is a major portion of the traffic volume.

For Handheld Devices

Table 10 shows that handheld internet traffic was quite different from that of traditional computers.

Table 10: *Category Breakdown of Top 100 Web Hosts - Handheld devices*

Site Category	Bytes	Site Category	Flows
News Portal	18.61%	Online Advertising	26.08%
Online Entertainment	16.39%	Social Networking	20.52%
Adult Site	15.67%	News Portal	14.76%
Search Engine	12.09%	Search Engine	12.36%
Online Repository	9.74%	Adult Site	8.53%
Online Advertising	6.45%	Online Entertainment	5.40%
Social Networking	4.92%	Online Portal	4.99%
Online App Store	4.75%	Online App Store	2.94%
Online Portal	3.9%	Video Sharing	1.40%
Video Sharing	2.23%	Online Shopping	1.34%
Software Update	2.11%	Financial News	1.12%
Financial News	1.63%	Online Repository	0.37%
Online Shopping	1.51%	Software Update	0.19%

The results showed that news portals like iltalehti.fi and hs.fi were read extensively on handheld devices. Most of these websites are mobile optimised. This may be because most people use the time they are commuting to work to read the news on their handheld devices. Online advertisement was again the source of most domain requests followed by social networks. This once again shows that social networks like Facebook are highly popular on mobile devices as well. It is seen that search was also very popular on handheld devices.

Rather surprisingly the consumption of Adult Sites was higher in handheld devices as compared to PC devices. People also seem to be using their handhelds to check the weather and get information about public transportation timings through online portals. Video sharing and online video streaming websites weren't seen to be very popular. Online shopping was also not very popular with handheld devices as compared to PC devices. This is understandable as people are still weary of security issues in handheld devices.

In terms of volume shares Online Repository was also quite significant. This is because the Nokia N900 model has a major share among the users of this network as will be shown in section 5.4. It online Maemo repository seems to be rather popular creating some significant traffic. This device is mostly popular among the more tech-savvy users who are also heavy users of the data services. This may explain its high share. There was also some traffic requests going to unknown IP addresses which are categorised under "Others".

5.3.4 Discussion

The HTTP header information gathered in the measurements has shown to contain very rich set of information that can be used to study the mobile web traffic in good detail.

The *UA* field was successful in identifying most of the traffic. The 16% unidentified traffic represents those *UA* strings that did not follow any standard or well-known format which would allow extracting the OS information from it. The remaining traffic had standard string formats from which the OS information can be reliably obtained. Hence, this allowed the separation of the PC and handheld traffic.

Only the HTTP header was used for OS identification and the OS distribution was done only for HTTP. This can be said to be a step back from previous studies that used TCP fingerprinting (Riikonen 2009). However, since this thesis aims to study the measurements from the Tstat tool, this has been accepted as a necessary compromise to focus on the potential the *UA* field has in device identification.

The comparison done between the content mix of the PC and handheld device HTTP traffic showed that the *Content-Type* field is very useful and mostly available. It is also to be noted that this information is not available for all the HTTP Responses in the header files as the browsers and web servers may be configured to hide this information. Though, this was quite low in this study. Furthermore, studies like Schneider et al. (2012) have shown that the information given in the *Content-Type* field is not always accurate.

Similarly, the *Host* field contains information about the websites visited. Using this and considering only HTTP requests of type *web/html*, the categorising the 100 most popular websites was done and the difference in online services consumed in these two types of devices was seen. People seem to use desktop OS based devices to consume rich media like videos whereas the handheld devices are used more to consumer news and do search for local information. There was no distinction made here between user HTTP request made by the user and those generated in the background.

Hence, using only three of the many available header fields, it has become possible to gain great insight in the difference between PC devices and handheld devices.

5.4 Handset Population

One of the main focus of this thesis has been to observe which handheld devices are consuming HTTP traffic. The results in the previous section show that they consumed around 3% of the HTTP bandwidth. This provides us with valuable information in getting a more holistic view of mobile Internet usage. The results obtained using the device identification method explained in section 4.6 is presented now. For the purpose of this thesis a look at the share of the major platforms and brands is shown. This provides a way to establish the usability of this device identification method and also get an idea of how the market it evolving after the rapid rise in popularity of smartphones. It is to be noted that all the results in this section are only for the HTTP traffic of handheld devices.

5.4.1 Brand Popularity

Finland is home to the biggest maker of mobile handheld devices in the world which is Nokia. Hence, it has always been the dominant brand in the Finnish market. Studies carried out in the MOMI project itself have shown this to be the case as well (Riikonen 2011). Table 11 shows the results for the most popular handset brands in our measured dataset in terms of number of flows and data volume.

