A Comparison of Indirect and Optimal Routing in Mobile Networks

by

Qiang Fu
B.E., Harbin Institute of Technology, P. R. China

Supervisor: **Professor Lang White**

Thesis submitted for the degree of
Master of Engineering Science

Department of Electrical and Electronic Engineering
Faculty of Engineering
Adelaide University

Adelaide, South Australia
September 2000

Contents

Abstract

When a mobile user is away from its home network, it is desired that the data transfer between the mobile user and its correspondent node could use the normal IP routing to establish routing paths between them by sophisticated routing algorithms according to the optimality principle. Unfortunately, in most of cases, the data routing under current mobile networking protocols is not so efficient as normal IP routing because packets have to be routed passing through the mobile user's home network. Here, we refer to the former as direct (optimal) routing and the later as indirect routing. Some mechanisms are suggested to enable the direct routing. However, they are not perfect and additional elements have to be introduced to support these mechanisms. Therefore, we would like to see how the indirect routing affects traffic performance and how much direct routing improves the performance.

In this thesis, we summarise the basic concepts of mobile networking protocols and involved routing issues and make a comparison of indirect and direct routing in traffic performance. The traffic performance is examined based on TCP and UDP performance. Furthermore, the performance is examined for broadband and narrowband mobile user respectively. The former has an emphasis on the bottleneck within the public Internet while the later has a bottleneck within the local area, the subnet where the mobile node is attached. We design four topology scenarios that describe the relative difference between the indirect and direct routing in the length of routing path (number of hops). From topology 1 to 4, the relative difference is 0%, 50%, 100% and 433% respectively (see Figure 3.1.1-2). Topology 1 has an emphasis on encapsulation overhead since there is no difference between the two routings in length. The thesis describes the performance under various background traffic loads put on the studied topologies. Given a certain background traffic load, traffic performance is examined by changing the radio link capacity, which is described by GPRS timeslot operation.

The thesis concludes that for UDP data transmission the effectiveness of performance improvement of direct routing over indirect routing is sensitive to the relative difference between the two routings in the length of routing path while the effectiveness for TCP data transmission more depends on background traffic load - the higher the background traffic load is, more apparent the effectiveness is. Furthermore, it is shown that the location of bottleneck for data transmission has a significant impact on performance improvement. The bottleneck within wireless local area usually consumes the performance improvement created by the direct routing.

Statement of Originality

I hereby declare that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Signature:

Date: 28th August, 2000

Acknowledgments

I wish to express my deepest and sincere appreciation to my supervisor, Professor Lang White, for his constant guidance, encourage and assistance throughout my Masters study. The discussions and arguments with him made it possible to refine and finally implement the ideas in the thesis smoothly.

I would like to thank Dr. Ken Sarkies for his heuristic suggestions, which helpfully improved my original modelling of the simulations in the thesis. My appreciation also goes to Professor Reg Coutts, Dr. Sergey Nesterov and Dr. Shao-Kai Yu (now at Lucent) at the Center for Telecommunication Information Networking for their helpful suggestions and comments.

Many thanks go to the staff and students at the Department of Electrical and Electronic Engineering. In particular, I would like to thank Professor Robert Bogner, Dr. Derek Abbott and Dr. Cheng-Chew Lim for their kindly help. To Wei Xu, Alex Lin, Hong-Gang Chew, Zhi-Shun She (now at Uni. of Wales), Jee-Gah Lim, Eric Chong, Ming-Yan Li (now at Washington State Uni.), Jun Fang (now at Motorola) and all in the Communications and Signal Processing Lab., I thank them for their help and all the fun with them.

Finally, special thanks to my parents and sister for their unfailing love and support.

List of Figures

List of Tables

List of Abbreviations

ACK	Acknowledge
AFA	Anchor Foreign Agent
APN	Access Point Name
ATM	Asynchronous Transfer Mode
AUC	Authentication Center
BG	Border Gateway
BSC	Base Station Controller
BSSGP	BSS GPRS Protocol
BSS	Base Station Subsystem
BTS	Base Transceiver Station
btl	Background Traffic Load
cwnd	Congestion Window
df	Degrees of Freedom
DHCP	Dynamic Host Configuration Protocol
ETSI	European Telecommunications Standardisation Institute
FA	Foreign Agent
FIFO	First In First Out
FTP	File Transfer Protocol
Gb	Interface between a SGSN and a BSS
Gc	Interface between a GGSN and an HLR
GFA	Gateway Foreign Agent
GGSN	Gateway GPRS Support Node
GGSN-H	GGSN in HPLMN
GGSN-V	GGSN in VPLMN
GMSC	Gateway MSC
Gn	Interface between two GSNs within the same PLMN

Gp	Interface between two GSNs in different PLMNs
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
GSN	GPRS Support Node - GGSN and SGSN
GTP	GPRS Tunnelling Protocol
HA	Home Agent
HLR	Home Location Register
HPLMN	Home PLMN
IETF	Internet Engineering Task Force
IMSI	International Mobile Subscriber Identity
IP	Internet Protocol
IPv4	IP version 4
IPv6	IP version 6
ISC	International Switch Center
ISDN	Integrated services Digital Network
ISO	International Standardisation Organisation
L1	Physical Layer in ISO reference model
L2	Data link layer in ISO reference model
LA	Location Area
LAN	Local Area Network
LEC	Local Exchange Center
LLC	Logical Link Control
MAC	Medium Access Control
MM	Mobility Management
MN	Mobile Node
MO	Mobile Originated (data transfer)
MS	Mobile Station
MSC	Mobile Switching Center
MSISDN	MS ISDN number
MSRN	MS Roaming Number
MT	Mobile Terminated (data transfer)
N-VLR	New VLR

NoP	Number of Packets
NS[1]	Network Simulator
NS[2]	Network Service
NSAPI	Network layer Service Access Point Identifier
NSS	Network Subsystem
O-VLR	Old VLR
OMC	Operation Maintenance Center
OSI	Open Systems Interconnection
OSS	Operation Support Subsystem
P-TMSI	Packet TMSI
PDN	Packet Data Network
PDP	Packet Data Protocol
PDU	Protocol Data Unit
PIN	Personal Identification Number
PLMN	Public Land Mobile Network
PPP	Point-to-Point Protocol
PSTN	Public Switched Telephone Network
PTM	Point to Multipoint
PTM-G	PTM Group call
PTM-M	PTM Multicast
PTP	Point To Point
PUK	Personal Unblocking Key
RA	Routing Area
RLC	Radio Link Control
Rlt Dif	Relative Difference
RO	Route Optimisation
Rtr Rate	Retransmission Rate
RTT	Round-Trip Time
RTSP	Real-Time Streaming Protocol
SGSN	Serving GSN
SGSN-H	SGSN in HPLMN
SGSN-V	SGSN in VPLMN

SIM	Subscriber Identity Module
SNDCP	SubNetwork Dependent Convergence Protocol
ssthresh	Slow Start Threshold
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TID	Tunnel Identifier
TLLI	Temporary Logical Link Identity
TMSI	Temporary Mobile Subscriber Identity
Topo (Top)	Topology
UDP	User Datagram Protocol
Um	Interface between the MS and the GPRS fixed network part
UMTS	Universal Mobile Telecommunications System
VLR	Visitor Location Register
VPLMN	Visited PLMN
W-CDMA	Wideband Code Division Multiple Access

Chapter 1

Introduction

The General Packet Radio Service (GPRS) [1] and Mobile IP [2] enable mobile computing and networking. The GPRS Tunneling Protocol (GTP) [3] introduces global scale IP mobility tunnels while Mobile IP's IP-in-IP encapsulation [21] does the same job. When a mobile user is roaming outside its home network, data routing for the mobile user introduces a tunnel into the routing path followed by the packets sent by the correspondent node to the mobile node. The routing path consists of two parts - the path from the correspondent node to the mobile node's home network and the tunnel from the home network to the visited network of the mobile node. In Mobile IP, the tunnel originates at the mobile node's home agent and terminates at its foreign agent while GGSN in the mobile node's home network (GGSN-H) and SGSN in the mobile node's visited network (SGSN-V) serve as the two endpoints of the tunnel in GPRS. In this case the routing path from the correspondent node to the mobile node has to pass through the mobile node's home network rather than reach the mobile node directly. We refer to this kind of routing as indirect routing, also called "dog-leg" routing, in contrast to the direct routing (optimal routing) between the correspondent node and the mobile node. In mobile IP, only mobile node terminated (MT) packet routing follows the indirect routing while mobile node originated (MO) packet routing uses the normal routing to deliver each packet to its destination directly. This asymmetry of MT and MO packet routing is named "triangular routing" in Mobile IP terminology. In GPRS, both MO and MT packet routing follow the indirect routing.

Thus, packets to or from (in the case of GPRS) the mobile node are generally routed along paths that are significantly longer than optimal. For instance, if a mobile node is visiting some subnet, even packets from a correspondent node on the same subnet must be routed through the Internet to the mobile node's home network, only then to be tunnelled back to the original subnet for final delivery. Furthermore, if a high volume of data transmission were going between the correspondent node and the mobile node, the indirect routing would be intolerable.

It is desired to introduce a mechanism enabling direct routing between the correspondent node and the mobile node without tunneling involved (in case of Mobile IP) or using GGSN in the mobile node's visited network (GGSN-V) instead of GGSN-H as an endpoint of the tunnel. Route Optimisation (RO) [23] and the mechanism of dynamic address allocation [1], which are designed for Mobile IP and GPRS respectively, are focusing or practically have effectiveness on addressing this problem. Route Optimisation enables the correspondent node to be informed of current care-of-address of the mobile node so that a communication pipe could be established between the correspondent node and the mobile node directly. However, the implementation of Route Optimization requires additional changes on any Internet node, which supports RO function. Apparently, it is not practical. On the other hand, although RO would be integrated into next generation IPv6, it is not clear when IPv6 could be extensively deployed in place of IPv4. As to RO itself, it has to face several problems. Firstly, in order to keep the correspondent node informed of its current address, the mobile node has to send some messages (such as binding update message) to all involved or potential correspondent nodes after each change of agent. Since these packets (messages) may get lost or badly delayed, the correspondent nodes would continue sending packets to the obsolete agent, and thus, increased packet transfer delay and packet loss rate would be introduced. Furthermore, frequent changes of agent may induce excessive message exchanges, which may exceed the advantages of RO, between the mobile node and the correspondent nodes. Another serious problem faced by the RO is security issues. In order to avoid intruders forging the RO operation to capture the connection between the mobile node and the correspondent node, an authentication mechanism has to be introduced for the RO operation. However, this authentication generally depends on a mobility security association established in advance between the sender and receiver of such messages. It is not an easy thing to establish a prearranged authentication mechanism between the mobile node and its potential correspondent nodes. As to GPRS dynamic address allocation, it enables the mobile node to obtain a dynamic address from the visited network (VPLMN).