Table 11: *Share of Mobile Brands*

Brand	Bytes	Brand	Flows
Nokia	44.48%	Nokia	45.62%
Apple	21.35%	Samsung	13.74%
Samsung	12.26%	Apple	10.87%
HTC	6.05%	HTC	10.09%
LG	5.81%	LG	7.52%
Unidentified	4.27%	Sony Ericsson	4.55%
Generic Android	2.48%	Unidentified	3.17%
Sony Ericsson	2.19%	ZTE	1.96%
ZTE	1.10%	Windows Mobile	1.51%
Other	˜2%	Others	˜2%

The results are quite interesting because although Nokia was the dominant brand both in terms of share of volume and the number of flows, it was not as big as expected. Of the devices identified correctly around 44% of them were Nokia devices generating around 46% of the traffic. Given that it was reported that around 90% of the handsets in Finland are Nokia phones in 2010 (Riikonen 2011), the share among devices using the mobile Internet in this network is low. Samsung was the next most popular brand with 14% of the flows. Apple that makes the popular iPhone and iPad devices was the next most popular brand with 11%, followed closely by HTC. It can be seen that Apple devices consume more bandwidth than Samsung and HTC, both of which predominantly use the Android OS.

There were also around 3% of the devices whose brand could not be identified from their UA strings, mostly because of the use of non-standard UAs. There are some UAs identified as Generic Android devices because they only consist Android OS information.

A look at the distribution of the device OS in the handheld device also agrees with the above observation. This is shown in Figure 15.

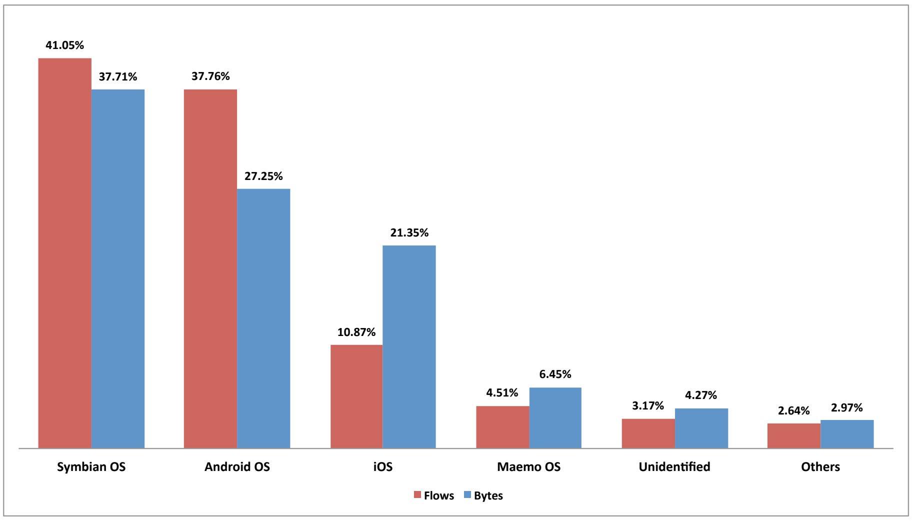


Figure 15: *Distribution of OS among Handheld traffic*

Most of the devices identified were Symbian devices which was the OS used by Nokia. Android was second followed by iOS which powers Apple devices. Again it is seen that despite only 10% of the flows coming from iOS devices, it consumed 21% of the traffic volume. The Maemo operating system was present is around 4% of the devices. Since, there is only one handheld device in the market that uses Maemo which is the Nokia N900, it can be said that all of its traffic is due to this one device. This is indeed the case as seen in the next section.

5.4.2 Which Devices Generate the Most Traffic?

Having gone through which is the most popular device manufacturing brand the top mobile handset models are now shown in Table 12. The share of the top 15 models out of over 300 devices identified correctly is given. We don't consider Unidentified devices and devices identified as *Generic* here.

Table 12: *Share of Top 15 Device Models (of Devices Identified)*

Model	Bytes	Model	Flows
Apple iPhone	18.92%	Samsung Galaxy S	11.69%
Nokia 5230	7.29%	Apple iPhone	8.35%
Nokia N8	6.70%	Nokia 5230	7.88%
Nokia N900	6.45%	Nokia N8	7.74%
Samsung Galaxy S	6.05%	Nokia N900	4.51%
HTC Desire	5.78%	HTC Desire	4.34%
LG Apex	3.60%	LG Apex	3.48%
Nokia N97	2.77%	Nokia N97	2.78%
Nokia N97 mini	2.63%	Nokia N97 mini	2.64%
Nokia 5800 XpressMusic	2.47%	Nokia C6	2.62%
Apple iPad	2.43%	Apple iPad	2.52%
HTC P3600	2.35%	Nokia E71	2.16%
Nokia C6	2.22%	Nokia E75	1.89%
Nokia 2605	2.12%	HTC Hero	1.64%
Nokia X6	1.80%	Nokia 5800 XpressMusic	1.58%

Though the bulk of the models were Nokia phones, the most popular model seemed to be the Samsung Galaxy S. It was then followed by Apple iPhone, Nokia 5230 and Nokia N8. It can be seen that the iPhone consumes by far the most bandwidth among all the identified devices. Infact, more than three times that of the Samsung Galaxy S. This is in line with the results of many market researches which say that iPhone users are the heaviest users of mobile Internet services, even back in 2010.