Therefore, the mobile node has the capability to access corresponding GGSN in the VPLMN through its SGSN. Accordingly, a direct communication pipe is established between the VPLMN and the correspondent node. However, only the mobile node can initiate such a direct communication because no Internet node knows its dynamic address before receiving packets from the mobile node. When the mobile node performs PDP context deactivation procedure (the node does not like to access the external network any more), the allocated dynamic address would be released for subsequent use by other mobile nodes. Therefore, only the mobile node's public home address can provide its availability for all the Internet nodes that know its home address. The allocation of dynamic address can address this problem only when the mobile node does not have any needs for connections initiated by the correspondent node.

We have seen that the above solutions either cost a lot or just have a limited ability to eliminate indirect routing. Thus, the question arises: is this cost worth it? We may be unable to offer a certain answer, but we would like to know how indirect routing affects traffic performance and how much direct routing improves the performance.

1.1 Research Objectives and Methodology

In this project, we study the influence of indirect routing on traffic performance, compared to direct routing between correspondent node and mobile node. Simulations are performed using Network Simulator version 2 (NS-2) from Lawrence Berkeley National Laboratory. We design four kinds of topologies and introduce different background traffic load to examine traffic performance based on indirect and direct routing. Two kinds of applications, one TCP [15] and one UDP [17] application, are mapped on indirect routing and direct routing respectively in our simulation work. From the simulation results we would like to determine how seriously the indirect routing degrades traffic performance, correspondingly how well the direct routing improves traffic performance and the effectiveness of the performance improvement. In the simulations we construct four topologies. Each topology represents a specific ratio of number of hops between the indirect and direct routing path. In topology 1, there is no relative difference between the two routings in length in order to examine the influence of tunneling overhead. The relative differences are 50%, 100% and 433% through topology 2 to topology 4 (see Figure 3.1.1-2). These four topologies basically cover the relationship between the two routings in length. The reasons we choose one TCP and one UDP application are because TCP and UDP applications are two important applications [26] and TCP and UDP performance should be dramatically different from each other. Furthermore, it permits investigation of effect of retransmissions due to packet loss.

The mobile node is connected to the visited network as broadband or narrowband mobile user. The main difference between them is the location of bottleneck on the routing path. The former is integrated into the visited network gateway and thus has a potential bottleneck on the intermediate Internet rather than on the visited network while the later has a potential bottleneck on its radio link. On the other hand, Mobile IP tunnel is introduced for the broadband mobile user while GPRS tunnel is for the narrowband mobile user. Given that Mobile IP can be used for wired mobile communication and has a focus on wireless (radio, infrared) LAN which is of much higher bandwidth than GPRS radio link, we introduce Mobile IP and GPRS functionality for broadband and narrowband mobile user respectively. The tunneling (encapsulation) overhead is 48 bytes (a GTP header plus an UDP header) for GPRS and 20 bytes (an IP header) for Mobile IP. The GPRS radio link capacity adopts coding scheme 2, that is, 13.3kb/s, and wireless loss is assumed to be perfectly recovered by local recovery mechanism. That is, no timeout, which would invoke TCP retransmission, is due to the local recovery procedure [36]. Therefore, all packet drops are due to traffic congestion.

1.2 Outline of Thesis

The thesis is organized as follows. Chapter 2 gives an overview of mobile networking protocols - GSM, GPRS and Mobile IP. Routing procedures are highlighted and we have an emphasis on indirect routing issues. Chapter 3 describes the simulation environment and the statistical methods used to analyze simulation results. Topology structure, link features, and background, TCP and UDP traffic features are presented. NS-2 simulation tool is also simply introduced in this chapter. The description of statistical methods focuses on the involved T-test and F-test. Chapter 4 presents the simulation results and detailed analysis of them. TCP and UDP performance are examined. A summary of simulation results and conclusions is presented in Chapter 5, and future possible research is introduced.

1.3 Terminology

It is not intended to be a detailed summary of terminology used in this thesis. It is intended to emphasize the terms, which

1) are important,
2) hold the same meaning but may be with different terms, or
3) are not exactly defined through the subsequent chapters.

General

Indirect Routing

The routing on the path where packets, sent by the correspondent node to the mobile node, are routed from the correspondent node to the mobile node's home network and then tunneled from the home network to the mobile node's visited network (see Figure 3.1.1-2).

Optimal Routing (called henceforth as Direct Routing)

The routing on the path where packets, sent by the correspondent node to the mobile node, are routed directly to the mobile node as standard IP routing without tunnel introduced or under the situation that tunnelling just happens within local service area (see Figure 3.1.1-2).

Tunnel

The path followed by a packet while it is encapsulated (tunneled) with one or more additional headers of which the encapsulated IP header temporarily supplants the original IP header to function during the tunnel.

Broadband Mobile User

The mobile user with broadband access to attached network.

Narrowband Mobile User

The mobile user with narrowband wireless access to attached network.

Length of Routing Path

The number of hops, used in the routing path, rather than its physical length.

GPRS

Logical Link Control (LLC)

A protocol that is responsible for maintaining a communication channel between an individual mobile station and the GPRS fixed network across the radio interface Um.

Packet Data Protocol (PDP)

Any protocol which transmits data as discrete units known as packets, e.g., IP or X.25.

PDP address

The network layer address of a GPRS subscriber. A GPRS subscriber could have one or more PDP addresses. A PDP address may be dynamic or static. A dynamic address is allocated by the operator (HPLMN or VPLMN) during PDP context activation. A static address is assigned permanently the HPLMN.

PDP context

Each PDP address is described by an individual Packet Data Protocol context in the MS, SGSN, and GGSN. Every PDP context exists independently in either active or inactive state. The PDP context must be active for data transmission using that PDP address. The PDP context contains routing information and a QoS profile, etc.

Mobile IP

Agent Advertisement

An advertisement message constructed by attaching a special Extension to a router advertisement [34] message.

Care-of-Address

The termination point of a tunnel toward a mobile node, for packets forwarded to the mobile node while it is away from home. The protocol can use two different types of care-of-address: a "foreign agent care-of-address" is an address of a foreign agent with which the mobile node is registered, and a "co-located care-of-address" is an externally obtained local address which the mobile node has associated with one of its own network interfaces.

Mobility Agent

Either a home agent or a foreign agent.

Mobility binding (binding)

The association of a home address with a care-of-address, along with the remaining lifetime of that association.

Chapter 2

Mobile Networking Protocols

2.1 Existing GSM Cellular Systems

Global System for Mobile Communications (GSM) is a wireless digital network standard that was developed by standardization committees from the major European telecommunications operators and manufacturers. For more details, please refer to the book by Michel Mouly and Marie-Bernadette Pautet, *The GSM System for Mobile Communications* [12].

2.1.1 System Architecture

The GSM technical specifications define the different entities that form the GSM network. The specifications not only specify the air interface and the message flow between mobile stations and the cellular network on that air interface, but also the whole infrastructure and all the other parts of the system described here. Basically, the GSM network can be divided into four chief parts:

- Mobile Station (MS).
- Base Station Subsystem (BSS).
- Network and Switching Subsystem (NSS).
- Operation Support Subsystem (OSS).

Mobile Station

A Mobile Station (MS) consists of two elements: Mobile Station Terminal Equipment (TE) and Subscriber Identity Module (SIM).

The types of terminals can be distinguished by power and application. The 'fixed' terminals, which are installed in cars, may have a maximum allowed output power of up to 20W. The GSM portable terminals (bag phones), which can also be installed in vehicles, can emit up to 8W and hand-portable terminals up to 2W.

The Subscriber Identity Module (SIM) is a smart card and provides the Mobile Station TE with an identity. By inserting the SIM card into the terminal, the subscriber can have access to the subscribed GSM services. Without the SIM card the terminal is not operational except for emergency calls. The SIM card identifies the subscriber to the network according to the International Mobile Subscriber Identity (IMSI) [7] stored on the card. The authentication key Ki is used in authentication procedure. Since the SIM card is the only element to personalize a terminal, it is possible for a roaming subscriber to travel only with the SIM card and rent a terminal at the destination. Because of the above, it is important to protect the SIM card from improper use. A four-digit Personal Identification Number (PIN) is stored on the card. The wrong PIN input will cause the card to block itself. An eight-digit Personal Unblocking Key (PUK), which is also stored on the card, is required to unblock the card.

Base Station Subsystem

The Base Station Subsystem (BSS) connects the Mobile Station and the Network and Switching Subsystem. It consists of many Base Station Controllers (BSC) which connect a single Mobile-services Switching Center (MSC). The main tasks of the BSC are frequency administration, handovers, exchange information and the control of Base Transceiver Stations (BTS).

Each BSC typically controls a group of BTSs, which are the Mobile Station's air interface to the network. The transmitting power of the BTS defines the maximum size of a cell. Each BTS has between one and sixteen transceiver according to the density of users in the cell, each of which represents a separate RF channel.

Network and Switching Subsystem

The Network and Switching System (NSS) executes the main switching functions of GSM and maintains the databases needed for subscriber data and mobility management. It is the responsibility of the NSS to handle the switching of GSM calls between external networks and the BSCs, and manage and provide external access to these databases. In the NSS there are three different databases: Home Location Register (HLR), Visitor Location Register (VLR), and Authentication Center (AUC). A description of different components of the NSS follows.

Mobile-services Switching Center (MSC) and Gateway MSC (GMSC)

GMSC is a node connecting the cellular network to the Public Switched Telephone Network (PSTN). With all its registers it is capable of routing calls from fixed network to a Mobile Station via the BSS. Depending on the network size, an operator might use one or several GMSCs to connect the fixed network. If the traffic capacity required is more than the capacity of the GMSCs, additional MSCs might be introduced, which do not directly the access to the fixed network. The call targeted to a GSM user is first routed to the GMSC. For the caller, the knowledge of whereabouts of the GSM user is not required. The GMSC is in charge of fetching the location information of the user and routing the call to the user through the MSC that is serving the user. There is a major difference between the GMSC and the MSC that the MSC has no related Home Location Register (HLR) but has a Visitor Location Register (VLR) linked, which might be shared by several adjoining MSCs.

Home Location Register

The HLR stores subscriber information relevant to the subscriber's identity and the provision of GSM services, and also some information of the current location of the subscriber. The former data are permanent such as the International Mobile Subscriber Number (IMSI), the phone number and permitted supplementary services. The latter are the temporary data such as the Mobile Station Roaming Number (MSRN).

Visitor Location Register

The VLR contains the necessary information that enable the subscribed GSM services to be provided to the visiting user. The VLR gets the information from the visiting user's HLR. With the information, the VLR does not need to interrogate the HLR each time when a communication is established. At the same time the VLR sends the necessary information to the visiting user's HLR for the routing of the calls to the user. The VLR holds the information of the visiting user's location at a more precise level than the HLR.

Authentication Center and Equipment Identity Register

The Authentication Center (AUC) is related to the HLR and used for security purposes. It handles the authentication and encryption keys for each subscriber. The AUC determines which algorithm should be used for a specific subscriber. The implementation of the Equipment Identity Register (EIR) is a relatively new security feature of the GSM system. The EIR identifies the terminal and can forbid calls from a stolen or unauthorised terminal.

Operation and Support Subsystem

The OSS supports one or several Operation Maintenance Centers (OMC) which are connected to the BSC and (G)MSC. The OSS monitors and maintains the performance of each MS, BTS, BSC and MSC within a GSM system. Because of the increasing number of Base Stations some of the tasks have been transferred to the BTS. This transfer decreases the costs of system maintenance.

2.1.2 Roaming and Call Routing

Location Updating

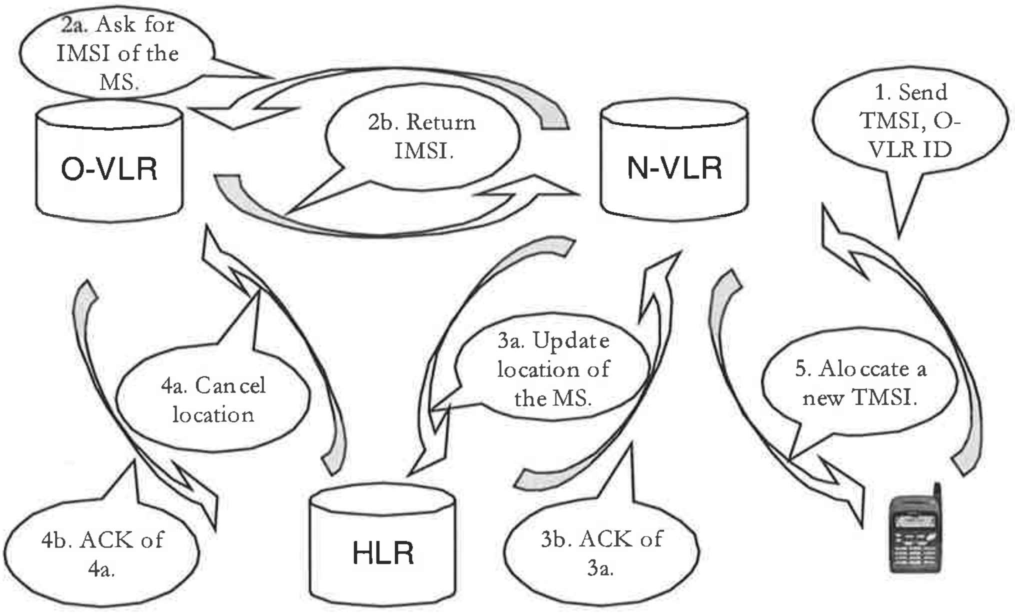


Figure 2.1.2-1: GSM location update procedure

In GSM systems, the HLR and VLR are in charge of maintaining the location information of a Mobile Station. To access a Mobile Station, location update procedure is performed to tell the system where to search for the MS during paging for an incoming call. Here a simple description of the update procedure follows.

When a Mobile Station move to a new location area. If the new location area and the previous location area are served by the same VLR, the MS is offered a new Temporary Mobile Subscriber Identity (TMSI) and its location is updated in the VLR. On the other hand, if the MS enters the scope of a new VLR, the location update procedure is a little complicated (Figure 2.1.2-1). The MS sends its current TMSI and the ID of the old location area, a "location update request" message, which is sent to BSS and then relayed to MSC. The MSC in turn alerts the VLR of the location update request. The TMSI are very useful to avoid transmitting the IMSI of an MS over radio path, thus, enhance the security of the IMSI. The new VLR (N-VLR) can identify the location of the old VLR (O-VLR) from the TMSI and the ID of the old location area. From the old VLR, the new VLR can obtain the IMSI of the MS. The new VLR identifies the address of the HLR of the MS from the IMSI and requires an authentication of the MS. If the authentication procedure is successful, the new VLR sends an update location message to the HLR. The HLR acknowledges the update location message if the update location request is accepted, meanwhile, the HLR sends a cancel location message to the old VLR. The old VLR, after getting the cancel location message from the HLR, deletes the record related to the MS and replies an acknowledgment to the HLR. After receiving the update location acknowledgment from the HLR, the N-VLR assigns a new TMSI to the MS via the MSC. After the location updating process the MS can access to the services of the new VLR.

Routing to a MS in its Home PLMN

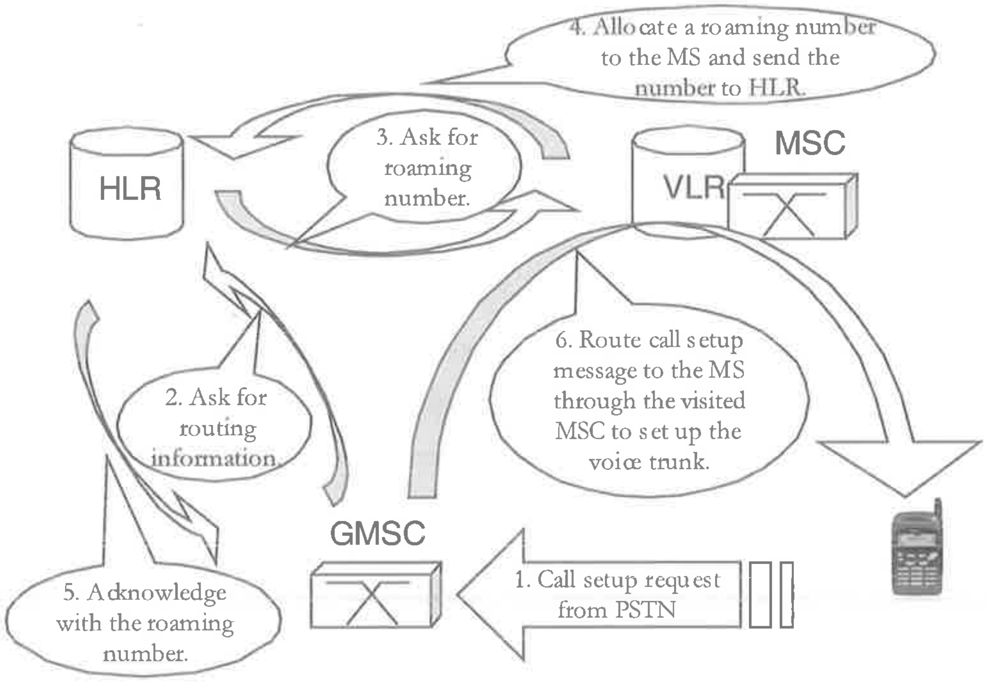


Figure 2.1.2-2 : GSM call setup procedure

A GSM subscriber A wants to set up a call to another GSM subscriber B within the same PLMN. The B's ISDN number (MSISDN) [7] is dialed and a call setup request is send to the Gateway MSC (GMSC). The GMSC asks the HLR for routing information. After receiving this request, the HLR interrogates the current VLR of B to allocate a roaming number (MSRN). The MSRN is allocated and sent to the GMSC via the HLR. The MSRN indicates the location of the visited MSC that is serving B. Using the MSRN, the GMSC sends the call set up request to B via the visited MSC. When the call is set up, voice traffic passes the circuit from the subscriber A, through the GMSC and the visited MSC, to the subscriber B. The routing of a call from the fixed network follows the same principle except that the GMSC is accessed from a PSTN or ISDN (Figure 2.1.2-2).

Routing for Roaming Subscriber

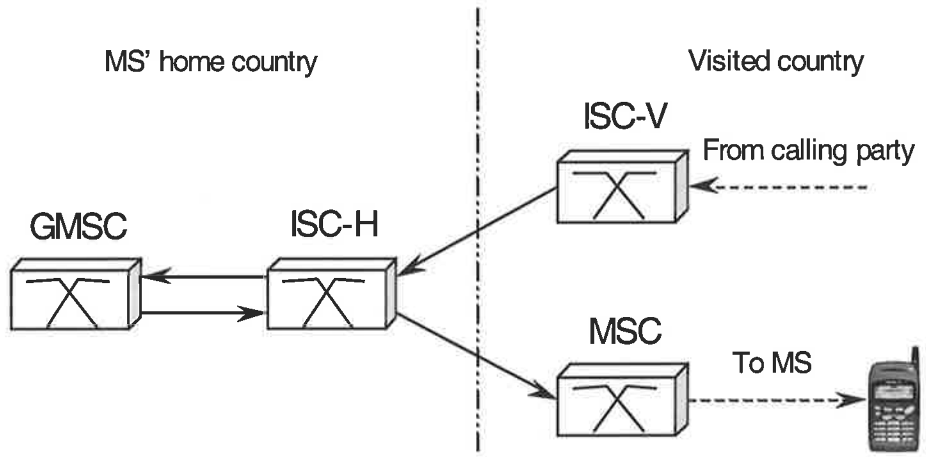


Figure 2.1.2-3: An example of international GSM call routing

The call routing for roaming subscriber is basically the same as the previous one with a difference that the inefficiency of call routing would be serious in some extreme cases. Every country has a national telecommunication network, and all the networks are connected to an international network. The International Switch Center (ISC) provides the internetworking function between the international network and the national network. When an MS is away from its home country and roaming in the country of the calling party, two ISCs are involved to set up the voice trunk. The two ISCs are between the MS's home GMSC and visited (G)MSC. The voice trunk is therefore set up as follows (Figure 2.1.2-3): from the calling party to the ISC in the visited country of the MS (ISC-V) to the ISC in the home country of the MS (ISC-H) to the home GMSC of the MS and then come back from the GMSC to the ISC-H to ISC-V to the visited (G)MSC of the MS finally to reach the MS. In this case, two international calls are involved instead of one local call.