Among other handset manufacturers HTC and LG handsets had good share of the market. Nokia 5230 was the most popular Nokia phone and the Nokia N8 was also seen to be amongst the top handsets. A tablet device, which is the Apple iPad also featured among the most popular devices for Mobile Internet usage though it got officially launched in Finland only after the measurements were carried out.

One observation regarding the device population is that touchscreen was a very important feature in smartphones. Of the HTTP requests observed in the network around 70% of these requests were generated by handhelds that have some variation of the touchscreen (with or without a keyboard).

5.4.3 Mobile Application Usage

As mentioned in Chapter 4, utilising the observations made about the generic patterns in *UA strings* for mobile browsers and applications, a study of mobile applications can be made. Although not the main focus of this thesis, an initial analysis carried out to see the viability of this method is done here. Figure 16 shows the share by volume of the identified browsers and applications.

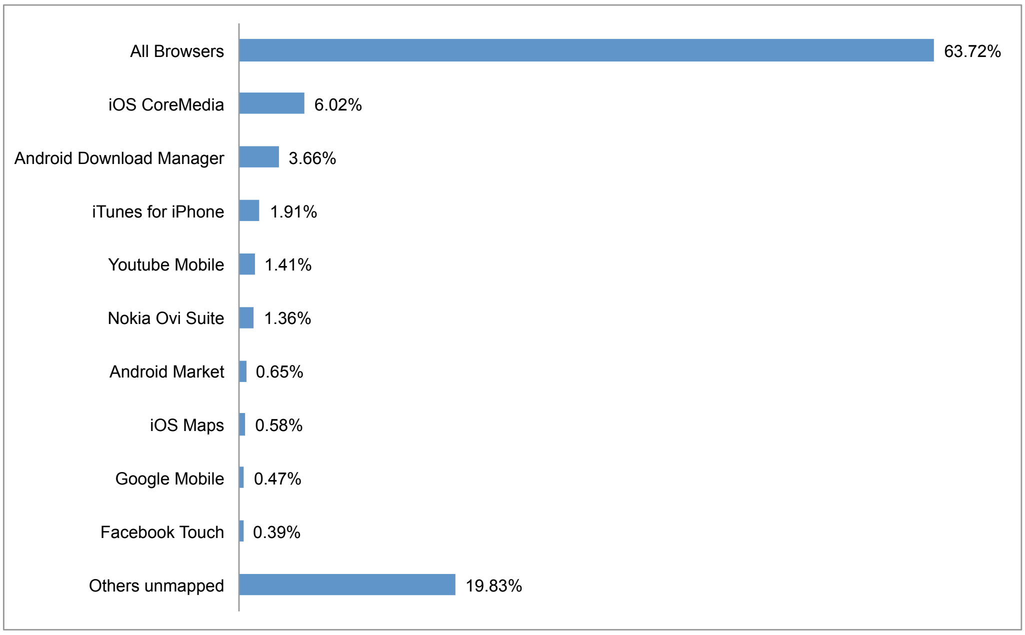


Figure 16: *Share of the 10 most popular applications (by Volume)*

All the browsers were considered together while the top 10 most popular application UAs were mapped manually to get their share of the traffic.

The browser was by far most popular application in handheld devices with around 64% of the total traffic volume. A closer look at these mobile browsers revealed Mobile Safari which is the default browser in Apple's iPhone, iPod Touch and iPad to be the most popular browser by volume with around 33% share. This was followed by the Android browser, the Maemo browser and then the Nokia browser. The Opera Browser which is a third-party found in almost all the major platforms was also quite popular.

Among other applications, iOS CoreMedia which was used by the media players in Apple's devices had the largest share of 6%. iTunes for iPhone also produced significant traffic, mostly because of the popular iTunes Store. The other popular applications after this were mostly Android based. Some traffic was also seen from Nokia's Ovi Store and Android Market. The Maps applications in Apple's devices was also seen to be popular along with the touch optimised Facebook application. The unmapped UAs represent the remaining traffic.

5.4.4 Discussion

The results obtained from the device identification process using the WURFL database were seen to be quite effective. This method was able to identify the UA from more than 90% of the HTTP traffic flow. However, the exact handset model was not possible for many cases which are termed as Generic by the WURFL engine. There are UAs from which only partial information like the OS or manufacturer could be observed but not the exact phone model.

With around 500 device features available and a subset (Appendix A) of the most useful of those features shown, this brings a lot of possibilities for research. As a showcase of this, the top device manufacturing brands, device models and device OS were given. These results clearly showed the existing domination of Nokia brand in Finland but also the rising popularity of iPhone and Android devices in 2010.

The most interesting result is that despite not being the top device by HTTP request flows, the Apple iPhone consumed the most data volume in the mobile network. It is to be noted that the different versions of iPhones are not differentiated here and taken as one iPhone device. The iPad tablet from Apple was also seen in the dataset.

Additionally, an analysis on how the UA string can be used to identify the browsers and applications in mobile handsets was done. Despite the popularity of applications in mobile devices, the browser was the dominant method to access the web. The top platform and add-on applications are shown. Manually mapping all the UAs is outside the scope of this thesis.