2.1.3 Indirect Routing in GSM

When an MS is away from its HPLMN, the call setup message has to set up a voice trunk to the MS's GMSC, then a voice trunk is established between the GMSC and the visited MSC in the VPLMN. Apparently, the routing is inefficient in some cases. The instance described in "Routing for Roaming Subscriber" has illustrated this routing inefficiency. In order to fix this problem, the concept of IPLMN [5] is introduced. IPLMN is a PLMN that has the capability to interrogate the HPLMN of the called MS for its location information. Therefore, a direct routing path could be established between the IPLMN and the VPLMN, which is serving the MS. However, the solution is not perfect. A fixed calling party does not have this privilege, since the Local Exchange Center (LEC) has no capability to interrogate the HPLMN of the called MS for routing information. [25] provides several solutions that are applied to the fixed calling party case. It has to introduce a roaming subscriber location cache that indicates the current location of the subscriber.

2.2 General Packet Radio Service

General Packet Radio Service (GPRS) [1] is a standard developed by European Telecommunications Standards Institute (ETSI). GPRS employs packet-switched mode and allows an efficient usage of the radio resource. Multiple radio timeslots can be allocated to a single MS and a same timeslot can be shared between several MSs. Theoretically, GPRS can supply a data rate of up to 170kb/s, which can be flexibly allocated according to actual user demands. By adding GPRS functionality to the PLMN, GSM license holders can offer their subscribers efficient access to external Packet Data Networks.

2.2.1 System Architecture

GPRS is implemented on the structure of existing GSM PLMN and introduces two new nodes to the PLMN to support packet traffic: the gateway GPRS support node (GGSN) and the serving GPRS support node (SGSN).

GPRS Support Nodes

The GGSN, which is the first point of external Packet Data Network (PDN) interconnection with GSM PLMN supporting GPRS, is connected to the SGSN via an intra-PLMN backbone. The GGSN stores routing information such as active PDP contexts. It is the responsibility of GGSN to tunnel mobile terminated Packet Data Units (PDUs) to the MS's current point of attachment, the SGSN, according to GPRS Tunneling Protocol (GTP)[3]. The GGSN can also query the HLR for the location information of a MS via the optional Gc interface.

The SGSN is the node that is serving the MS. The SGSN, in turn, is connected to the BSS and belongs to the same hierarchical level in the network as the MSC/VLR. The basic functions of the SGSN are mobility management, paging, ciphering, data compression, traffic measure and charging. At PDP Context Activation, the SGSN establishes a PDP context for the purposes of routing. It keeps the track of the location of the MS and transmits data packets to or from the MS. To a large extent, SGSN does for packet data service what the MSC/VLR does for traditional GSM circuit-switched service.

GPRS Backbone Networks

There are two kinds of GPRS backbone networks: the intra-PLMN backbone and the inter-PLMN backbone. The intra-PLMN backbone network is an IP network connecting GPRS Support Nodes (GSNs) within the same PLMN. The inter-PLMN backbone network is an IP network connecting GSNs and intra-PLMN backbone networks via Border Gateways (BGs) in different PLMNs. Every intra-PLMN backbone network is a private IP network dedicated to GPRS data and signaling only. In order to achieve required security, some access control mechanism is applied to the intra-PLMN backbone network. The inter-PLMN backbone network can be a PDN such as public Internet or a leased line and is selected by a roaming agreement.

Home Location Register

The HLR contains GPRS subscription data and routing information. It connects to the SGSN and GGSN via Gr and Gc interfaces respectively. The Gr interface is used to exchange the data related to the location of the MS and to the management of the subscriber. The Gc interface is used by GGSN to retrieve information about the location and supported services for the subscriber.

GPRS Mobile Stations

In the GPRS standard, three modes of operation are defined [11]. The mode of operation depends on the services that the MS is attached to.

- Class A mode of operation: the MS supports simultaneous circuit switched GSM and packet switched GPRS services.
- Class B mode of operation: the MS supports both circuit switched GSM and packet switched GPRS services and can be attached to these two kinds of services at the same time, but does not support these two kinds of services simultaneously.
- Class C mode of operation: the MS is attached either to circuit switched GSM or packet switched GPRS services.

2.2.2 Protocol Architecture

The GPRS transmission plane consists of the well-known layered protocol structure according to the International Organization for Standardization/Open Systems

Interconnection (ISO/OSI) reference model. The proposed transmission plane (Figure 2.2.2-1) is up to network layer. Above the network layer, many standardized protocols may be used. The specification of these protocols is beyond of the scope of GPRS specification.

Between GGSN and SGSN is Gn interface (if the two GSNs are within the same PLMN) or Gp interface (if the two GSNs are in different PLMNs). GPRS Tunneling Protocol (GTP) [3] is employed to tunnel PDUs through GPRS backbone network. Below the GTP the transmission Control Protocol/User Datagram Protocol (TCP/UDP) [15] [17] and the Internet Protocol (IP) are used as the GPRS backbone network layer protocols. Here the GTP is not on top of IP. GPRS needs to have own fragmentation and flow control mechanism such as that provided by UDP and TCP layer. Furthermore, the IP protocol has fewer protocol identities and it is easier to use UDP/TCP ports. The GPRS backbone network may initially be based on the IPv4 protocol [18], and ultimately, IPv6 [19] may be used. Below the IP protocol, many protocols such as Ethernet, ISDN or ATM-based protocols can be used depending on the operator's network infrastructure.

Between SGSN and BSS, Base Station System GPRS Protocol (BSSGP) is used to convey the information related to routing and QoS. Network Service (NS) layer is based on the Frame Relay connection between BSS and SGSN and is used to transport BSSGP PDUs.

Gb interface connects SGSN and BSS while Um interface is between BSS and MS. Subnetwork Dependent Convergence Protocol (SNDCP) [10] maps network-level protocol characteristics onto the underlying logical link control [9] and provides multiplexing of network layer messages onto a single virtual logical connection, encryption, segmentation and compression functionality. Between BSS and MS is Um interface [6] where the radio link control/medium access control (RLC/MAC) [8] sublayers enable a multitude of MSs to share a common transmission medium. Some simulation works on the interface are presented in [13] and [14].

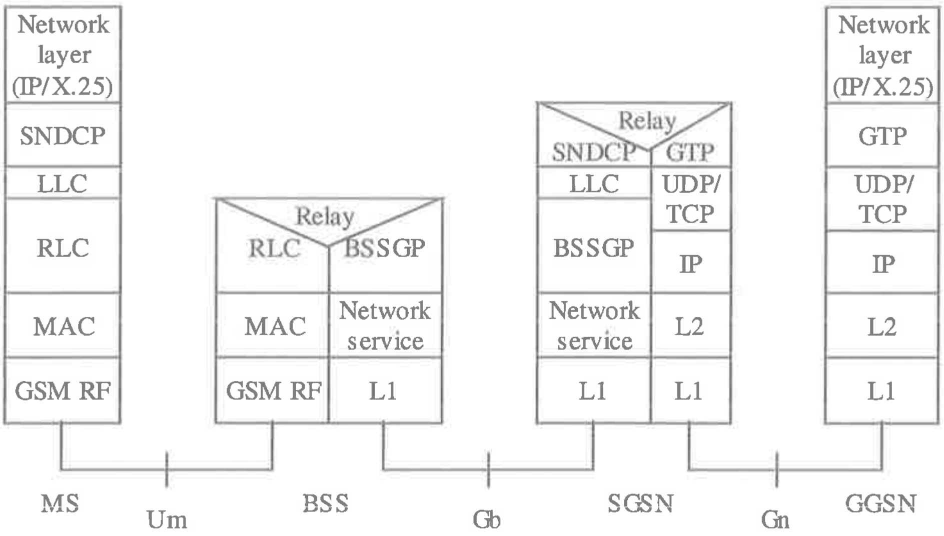


Figure 2.2.2-1: GPRS transmission plane

2.2.3 Mobility Management

Before an MS is able to send data to the GPRS network, it has to attach to the GPRS system via the GPRS attach procedure. In order to reach the MS, the GPRS network has to know the current location of the MS. A GPRS detach procedure is executed when the MS wants to leave the system. In the GPRS specification, three different Mobility Management (MM) states are defined - Idle state, Standby state and ready state.

2.2.3.1 Mobility Management States

Idle State

In Idle State, the MS is not attached to the GPRS mobility management. The MS is unreachable from the GPRS network. The location and routing information stored in the MS and SGSN is outdated and non-valid. In this state, the mobility management procedures are not performed but PLMN selection and GPRS cell selection procedures can be performed. The MS may receive Point To Multipoint Multicast (PTM-M) transmissions but the Point To Point (PTP) and Point To Multipoint Group Call (PTM-G) transmissions from or to the MS are unavailable. In order to establish MM contexts and make the MS reachable, a GPRS attach procedure has to be executed.

Standby State

In Standby State, the MS is attached to the GPRS mobility management and can be reached by the GPRS network. The MS and SGSN have established MM and PDP contexts. In this state, the MS may receive PTM-M and PTM-G transmission and the pages for incoming PTP and PTM-G data or signaling information may be received. However, the PDP transmissions and receptions, and PTM-G transmission are not possible.

In the state, the paging occurs within the whole routing area where the MS is staying. The MS only executes Routing Area (RA) update rather than cell update although the GPRS cell selection and reselection processes can be performed. When the MS enters a new routing area the MS executes the routing area update and informs the SGSN of the identity of the new routing area. Therefore, the location information in the SGSN MM context contains only the GPRS Routing Area Identity (RAI) for the MS.

When the MS successfully responds to a paging the MM state in the MS is changed to the Ready state, and when the response is received by the SGSN the MM state in the SGSN is also changed to the ready state. Similarly, the MM state in the MS and SGSN is respectively changed to the Ready state when data or signaling information is sent from the MS and received by the SGSN. The Standby State can be changed to Idle State if the MS or the network initiates a GPRS Detach Procedure. After expiry of the mobile reachable timer the SGSN may execute an implicit detach in order to return the MM state in the SGSN to Idle State. The SGSN does not maintain the MM and PDP contexts in Idle State.

Ready State

In Ready State, the MS performs mobility management procedure at the level of cell update. An identifier of the cell, the Cell Global Identity including Routing Area Code (RAC) and Location Area Code (LAC), is included in the BSSGP header of the data packet from the MS and sent to the SGSN. The SGSN MM context stores the location information for the MS on cell level. In the state, a cell update occurs when the MS enters a new cell within the current routing area. If the new cell is inside a new routing area, a routing area update takes place instead of a cell update.