6 Conclusion

This chapter concludes the thesis by summarising the major findings made from the results of the research. Some discussion in then made regarding implications of these findings and finally some recommendations are made to the stakeholders involved on how the findings may be utilised and possible future research areas are pointed out.

6.1 Summary

This thesis aims to create a framework to carry out holistic understanding of Mobile Internet usage using network measurement data. This was carried out using a modified version Tstat traffic classifier tool to additionally capture HTTP header information in a mobile cellular network for a period of one week in 2010. Since, this is the first time such a measurement setup is used in a cellular network, the thesis mainly focusses on verifying the reliability of the new dataset, comparing the results with those from previous methods, exploring the possibilities the new dataset provides to study various aspects of the Mobile Internet traffic and also developing a method to use the HTTP headers to identify mobile handsets and their features.

For the purpose of fulfilling the objectives, the analysis went through several stages. This began by looking at the different types of datasets provided by the Tstat measurement and assessing ways to perform a sanity checks on the data. Looking at the general traffic characteristics, the traffic was seen to be asymmetric in the downlink direction as expected. There was an increase in the downlink traffic (85%) as compared to the previous study (75%) which may generally be attributed to increase in online streaming media traffic. TCP was still the dominant transport layer protocol with 83% of the traffic but the UDP share had increased, possibly due to increase in online telephony services. Looking at the aggregated 24-hour usage traffic pattern showed increased usage during the morning hours, with more or less same usage during the daytime. The period of highest traffic was during the evening peaking at 9 P.M. and finally decreasing sharply after midnight.

The Tstat tool is capable of classifying the network level traffic looking at signatures in the pattern of the traffic. This formed the basis of the next analysis done. HTTP had the largest share of the traffic in the network with 46% of the total traffic volume. The Bittorrent P2P protocol was dominant in terms of number of flows with around 54% of the traffic, while Other P2P were not as significant. The tool was unable to classify around 19% of the traffic. Direct comparison with the port-based method showed how much more effective the Tstat tool is in correctly identifying the application, especially for P2P based traffic. The results for daily usage characteristics showed how different applications were used throughout the day. HTTP traffic was generated most significantly during the *waking hours* since mostly people may be browsing. The rise in P2P traffic during the *late hours* of the day indicate that most of the traffic may be PC based. The Volatility and PA ratio results show that messaging, email and streaming usage was spread in shorter intervals of use throughout the day, while HTTP and P2P usage remains more evenly spread out.

The HTTP headers obtained were then analysed to see what kind of information we can get from them regarding web usage. The OS information obtained from the *UA* field were used to identify the OS in around 84% of the HTTP data volume. Desktop based OS was dominant with 80% of the traffic flows while 4% was from handheld devices. Some comparisons were then carried out between the PC based and handheld based Internet traffic. For PC devices, video content was seen to consume most bandwidth with 36% of the share. The same was also seen for the handheld traffic, although the share of the traffic volume was more comparable to images. The category breakdown of the top 100 websites show that video heavy websites like youtube, vimeo and online entertainment and news sites that have a lot of video content dominate the bandwidth consumption. Social Networking sites were very popular and had very high traffic flows. For handheld devices however, video sharing sites weren't as popular as in PC, although News Portal and Online Entertainment websites had most of the traffic volume. Social Networking sites like Facebook were again seen to be very popular in handheld devices as well.

Hence, these results showed that the Tstat traffic classifier contained a rich set of data that can be reliably used to carry out Mobile Internet usage studies. It was also more accurate in identifying the various application level protocols of modern day Internet traffic as compared to the port-based method.

The *UA field* in the HTTP headers was then used to identify the mobile handset devices. It was seen that Nokia is the dominating device manufacturer with around 45% of the traffic flow and the Symbian OS used by most of these devices has the top share of the traffic. The growth of the iOS and Android platform can easily be seen here. With many manufacturers like Samsung, HTC, LG etc. using the Android OS in their smartphones they were already close to Nokia with these Android taking 37% of the traffic flow share. Apple devices that run on iOS were also very popular with these devices second only to Nokia in terms of traffic volume consumed. In terms of device models, Samsung Galaxy S and Apple iPhone lead although most of the devices in the study are seen to be Nokia devices. This show that individually they were already more popular than any single Nokia handset. Around 2% of the traffic also came from Apple's iPad tablet. This was surprising because it was not introduced in Finland when the measurements were taken. This indicates people buying these devices from other countries. Finally, a brief analysis carried out to separate browser and application related traffic from the dataset showed that browser traffic constituted more than 63% of the traffic volume. The UAs related to the top 10 handset applications done was also studied to see if the application names can be derived from them. The results showed that this information was indeed available from the UA. However, maintaining the complete mapping of UA to application is outside the scope of this research.

6.2 Implications of the Results

The purpose of this results as stated in Chapter 1 was to first check if the Tstat measurements can be used to perform studies on Mobile Internet traffic reliably and then if it can be, to create the necessary analysis framework to further develop the methods to study this traffic and enrich the understanding of Mobile user behaviour. Identifying the handsets in the network and studying some of the device features was also another major goal.

Looking at the results obtained for the general traffic characteristics, the uplink and downlink symmetry of the network layer protocols and daily usage pattern were in line with what is expected. Studies carried out with residential Internet data from large ISPs(Schneider 2010) also found the same kind of characteristics. This indicates that the measurement data can indeed be relied upon to form the basis for further analysis. Also, the traffic classification algorithm in the Tstat tool was found to be very capable and was a definite improvement over the port-based method. It can successfully identify most of the network level traffic and can even recognise most of the important P2P protocols. The previous method used did not work very well for P2P traffic identification. The results show how popular P2P was, along with other streaming and messaging (text and video) applications. By coming up with a way to directly compare how much better the new tool was in identifying the traffic and creating daily aggregated application usage served as the first step in carrying out the objective of setting up a methodology to gain new insights into the data. The TCP and UDP logs have a lot of information to study the various aspects of the traffic on the network layer.

Another addition to set of measurements available was the HTTP header information. The results from the network layer show HTTP to be dominant protocol for Internet traffic. Not only this, a lot of other Internet protocols are seen to be running on top of HTTP. These header fields can be mapped with the corresponding TCP flow and provide us with accurate information about the data volume and flow distribution for different OSes, content types, websites and handheld devices. The results obtained from this process was significant and gave valuable insights into the current state of Mobile Internet usage in Finland. Even when only the first request/reply pair was considered, this information was adequate for carrying out different studies. It shows that most of the Mobile Internet traffic is dominated by PC devices using 3G cards/dongles, though handheld devices were increasingly being used to access the Internet. Also, the rise of Online Video and Social Networking sites was seen from the results. One important consideration is this measurement was from the year 2010 and this trend is only increasing as time goes by. The methods developed in this thesis can be used to analyse the same kind of data for any future measurements.

The handset population analysis implies that the HTTP header fields (the UA field in particular) can be used with the WURFL tool to identify most of the handheld devices. Despite the fact that there are a lot of malformed and non-standard UAs, it is possible to get correct information on most of the modern handheld devices. Looking at the results several inferences are possible about the state of the Finnish handset market. Even though Nokia was the market leader in terms of devices seen and data usage, Apple's iOS devices (iPhone and iOS) and handsets using Google's Android OS were very popular. The data usage from these two platform was very heavy, especially for iOS. Another observation was that most of the popular handsets are fully touchscreen based or have touchscreen capabilities with a keyboard. The look at browser and mobile application traffic show that even though browser was the primary means of accessing the web services on handheld devices, the apps were increasingly being used.

6.3 Reliability and Validity of the Results

As with any research done, this research also comes with a certain possibility of error. Although great care had been taken during every stage of the research to ensure accuracy and minimisation of errors, several sources of random, technical, tool or human error can affect the results seen. The fact the this was the first time that the Tstat traffic classifier tool and the HTTP headers were used in this project further compounds these chances. Only by repeating these methodology a few times on other dates and operators can it be proven to be truly valid.

Although as discussed the general characteristics of the data available were seen and they make sense, this was only a high level aggregation that was obtained. As a whole it can be said that there are no errors that affect the entire dataset, but nothing can be said about any missing or corrupt data on any subset of the data. It has been mentioned in section 3.3.1, that there are certain errors related to the data captured by the tool. These are taken into consideration during the analysis and only after concluding that the end results will not be greatly changed by this, the analysis is performed.

The traffic classification although shown to be a vast improvement on the previous method cannot be 100% accurate. The results offered no great deviations from what is normally expected to indicate that any classification was significantly faulty. The same can be said of HTTP headers. Although, there are certain standards that these header fields are expected to follow, the browser and server vendors are open to make their own tweaks and changes. So, for both PC and handheld HTTP headers may not have adequate fields available, as many header fields are optional or can be false. The methods used chose the headers which have most of the important header fields available within one single TCP flow to minimise such cases.

The UA field which was the basis of device identification is also another possible source of inaccuracy. Many devices and applications do not follow the standard UA string patterns and provide a challenge to derive information from it. The WURFL database which is the basis of this device identification algorithm may have faults within itself which might cause inaccuracies in identifying the correct device model. The analysis mentions this but the results show that overall the results don't suffer greatly from these issues.

Another factor to consider is that this measurement was carried out in a single GGSN of an operator in Finland for a period of one week. This is by no means a big enough representation of the Mobile Internet population of Finland. The measurement has to be carried out in other operators for a longer period of time to give a more reliable insight into Mobile Internet traffic.

Despite the issues discussed, the results showed that the thesis has been largely successful is showing that this measurement setup and the analysis framework established can indeed provide accurate and repeatable results on Mobile Internet usage.

6.4 Recommendations to Stakeholders

After the discussion on the implication of the results obtained from the research carried out in this thesis, certain recommendation can be made to the stakeholders of the MOMI project as well as other network operators and players in the Mobile Internet market.