The MS may receive PTM-M and PTM-G data, and also send and receive PTP data in Ready State. The SGSN transfers data to the MS via the BSS that is currently serving the MS. The MS may activate or deactivate PDP contexts while in Ready State. An MM context moves from Ready State to Standby State when the Ready Timer expires. Performing a GPRS Detach Procedure can move the MM context from Ready State to Idle State.

2.2.3.2 Attach and Detach Procedure

Attach Procedure

Before an MS accesses the services of a GPRS network, a GPRS Attach Procedure has to be performed. The GPRS attach is made to the SGSN. During the attach procedure, the MS should provide its identity and indicate the type of attach - a combined GPRS/IMSI attach or a GPRS attach only. Here, we just take pure GPRS attach into consideration. The identity provided to the network should be the valid Packet TMSI (P-TMSI) or IMSI of the MS. The P-TMSI should be sent to network with the routing area identity to identify the latest SGSN that served the MS before the latest GPRS detach procedure is executed. If the MS has not a valid P-TMSI, then the MS has to provide its IMSI. The P-TMSI is similar to the TMSI in circuit-switched GSM services. It can avoid the MS's real identity (IMSI) being sent over the radio path. Another important function of P-TMSI is to produce Temporary Logical Link Identity (TLLI) which is used to identify the MS at and below the LLC [9] level.

After having performed GPRS attach procedure, the MS is in Ready State and MM contexts are established in the MS and the SGSN. Sequentially, the MS can activate PDP contexts.

Detach Procedure

The GPRS Detach Procedure can be initiated by the MS to inform the network that it wants to detach from the GPRS system, and also can be initiated by the network (SGSN and HLR) to inform the MS that it has been GPRS-detached by the network. The GPRS detach is grouped into two classes: Explicit Detach and Implicit Detach. Explicit Detach is the detach that the network or the MS explicitly requests. Implicit Detach occurs when the network detaches the MS, without notifying the MS, such as when a configuration-dependent time after the mobile reachable timer expired, or after an irrecoverable radio error causes disconnection of the logical link.

When the MS requests a detach, the MS is required to indicate if the detach is due to switch off or not. If it is due to switch off the Detach Accept message is not required. If not, the Detach Accept message should be returned. In the network-initiated Detach Request message, there may be an indication to tell the MS that it is requested to initiate GPRS Attach and PDP Context Activation procedures for the previously activated PDP contexts. If there is, the attach procedure should be executed when the detach procedure is completed.

2.2.3.3 Location Management

The GPRS Location Update is grouped into Cell Update and Routing Area Update. The MS detects if it enters a new cell by comparing the cell's identity with the cell identity stored in the MS's MM context. If cell update is required the cell with the strongest

Packet Broadcast Control Channel (PBCCH) signal is chosen. The MS detects if a new routing area has been entered by periodically comparing the RA identity stored in the MS's MM context with the RA identity which is received from the new cell. Similarly, the RA with the strongest cell signal is chosen when a RA update is required.

For a GPRS user, it is very common to change the cell after the MS was attached to the GPRS. At the same time, sending uplink data consumes much battery power. It is important to determine when to perform cell update and RA update. If the MS updates its location based on every cell change, the network knows the MS's location in the accuracy of cell. Thus, paging is not needed and data delivery is fast without waiting paging response. However, battery power and uplink radio resources are seriously consumed because of frequent cell changing. On the other hand, if the MS updates its location based on a big area change such as Location Update in GSM, it decreases the number of location update but increases the scope of paging area. The MS saves a lot of battery power and uplink radio resource is also saved. However, every downlink packet requires paging of the MS because the exact location of the MS is not available, and uplink radio resource is wasted waiting for paging responses. Consequentially data transfer delay is increased dramatically.

In order to implement efficient mobility management, the concepts of Routing Area and Ready and Standby State are introduced to the GPRS. We have described the Ready, Standby and together with Idle State in the section of Mobility Management States. When the MS is in the Standby State, the MS reports its location based on a big area change. This big area is called Routing Area (RA) in GPRS. A RA includes more than one cell while an SGSN controls more than one RA. When the MS is activated, that is, the MS is in Ready State, the MS informs the network in every cell change. The MS changes from Standby State to Ready State when the MS sends a packet. Meanwhile, the Ready Timer is started up. Every packet sending resets the timer and at the SGSN the timer is reset when a packet is correctly received. The length of the timer is the same in the MS and SGSN. The initial value of the length is defined by a default value. Only the SGSN can change the length of the Ready Timer by transmitting a new value in the Attach Accept, Routing Area Update Accept, or Anonymous Access PDP

Context Accept messages. When the value is set to 0 the MS is forced to the Standby State immediately. If the value is set to 1s in all bits the timer doesn't run and the MS remains in the Ready State until another value is received from the SGSN. The MS changes from Ready State to Standby State when the Ready Timer expires.

Intra SGSN Routing Area Update

We have mentioned that an SGSN controls more than one RA. Thus, the Intra SGSN routing area update happens when the MS changes from one RA to another while both RAs are handled by the same SGSN. The following description of the update procedure is taken from [1].

1) The MS send a Routing Area Update Request to the SGSN. The message includes the old RAI, old P-TMSI Signature, and Update Type. The Update Type indicates RA update or periodic update. The BSS adds the Cell Global Identity including the RAC and LAC of the cell where the message was received before passing the message to the SGSN.
2) Security functions may be executed.
3) The SGSN validates the MS's presence in the new RA. If due to regional subscription restrictions the MS is not allowed to be attached in the RA, or if subscription checking fails, then the SGSN rejects the routing area update with an appropriate cause. If all checks are successful then the SGSN updates the MM context for the MS. A new P-TMSI may be allocated. A Routing Area Update Accept (P-TMSI, P-TMSI Signature) is returned to the MS.
4) If P-TMSI was reallocated, the MS acknowledges the new P-TMSI with Routing Area Update Complete (P-TMSI).

Inter SGSN Routing Area Update

Because no routing area is shared by more than one SGSN, an Inter SGSN Routing Area Update is performed when the MS changes from an SGSN to another. In contrast to the Intra SGSN Routing Area Update procedure, the Inter Update not only involves the current (new) SGSN, and also the old SGSN, the associated GGSN and the MS's HLR. The update procedure is described in the following steps. The description is not complete but is intended to provide the most important information and quick understanding. Detailed description is available from [1]. The update procedure is illustrated in Figure 2.2.3-1.

1) When the MS moves to a new routing area, the MS sends its P-TMSI signature and RAI to the new SGSN through the Routing Area Update Request. The P-TMSI is optionally sent by the SGSN to the MS in Attach Accept and Routing Area Update Accept messages. It is only known by the SGSN and the MS. Since Routing Area Update Request and Attach Request are not ciphered, hackers could send fake requests to disturb the real users. The introduction of P-TMSI signature allows the SGSN to authenticate the MS with the unciphered requests.
2) From the RAI the new SGSN identifies the old SGSN and sends SGSN Context Request to the old SGSN to get the MM and PDP context for the MS. The authentication of the MS is executed at the old SGSN using the P-TMSI signature. If no security problem, the old SGSN responds with SGSN Context Response including the MM context, PDP context and LLC ACK. The old SGSN stores the address of the new SGSN in order to forward packets addressed to the MS to the new SGSN. The security functions may be performed, initiated by the new SGSN between the MS and its HLR. After receiving SGSN Context Response, the new SGSN send an acknowledge message to indicate it is ready to receive packets from the old SGSN. Thus, the packets belonging to the MS are forwarded to the new SGSN from the old SGSN.

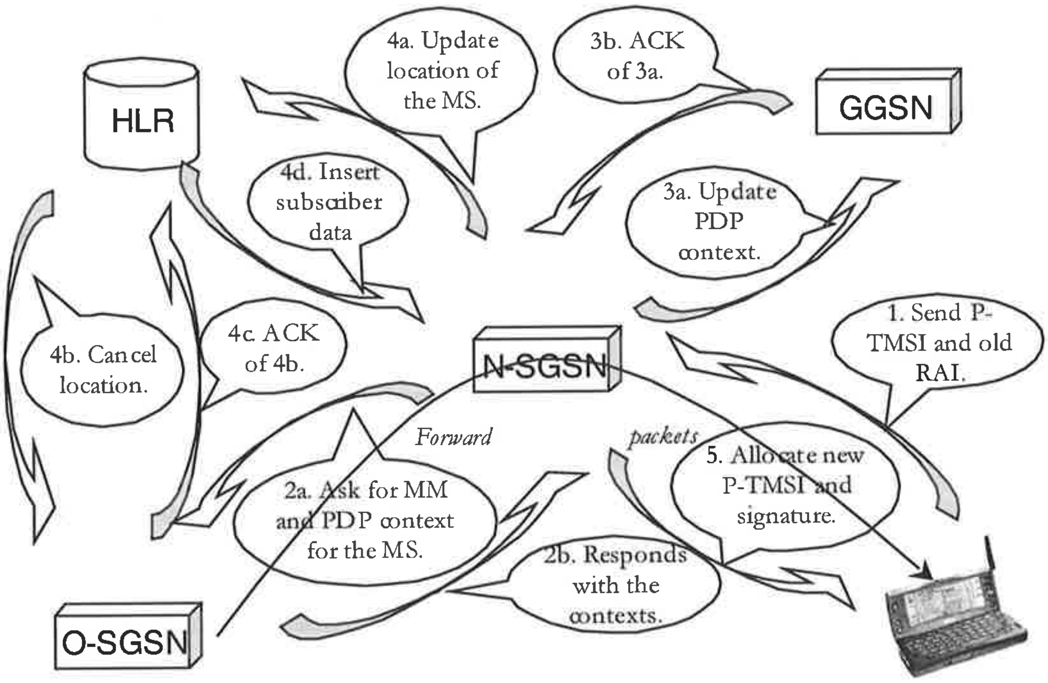


Figure 2.2.3-1: Inter SGSN Routing Area Update procedure

3) In order to point the associated GGSNs to the new SGSN, PDP contexts are updated in the GGSNs. The SGSN sends Update PDP Context Request containing new SGSN Address and TID to the GGSNs. The GGSNs then responds with TID included. With the TIDs, the tunnels between the new SGSN and the GGSNs are established.