For the MOMI project, it can be seen that using the Tstat traffic classifier and HTTP headers have enrich the results. This was the first time this measurement setup was being used and the initial results are indeed promising. In addition to improving the accuracy of the previous methodology, the dataset available can be used to carry out analysis on the handset devices, the web content etc. as shown here. Continuing using this measurement setup and improving the methodology with some of the possible improvements identified will enable all the stakeholders to understand the market and users better.

For example, the time of day analysis showed how the traffic varies over the day and how the applications are being used. The listing of the top website categories and device models also show trends in user behaviour and their changing tastes. This information can be used by operators should upgrade and reassign their resources to give the best network experiences. The increase in data usage from handheld devices show that future networks should be able to handle high increase in data traffic. The popularity of touchscreen devices means that manufacturers and operators should push to introduce more devices with touchscreens. The knowledge about the increasing use of mobile applications as a means to access web-based services may tell the content-providers to focus on making better performing applications and advertisers to focus on application in addition to browsers.

6.5 Future Work

The work done in this thesis was the first exploration of the potential provided by the Tstat traffic classifier tool and the HTTP headers. There is a lot of room to improve the whole process. Some of the possible areas of future improvements to continue with this research are given now.

There were some issues encountered with this measurement. One of the major issues was the missing HTTP data and Tstat histograms. The reason for these data holes need to be identified and corrected to give a complete dataset. Also, the *Status code* header should be added because it tells us if the traffic was from a successful HTTP request or not. This may provide means to increase the accuracy of the HTTP analysis.

The method used to identify the handset devices using WURFL database though fairly effective can be improved further. This thesis used an unofficial python-based API for device identification. However, using the official Java APIs is recommended as it will get long term updates and improvements in the future. Infact, this has already been done in the MOMI project (Adhikari 2012). This already shows good improvement over the initial method established here and should be continued.

Other possible areas of research can be to take the methodology created here and make it more modular and organise it into a proper tool to save manual work and sources of possible human error. Some areas like UA to application name mapping should be done to create a more comprehensive database of the application. The study of websites and website categories was done by making simplification to the nature of HTTP data. It does not differentiate between user generated requests for HTTP web pages and the background requests that are generated such as the images, videos, advertisements etc. embedded in the web page. Also, modern web pages are more complex and provide many client-side interaction features using javascript. So, in order to more closely model todays websites, methods such as that developed in (Ihm & Pai 2011) can be used.

All of the above mentioned results are on an application session level. This means that we can only distinguish one application session from another. However, the RADIUS log provides an opportunity to make the analysis on a per user basis. If it can be established that these radius sessions represent a single user session then the application flows can be mapped to these sessions. Thus, giving us results on the user level itself, where each user may be using several applications. These can be mapped by using the IP address of the device and the corresponding time stamps. Hence, all the application flows from one IP within the time frame of the user session with the same IP can be said to occur within that session. In the current analysis the OS identification is performed only for the HTTP traces. So, we have no information about the OS used by other protocols and indeed a user. The OS for the user can be mapped to the corresponding user session using the IP and time stamp information to get the OS used by the user. Hence, deeper research on the RADIUS data and mapping it with TCP and UDP flows can be done.

References

Adhikari, A. (2012), Mobile Device Identification from Network Traffic Measurements: A HTTP User Agent Based Method, Master's thesis, Department of Communications and Networking, Aalto University.

Bonfiglio, D., Mellia, M., Meo, M., Rossi, D. & Tofanelli, P. (2007), 'Revealing skype traffic: when randomness plays with you', *ACM SIGCOMM Computer Communication Review* **37**(4), 37–48.

Cascarano, N., Ciminiera, L. & Risso, F. (n.d.), 'Accelerating DPI Traffic Classifiers'.

Cisco VNI: Usage Study (2010).

Clegg, R., Withall, M., Moore, A., Phillips, I., Parish, D., Rio, M., Landa, R., Haddadi, H., Kyriakopoulos, K., Augé, J. et al. (2009), 'Challenges in the capture and dissemination of measurements from high-speed networks', *Communications, IET* **3**(6), 957–966.

Cohen, E., Kaplan, H. & Oldham, J. D. (1999), Managing tcp connections under persistent http, *in* 'In Proceedings of the World Wide Web-8 Conference'.

CoralReef homepage (2010), http://www.caida.org/tools/measurement/coralreef/.

Crovella, M. & Krishnamurthy, B. (2006), *Internet measurement: infrastructure, traffic and applications*, John Wiley & Sons, Inc. New York, NY, USA.

Erman, J., Gerber, A. & Sen, S. (2010), HTTP in the home: it is not just about PCs, *in* 'Proceedings of the 2010 ACM SIGCOMM workshop on Home networks', ACM, pp. 43–48.

Falaki, H., Lymberopoulos, D., Mahajan, R., Kandula, S. & Estrin, D. (2010), A first look at traffic on smartphones, *in* 'Proceedings of the 10th ACM SIGCOMM conference on Internet measurement', IMC '10, ACM, New York, NY, USA, pp. 281–287.

FICORA (2010), 'Communications markets in finland, 2010 annual report', http://www.ficora.fi/en/index/tutkimukset/generalmarketinformation.html.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. & Berners-Lee, T. (1999), 'Hypertext transfer protocol – http/1.1'.

Finamore, A., Mellia, M. & Meo, M. (2011), Mining unclassified traffic using automatic clustering techniques, *in* J. Domingo-Pascual, Y. Shavitt & S. Uhlig, eds, 'Traffic Monitoring and Analysis', Vol. 6613 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 150–163.

Finamore, A., Mellia, M., Meo, M., Munafò, M. & Rossi, D. (2010), Live traffic monitoring with tstat: Capabilities and experiences, *in* 'Wired/Wireless Internet Communications', Vol. 6074 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 290–301.

Habuchi, I., Dobashi, S., Tsuji, I. & Iwata, K. (2005), 'Ordinary usage of new media: Internet usage via mobile phone in Japan', *International Journal of Japanese Sociology* **14**(1), 94–108.

Heikkinen, M., Kivi, A. & Verkasalo, H. (2009), 'Measuring mobile peer-to-peer usage: Case Finland 2007', *Passive and Active Network Measurement* pp. 165–174.

Holma, H. & Toskala, A. (2000), *Wcdma for Umts*, Vol. 4, Wiley Online Library.

Ihm, S. & Pai, V. S. (2011), Towards understanding modern web traffic, *in* 'Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference', IMC '11, ACM, New York, NY, USA, pp. 295–312.

Ishii, K. (2004), 'Internet use via mobile phone in Japan', *Telecommunications Policy* **28**(1), 43–58.

ITU (2012), http://www.itu.int/ITU-D/ict/statistics/atglance/KeyTelecom.html—.

Jain, R. & Routhier, S. (2002), 'Packet trains–measurements and a new model for computer network traffic', *Selected Areas in Communications, IEEE Journal on* **4**(6), 986–995.

Kaaranen, H. (2005), *UMTS networks: architecture, mobility, and services*, Wiley.

Kalden, R. (2004), Mobile internet traffic measurement and modeling based on data from commercial GPRS networks, PhD thesis, University of Twente.

Karagiannis, T., Broido, A., Brownlee, N., Claffy, K. & Faloutsos, M. (2005), Is p2p dying or just hiding?[p2p traffic measurement], *in* 'Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE', Vol. 3, IEEE, pp. 1532–1538.

Karagiannis, T., Broido, A. & Faloutsos, M. (2004), Transport layer identification of P2P traffic, *in* 'Proceedings of the 4th ACM SIGCOMM conference on Internet measurement', ACM, pp. 121–134.

Kivi, A. (2007), Diffusion and Usage of Mobile Browsing in Finland 2005-2006, *in* 'Proceedings of the 4th CICT Conference', Citeseer, pp. 29–30.

Kivi, A. (2009), 'Measuring mobile service usage: methods and measurement points', *International Journal of Mobile Communications* **7**(4), 415–435.

Li, W., Moore, A. W. & Canini, M. (2008), Classifying http traffic in the new age, *in* 'ACM SIGCOMM', Vol. 8, pp. 17–22.

Maier, G., Feldmann, A., Paxson, V. & Allman, M. (2009), On dominant characteristics of residential broadband internet traffic, *in* 'IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference', ACM, New York, NY, USA, pp. 90–102.

Maier, G., Schneider, F. & Feldmann, A. (2010), A first look at mobile hand-held device traffic, *in* A. Krishnamurthy & B. Plattner, eds, 'Passive and Active Measurement', Vol. 6032 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 161–170.

McAfee TrustedSource (2010), http://www.trustedsource.org/en/feedback/url.

Moore, A. & Papagiannaki, K. (2005), 'Toward the accurate identification of network applications', *Passive and Active Network Measurement* pp. 41–54.

Postel, J. (1980), 'RFC 768: User Datagram Protocol (UDP)', *ISI* .

Postel, J. (1981*a*), 'RFC 791: Internet protocol'.

Postel, J. (1981*b*), 'RFC 793: Transmission control protocol'.

pywurfl homepage (2010), http://celljam.net/.

Radcom Homepage (2010), http://www.radcom.com/.

Ricciato, F., Svoboda, P., Motz, J., Fleischer, W., Sedlak, M., Karner, M., Pilz, R., Romirer-Maierhofer, P., Hasenleithner, E., Jager, W. et al. (2006), 'Traffic monitoring and analysis in 3G networks: lessons learned from the METAWIN project', *e & i Elektrotechnik und Informationstechnik* **123**(7), 288–296.

Rigney, C. (2000), 'RFC2866: RADIUS Accounting', *RFC Editor United States* .

Rigney, C., Willens, S., Rubens, A. & Simpson, W. (2000), 'Remote authentication dial in user service (RADIUS)'.

Riikonen, A. (2009), Mobile Internet Usage - Network Traffic Measurements, Master's thesis, Department of Communications and Networking, Helsinki University of Technology.

Riikonen, A. (2011), 'Mobile handset population in finland 2005-2010'.

Riley, M. & Scott, B. (2009), 'Deep Packet Inspection: The End of the Internet as we Know it'.