4) Similarly, the HLR of the MS should be informed of the location change. From the MM context the new SGSN identifies the MS's HLR and sends a location update message to the HLR. The HLR then sends a location cancel message to the old SGSN, using IMSI to identify the MS. After completing the forwarding of packets addressed to the MS, the SGSN removes the concerned contexts of the MS and acknowledges with a location cancel ACK message containing the IMSI of the MS. Meanwhile, the HLR sends the MS's GPRS subscription data to the new SGSN. The new SGSN validates the MS's presence. The new SGSN sends an ACK message and indicates whether there is any limitation on the operation of the MS in the routing area.

Figure 2.2.4-1 illustrates the routing procedure in GPRS. For MS Terminated (MT) data transmission, the GGSN intercepts the incoming packet from the external packet data network (PDN) [4] and detects the destination IP address in the IP header of the packet. If the corresponding PDP context of the IP address has been activated (general cases), the GGSN encapsulates the packet with a GPRS Tunneling Protocol (GTP) header, and inserts this GTP packet in a UDP or TCP packet that again is inserted in a IP packet. Generally, for the packet from IP networks UDP is required under the GTP and for the packet from X.25 networks TCP is suggested under the GTP. After the encapsulation procedure, a GTP, UDP or TCP and IP header are added to the original packet. The IP header identifies the address of the SGSN that is serving the MS. The GTP header contains tunnel endpoint identifier (TID) that uniquely addresses the MS' PDP context in the SGSN. Thus, with the information provided by the encapsulation procedure the GTP packet can be transferred to proper SGSN. After receiving the GTP packet, the SGSN finds the corresponding PDP context according to the TID information in the GTP header. From the PDP context SGSN can fetch a Temporary Logical Link Identity (TLLI) and a Network layer Service Access Point Identifier (NSAPI). TLLI is derived from the P-TMSI. An NSAPI is assigned when the MS initiates the PDP context activation procedure. The TLLI/NSAPI pair is unique within the routing area. The TLLI unambiguously identifies the logical link between the MS and the SGSN. In the MS the NSAPI identifies the PDP-SAP (PDP-Service Access Point). At the same time, the SGSN decapsulates the GTP packet and then delivers the packet to the MS with Subnetwork Dependent Convergence Protocol (SNDCP).

Mobile Originated data transfer

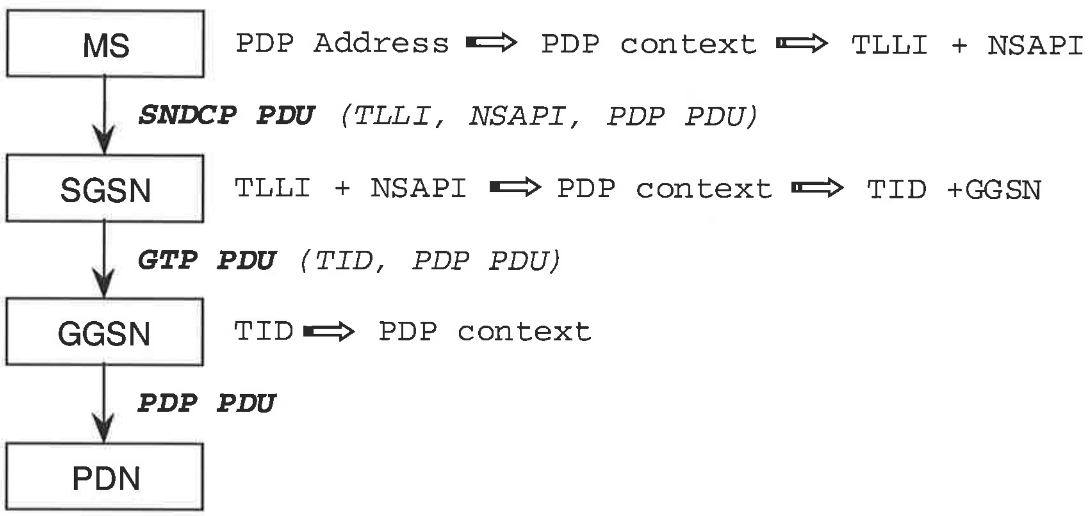


Figure 2.2.4-2: Mobile originated data routing

In the case of an MS originated transmission (Figure 2.2.4-2), the MS identifies the corresponding PDP context with the address which it uses as source address in IP header. From the PDP context the TLLI/NSAPI pair is available. The packet is sent to the SGSN as SNDCP packet that consists of TLLI, NSAPI and original packet. Using the TLLI/NSAPI pair the SGSN can identifies corresponding PDP context in its dictionary. Accordingly, corresponding TID and GGSN IP address can be identified. After an encapsulation procedure, the GTP packet is produced and sent by the SGSN to the addressed GGSN where the GTP packet is decapsulated and forwarded to the correct external packet data network.

2.2.4.3 Data routing for a roamer

In this case, the MS has moved to another PLMN. The SGSN that is currently serving the MS is located in the visited PLMN while the GGSN is in the home PLMN. In principle, the routing for a roamer is almost the same as the previous one. However, routing in this case is inefficient because the packet from or to the MS has to pass through the GGSN in the HPLMN of the MS. We refer to the routing as indirect routing.

2.2.5 Indirect Routing in GPRS

Figure 2.2.5-1 illustrates the indirect routing in GPRS. When an MS is away from its home network whether MO or MT packets have to pass through the GGSN in the MS's HPLMN. For MT packets, the GGSN sends the packets to the SGSN, which is currently serving the MS in the VPLMN, through intra-GPRS backbone and border gateway (BG) in the HPLMN, inter-GPRS backbone, and BG and intra- GPRS backbone in the VPLMN. And then, the packet data are delivered to the MS. For MO packets, the routing is in the same way with an opposite direction. In this scenario the correspondent node cannot directly communicate with the MS via a GGSN in the VPLMN. If the correspond node is a GPRS subscriber of the same VPLMN, the routing inefficiency is very serious.

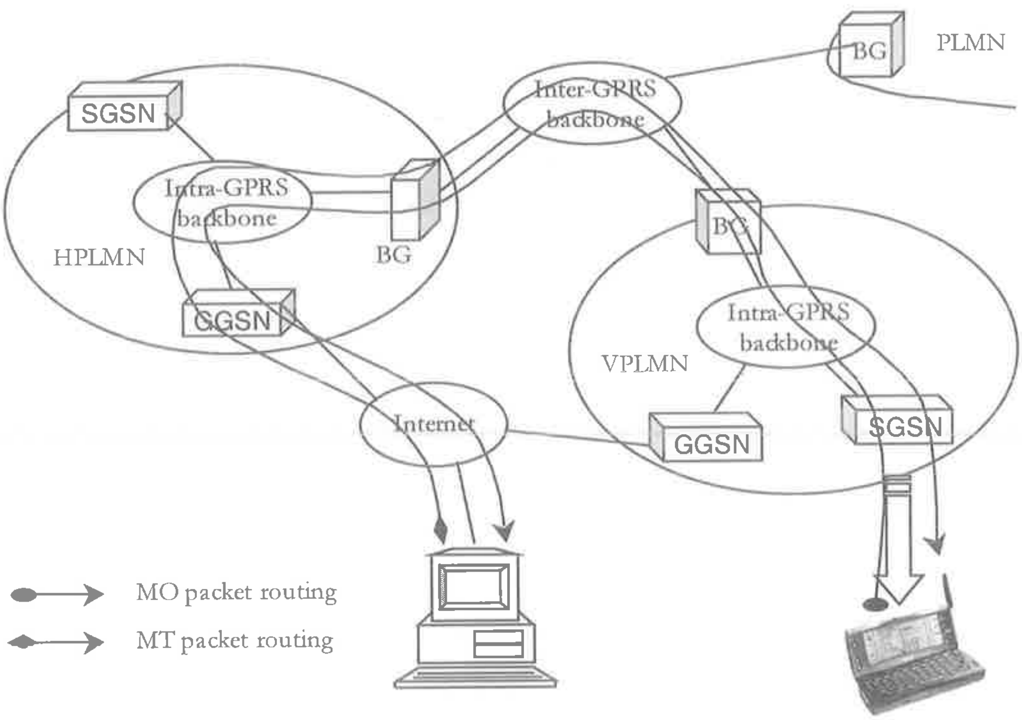


Figure 2.2.5-1: Indirect routing in GPRS

In order to eliminate indirect routing and establish direct routing, the GGSN in the VPLMN (GGSN-V) should be accessible for MO and MT data transfer. Therefore, the MS should be assigned a dynamic address by the VPLMN. This dynamic address indicates its current location. It is allocated by the GGSN-V to the MS during PDP context activation procedure. A PDP context is thus created for the dynamic PDP address at the GGSN-V. This PDP context enables the MS to use the GGSN-V for data transmission, that is, the MO and MT data packets don't need to pass through the GGSN in HPLMN (GGSN-H) to access to external data networks after the PDP context activation procedure. The dynamic address is released by the GGSN-V when a PDP context deactivation procedure is performed. The deactivation procedure may be initiated by the GGSN-V, SGSN or the MS itself. Afterwards, if the MS wants another communication under direct routing, it should request a new dynamic address from the VPLMN during the PDP context activation procedure. According to the above description, we can see that the routing for MO and MT data packets can be optimised if the activated PDP address is dynamically assigned by the VPLMN. However, the optimal routing implemented by the dynamic PDP address assumes that the MS does not have any needs for communications originated by the correspondent node, otherwise the MS has to activate its static home address and thus data transfer has to follow the indirect routing. In order to make optimal routing available for the communications originated by the correspondent node, new mechanism has to be introduced. The mechanism may be a little like Route Optimisation [23] suggested for Mobile IP.

2.3 Mobile IP

Mobile IP [2] [24] developed by Internet Engineering Task Force (IETF) is an extension of IP. It enables mobile node to move from one IP subnet to another without its computing activities being disturbed. Mobile IP allows the mobile node to use two IP addresses: a fixed home address and a care-of-address. The former is used to identify the mobile node. The later identifies the current attachment point of the mobile node to the Internet, and therefore enables packets addressed (home address) to the mobile node to be routed (tunneled) to the attachment point.

2.3.1 Mobile IP architectural entities

Mobile IP introduces three new architectural entities: mobile node, home agent and foreign agent. They are defined as follows [2].

Mobile Node (MN)

A host or router that changes its attachment point from one network to another. It may change its location without changing its IP address and continue to communicate with other Internet nodes using this IP address. It is assumed that link-layer connectivity to a point of attachment is available.

Home Agent (HA)

A router on a mobile node's home network which tunnels datagrams addressed to the mobile node for delivery to the mobile node when it is away from home. Home agent maintains current location information for the mobile node.

Foreign Agent (FA)

A router on a mobile node's visited network which provides routing services to the mobile node while registered. The foreign agent detunnels and delivers datagrams to the mobile node that were tunneled by the mobile node's home agent. For datagrams sent by the mobile node, the foreign agent may serve as the default router.

A mobile node is given a long-term IP address on a home network. This home address is administered in the same way as a "permanent" IP address is provided to a stationary host. When away from its home network, a "care-of-address" is associated with the mobile node and reflects the mobile node's current point of attachment. The mobile node uses its home address as the source address of all IP datagrams that it sends, except where otherwise described in this document for datagrams sent for certain mobility management functions.

2.3.2 Protocol overview

There are three main operations of Mobile IP, i.e. Agent Discovery, Registration and Tunneling. Along with the description of how Mobile IP operates, we present these three operations.

Agent Discovery

Using Agent Discovery Process, a Mobile Node can determine whether it is on its home network or a foreign network, and obtains a care-of-address when connected to a foreign link.

Home agents and foreign agents may advise their presence via Agent Advertisement messages. The messages are periodically transmitted as multicasts or broadcasts on each link on which they provide service. This allows a mobile node that is connected to such a link to determine whether any agents are available. Optionally, if a Mobile Node is impatient to wait around for the next periodic transmission of an Agent Advertisement, it can solicit an Agent Advertisement message from any locally connected mobility agents (home agent or foreign agent) through an Agent Solicitation message. This is useful in those situations where the frequency at which agents are transmitting is too low for a mobile node that is moving rapidly from link to link.

The introduction of Agent discovery provides a mechanism by which a mobile node can determine whether it is connected to its home link or a foreign link, and discover mobility agents that are present on its current link. The mobile node can passively listen for an Agent Advertisement message or positively by sending an Agent Solicitation message. After receiving the Agent Advertisement message, the mobile node detects its location - on its home network or a foreign network. When the mobile node is located on its home network, it operates without mobility service. If the mobile returns to its home network from being registered elsewhere, the mobile node has to deregister with its home agent. This process is performed by exchange of a Registration Request and Registration Reply message between the mobile and its home agent. These two messages are also used in the following Registration process. When the mobile node is visiting at a foreign network, it obtains a care-of-address on the foreign network. The mobile node can read the care-of-address directly from an Agent Advertisement sent by a foreign agent on the link. Or, if the mobile node connects to a link on which no advertisements are being sent, the mobile node can use transport and link-layer indications to determine its current location, and then acquire a care-of-address using DHCP [20] or manual configuration. In order to communicate with other Internet nodes, the mobile node has to validate its care-of-address. Therefore, the following Registration process is necessary.

Registration

Thus, the mobile node has to register its care-of-address with its home agent so that it can be found for communication. Depending on the method by which the mobile node acquires its care-of-address, it registers either directly with its home agent, or through a foreign agent that forwards the registration to the home agent. The registration process is implemented through exchange of a Registration Request and Registration Reply between the mobile node and its home agent directly or through a foreign agent.

If the mobile node's care-of-address is provided by its foreign agent through an Agent Advertisement message, the registration messages are relayed by the foreign agent from and to the mobile node. This kind of address is referred to as "foreign agent care-of-address". It is an IP address of the foreign agent. Foreign agent care-of-address is preferred because it allows many mobile nodes to share the same care-of-address and therefore doesn't place unnecessary demands on the already limited IPv4 address space. If the care-of-address is an IP address obtained by the mobile node as a local IP address through some external mechanism, the registration process is performed between the mobile node and its home agent directly. The care-of-address acquired by this mode is called "co-located care-of-address". Co-located care-of-address may be dynamically assigned as a temporary address to the mobile node such as through DHCP or may be owned by the mobile node as a long-term address for its use only when it is away from home. It allows a mobile node to function without a foreign agent, however place additional burden on the IPv4 address space because additional address pool is required in the foreign network for visited mobile nodes. During the registration process, encryption is used to authenticate the registration information. The registration process establishes a mobility binding consisting of the mobile node's home address, its care-of-address, and the lifetime of the registration. When the registration lifetime is near expiration, the registration should be renewed. And whenever the mobile node detects a change in its network connectivity it should initiate a registration process. When the node is away from home, the registration process allows its home agent to create or modify a mobility binding for it. When it is at home network, the (de)registration process allows its home agent to delete any previous mobility binding(s) for it. Thus, the mobile node operates normally using its home address without mobility services.

Since the registration process, the home agent has identified current location of the mobile node and can communicate with the node. In the following section, Tunneling, we present how the home agent forward datagrams to the mobile node.

Tunneling

When the mobile node's home agent intercepts datagrams sent to its home address, the home agent tunnels them to the mobile node's care-of-address where the tunneled datagrams are decapsulated and finally delivered to the mobile node. The home agent and the care-of-address are the two endpoints of the tunnel. We've mentioned that there are two kinds of care-of-address: foreign agent care-of-address and co-located care-of-address. When using the former, the foreign agent serves as the endpoint of the tunnel and decapsulates the tunneled datagrams and delivers the inner original datagram to the mobile node. When the later is used, the mobile node self serves as the endpoint of the tunnel and performs the decapsulation process.

A datagram sent to the mobile node from an Internet host (correspondent node) consists of IP header and IP payload. The IP header has a Source Address field loaded by the correspondent node's IP address, a Destination Address loaded by the mobile node's home address and a Protocol Identifier which indicates the next higher lay protocol (such as TCP, UDP, ICMP, etc.) that comprises the content of the packet payload. The datagram is routed to the mobile node's home network using the normal IP routing. After intercepted by the home agent, using IP-within-IP encapsulation [21] the home agent insert a new IP header - tunnel header in front of the IP header of the datagram. Within the new IP header, the source address is the home agent's IP address, the destination address is the mobile node's care-of-address and the protocol identifier indicates the use of IP-within-IP encapsulation. With the encapsulation process, the original datagram is tunneled to the other endpoint of the tunnel - tunnel destination point (the foreign agent or the mobile node itself, depending on which kind of care-of-address is used). At the tunnel destination point, in order to recover the original datagram the decapsulation process is performed to eliminate the tunnel header. When the foreign agent is the tunnel destination point it has to compare the inner destination address - the home address of the mobile node to those entries in its visitor lists. If the destination address matches one in the visitor list it delivers the original datagram to the node. If none match the destination address in the visitor list, the foreign agent discards the datagram rather than forwards it without modifications to the original header, because otherwise a routing loop may be introduced.

For datagrams sent by the mobile node, they are routed as general datagrams. In their IP headers, the mobile node's home address and the correspondent node's IP address are used as the Source Address and Destination Address respectively. The foreign agent may serve as a default router to forward the datagrams to the correspondent node.

We use IP-with-IP as an example of encapsulation mechanisms above. Alternatively, a protocol called Minimal Encapsulation within IP [22], which combines the information of the tunnel header with the original header to reconstitute the original header, accomplishes the same function but has less protocol overhead by eliminating some redundant header information.

2.3.3 Indirect routing in Mobile IP

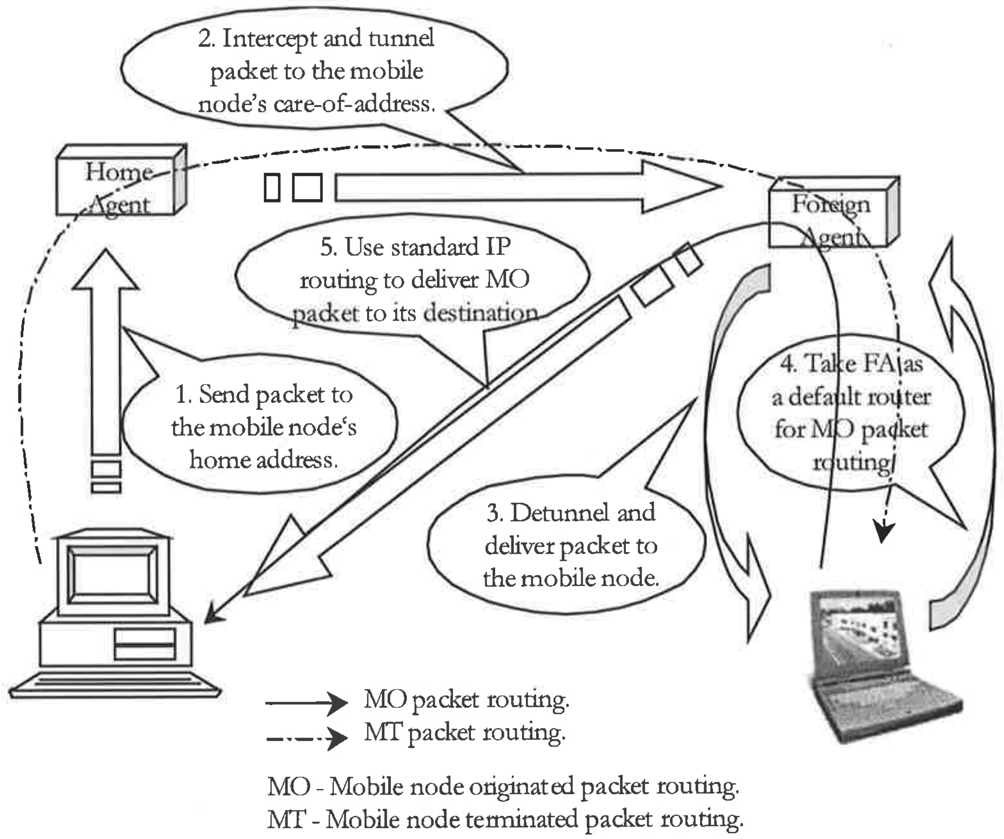


Figure 2.3.3-1: Mobile IP routing procedure

Figure 2.3.3-1 illustrates Mobile IP routing procedure. MT packets are routed to the mobile node via its home agent and foreign agent while the MO packets are carried to the correspondent node directly without passing through its home agent. This asymmetry between MO and MT packet routing is captured by the term of "triangular routing" in Mobile IP terminology. In GPRS, both MO and MT packet routing are in the same way and therefore both routings suffer from routing inefficiency when the MS is away from its home network. In Mobile IP, when the mobile node moves to a foreign network, only MT routing suffers from this inefficiency. In convenience of description, we use "indirect routing" as used in GPRS instead of "triangular routing" to describe this routing inefficiency in Mobile IP.