Risso, F., Baldi, M., Morandi, O., Baldini, A. & Monclus, P. (2008), Lightweight, payload-based traffic classification: An experimental evaluation, *in* 'Communications, 2008. ICC'08. IEEE International Conference on', IEEE, pp. 5869–5875.

Salgarelli, L., Gringoli, F. & Karagiannis, T. (2007), 'Comparing traffic classifiers', *ACM SIGCOMM Computer Communication Review* **37**(3), 65–68.

Schneider, F. (2010), Analysis of New Trends in the Web from a Network Perspective, PhD thesis, Technische Universitat Berlin.

Schneider, F., Ager, B., Maier, G., Feldmann, A. & Uhlig, S. (2012), Pitfalls in http traffic measurements and analysis, *in* 'Proceedings of the 13th international conference on Passive and Active Measurement', PAM'12, Springer-Verlag, Berlin, Heidelberg, pp. 242–251.

Svoboda, P. (2008), Measurement and Modelling of Internet Traffic over 2.5 and 3G Cellular Core Networks, PhD thesis, Faculty of Electrical Engineering and IT, Vienna Univeristy of Technology.

Svoboda, P., Ricciato, F., Pilz, R. & Hasenleithner, E. (2006), Composition of GPRS, UMTS traffic: snapshots from a live network, *in* 'IPS-MOME 2006 4th International Workshop on Internet Performance', Salzburg Research Forschungsgesellschaft, Salzburg, Österreich, pp. 40–51.

Tian, Y., Xu, K. & Ansari, N. (2005), 'TCP in wireless environments: problems and solutions', *Communications Magazine, IEEE* **43**(3), S27–S32.

Torres, R., Hajjat, M., Rao, S., Mellia, M. & Munafò, M. (2009), Inferring undesirable behavior from P2P traffic analysis, *in* 'Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems', ACM, pp. 25–36.

Trestian, I., Ranjan, S., Kuzmanovic, A. & Nucci, A. (2009), Measuring serendipity: connecting people, locations and interests in a mobile 3g network, *in* 'IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference', ACM, New York, NY, USA, pp. 267–279.

Tstat homepage (2010), http://tstat.tlc.polito.it/logos.shtml.

UMTS Forum (2011), 'Mobile traffic forecasts 2010-2020'.

Varga, T., Haverkamp, B. & Sanders, B. (n.d.), Analysis and modeling of WAP traffic in GPRS networks, *in* 'ITC Specialist Seminar', Citeseer.

Verkasalo, H. (2006), Empirical observations on the emergence of mobile multimedia services and applications in the US and Europe, *in* 'Proceedings of the 5th international conference on Mobile and ubiquitous multimedia', ACM, p. 3.

Verkasalo, H. & Hämmäinen, H. (2007), 'A handset-based platform for measuring mobile service usage', *info* **9**(1), 80–96.

Williamson, C., Halepovic, E., Sun, H. & Wu, Y. (2005), Characterization of CDMA2000 cellular data network traffic, *in* 'Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on', IEEE.

WURFL homepage (2010), http://wurfl.sourceforge.net/.

Zhu, Y. & Zheng, Y. (2008), Research on Intrusion Detection System based on pattern recognition, *in* 'Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on', Vol. 1, IEEE, pp. 609–612.

A WURFL Feature List

Table 13: *List of device features extracted from WURFL*

Feature Name	Description
brand name	Brand (e.g. Nokia)
marketing name	In addition to Brand and Model, some devices have a marketing name (e.g. HTC Hero)
model name	Model (e.g. N95)
device os	Information about OS of the device
device os version	Information about the version of the device os
mobile browser	Information about the device browser (e.g. Opera, Mobile Safari)
mobile browser version	Which version of the browser
model extra info	In addition to Brand and Model (and possibly a marketing name), some may be characterized by extra info (es: Nokia N95 8G, iPhone 16 GB)
pointing method	GUI navigated with either a stylus, a finger, a joystick or a BlackBerry-style clickwheel
has qwerty keyboard	If the device has a full qwerty keyboard. This can be both a real keyboard as well as a virtual one
is tablet	If a device is a tablet computer (iPad and similar, regardless of OS)
has cellular radio	Device has cellular technology (most probably a phone, but not necessarily. May be a data-only device such as Kindle or iPod touch
max data rate	Maximum bandwidth reachable by the device. HSDPA = 1800/ 3600/7200/14400, UMTS(3G) = 384, EGPRS/EDGE = 200, GPRS = 40
wifi	If device can access Wifi.
dual orientation	If the device GUI has both potrait and landscape mode
physical screen height	Screen height in millimiters
physical screen width	Screen width in millimiters
resolution height	The screen height expressed in pixels
resolution width	The screen height expressed in pixels
full flash support	If device supports the full version of Flash
built in camera	If the device has a built-in camera
built in recorder	If the device has a built-in audio recorder
receiver	May receive MMS messages
sender	May send MMS messages
can assign phone number	If device is a mobile phone and may have a phone number associated to it
is wireless device	If a device is wireless or not. Which all devices in this part of the analysis are
sms enabled	If phone supports SMS