Route optimization [23] is introduced to eliminate the indirect routing. The main idea of route optimization is to establish a binding cache in any Internet node supporting route optimization. The binding cache contains the care-of-address of one or more mobile nodes. The binding cache is created or updated only when the Internet node has received and authenticated the mobile node's mobility binding. In the absence of any binding cache entry, packets targeted to the mobile node are routed as usual Mobile IP without route optimization support. Route optimization also support "foreign agent smooth handoff". That is a means to keep the mobile node's previous foreign agent being informed its new mobility binding so that the packets in flight to its previous foreign agent can be forwarded to its new care-of-address. However, route optimization has a fatal weakness - require route optimization support to be added to any Internet node which wants to communicate with the mobile node directly. If next generation IPv6 can be successfully deployed, the functionality of route optimization would be realistic since it is possible to integrate route optimization into IPv6 [24].

Chapter 3

Simulation Environment and Statistical Methods

This chapter describes the simulation environment to perform the experiments and the statistical methods used to analyze the data collection from the experiments.

3.1 Simulation Environment

The simulations are performed using the Network Simulator version 2 (NS-2) from Lawrence Berkeley Labs. NS-2 is a discrete event simulator targeted at networking research and provides substantial support for simulation of TCP, routing, and multicast protocols. It is implemented in C++ that uses Tool Command Language (TCL) and MIT's Object TCL (OTCL) as the command and configuration interface. Figure 3.1.1-1 is a glimpse of TCL code for the simulations.

In this section, the description focuses on the topology structure, link feature, background traffic and studied TCP and UDP traffic features. The background traffic is created at the head of each link and received at the end of the link. Because the end of a link is the head of the next link, the packets received and created at the node are put into the same queue. The received packets are destroyed after complete reception and don't contribute to the creation of packets at the node. The studied TCP and UDP traffic is produced at source (sending) node and terminated at destination (receiving) node. The source node is integrated into the correspondent gateway while the destination node is included in the visited network gateway with a 0.01ms distance for radio link between them. Under direct routing packets are transferred to the destination node without passing through the home network gateway - no tunnelling occurs. It accords with normal IP routing. Under indirect routing the packets have to pass through the home network gateway where tunnelling begins. The tunnelling ends at the visited network gateway where decapsulation occurs and finally the packets are delivered to the destination node. In order to establish a clear understanding of the difference

```
    ...

# Create Topology and Background Traffic

    ...

for {set i 0} {$i < $hc } {incr i} {
  set ii [expr $i+1]
  $ns_ duplex-link \
    $n($i) $n($ii) $BW $DY DropTail;     # Create a bi-directional
                                         # link between node i $n(i)
                                         # and node ii $n(ii).

  # Monitor queue BS-Buffer Size
  # Forward
  $ns_ duplex-link-op $n($i) $n($ii) queuePos 0.5
  $ns_ queue-limit $n($i) $n($ii) $BS;  # Define buffer size.
  # Reverse
  $ns_ duplex-link-op $n($ii) $n($i) queuePos 0.5
  $ns_ queue-limit $n($ii) $n($i) $BS

  for {set j 0} {$j < 5} {incr j} {
    set i_j $i
    set ii_j $ii
    set vsr [append i_j - $a]
    set vsi [append ii_j - $a]
    set src($vsr) [new Agent/UDP];      # Source j of background traffic.
    set sink($vsi) [new Agent/UDP];     # Sink j of background traffic.
    $ns_ attach-agent $n($i) $src($vsr)
    $ns_ attach-agent $n($ii) $sink($vsi)
    $ns_ connect $src($vsr) $sink($vsi)
    set e($vsr) \
      [new Application/Traffic/Pareto]; # Create Pareto traffic.
    $e($vsr) attach-agent $src($vsr);   # Bind Pareto traffic to
                                        # source node.
    $e($vsr) set packet_size_ $pkts
    $e($vsr) set burst_time_ $bst
    $e($vsr) set idle_time_ $idt
    $e($vsr) set rate_ $brate
    $e($vsr) set shape_ $pshape

    $ns_ at $bt_start "$e($vsr) start"; # Start background traffic flows.
    $ns_ at $bt_stop "$e($vsr) stop";   # Stop backgrond traffic flows.
  }
}

    ...

# User Traffic (Reno TCP)
set tcp_ [$ns_ create-connection TCP/Reno \
  $n(c) TCPSink $n(m) 1];               # Set up TCP connection between
                                        # Correspondent Node $n(c) and
                                        # Mobile Node $n(m).
$tcp_ set window_ $ws
set tcp_source_ [$tcp_ attach-app FTP]; # Bind FTP traffic to TCP

    ...
```

Figure 3.1.1-1: A glimpse of TCL code

between indirect and direct routing, the following limitations are imposed during a single simulation:

a. The average background traffic load on each link keeps same and equivalent with each other.

b. No alternative routing path is available whether how much the background traffic load is.

The above two limitations assume that the Internet traffic load and structure are uniformly distributed. We note that this assumption is away from the real features of the Internet with unbalanced structure and traffic distribution. However, this assumption would help us make sense of the difference between the indirect and direct routing in traffic performance.

3.1.1 Topology Structure

Four topology scenarios are designed according to various distributions of three studied nodes: Correspondent Gateway (CG), Home network Gateway (HG) and Visited network Gateway (VG - including mobile node). Figure 3.1.1-1 shows these four topologies. **Topology 1** indicates that the distance (number of hops) between CG and VG equals to the distance from CG to HG plus the distance from HG to VG, that is, the indirect routing path equals the direct routing path in length (number of hops). In the topology, we make the distance between HG and VG much longer than the distance between CG and HG. Therefore we can examine the influence of tunnelling (encapsulation) overhead on traffic performance. The direct routing path consists of 8 hops. The path between CG and HG consists of 1 hop and thus between HG and VG are there 7 hops. There is no relative difference in the path length between the two routings. **Topology 2** shows that the direct routing consists of 6 hops while the indirect routing consists of 9 hops. The length of indirect routing is 1.5 times as much as the length of direct routing. HG is located in the middle of the path. In the scenario the length of indirect routing is slightly longer than direct routing. The relative difference is 50%. **Topology 3** shows that direct routing consists of 6 hops while indirect routing consists of 12 hops. Indirect routing path is 2 times as much as direct routing path in length. HG is located in the middle of the path. In the scenario, the difference in path length between the two routings is intermediate. The relative difference is 100%. **Topology 4** indicates that indirect routing path is much longer than direct routing, about 5 times. Indirect routing consists of 16 hops while direct routing consists of 3 hops. Same as the above, HG is located in the middle of the path. In the scenario the relative difference is 433%. The reason that we design these four topology scenarios is to examine the effect of various relative differences between the two routings. For the narrowband mobile user (see page 6), the mobile node is connected to the VG via a GPRS link.

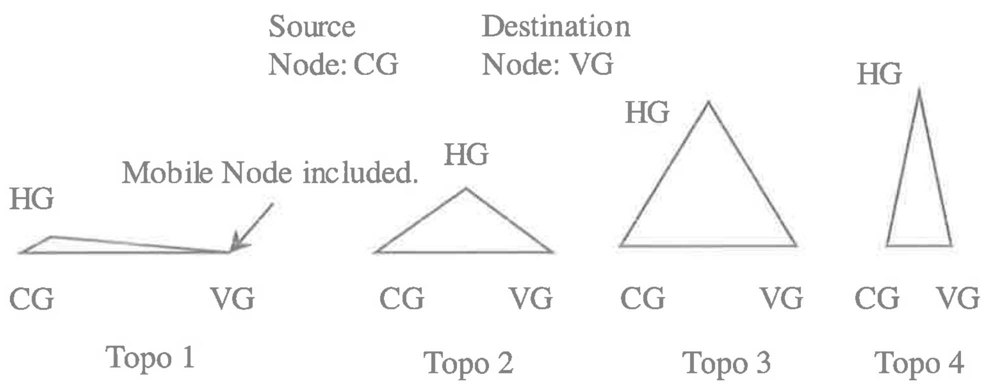


Figure 3.1.1-2: Location distributions between Correspondent Gateway (CG), Home network Gateway (HG) and Visited network Gateway (VG)

3.1.2 Link Feature

Link (hop) provides the ability to connect two neighboring nodes. All the links except GPRS radio link, which build up topology, hold the same link parameters. The link is duplex. The forward link and the backward link are separate. The link characters are described by the link delay (link distance), speed (capacity), buffer space, and link load.

The link delay is kept 4ms that corresponds a physical distance of around 1,000km. The capacity (speed) for each link is 2Mbits/s that follows T1(1.544Mb/s)/E1(2.048Mb/s) networks. In order to avoid high packet drop rate, buffer space is set to 300kbytes for each link since retransmission procedure would have influence on packet transfer delay. The drop-tail (FIFO) queuing discipline is adopted for buffer management. Link load means how much link capacity has been occupied. It is described by average background traffic load from 0.1 to 0.7 with an interval of 0.1. During a single simulation all links hold the same average link load.

The process of a packet of the studied TCP and UDP traffic passing through a link successively consists of enqueuing at the head node of the link, transferred by the node and propagating on the physical medium between the head and tail node.

The GPRS radio link [6] is not intended to be a realistic wireless link and implement the GPRS Um interface functions. It is intended as a simple link to emphasize the effect of bottleneck of GPRS radio link. Perfect link-layer error recovery is assumed. Coding scheme 2 is adopted to define the maximum bit rate, namely 13.3kbps per timeslot. The link delay is 0.01ms.

3.1.3 Background Traffic

Background traffic is aggregated by 5 traffic sources on each link. Each of these 5 traffic sources generates traffic according to a Pareto On/Off distribution. The average “on” and “off” time is 500ms. Both "on" and "off" periods are taken from a Pareto distribution. Packets are sent at a constant rate during the "on" period, and no packet is sent during the "off" period. The "shape" parameter used by the distribution is set to 1.4. Under 0.1 background traffic load, the data rate for each traffic source is 80kb/s. Therefore, under 0.2 through 0.7 background traffic load the data rate is 160kb/s, 240kb/s, 320kb/s, 400kb/s, 480kb/s and 560kb/s for each traffic source. Packets for the background traffic are constant size of 1kbytes.

Traditionally, exponential on/off distribution, which has a finite variance for the on/off periods, has been adopted to model wide-area traffic. Recent studies [27][28][29] on Internet traffic indicates that the distribution of on/off periods could have infinite variance as opposed to the finite variance assumption in the traditional voice traffic models. The Pareto distribution used in the simulations is a well-known heavy-tailed distribution. The "shape" parameter used in Pareto distribution describes the "heaviness" of the tail of the distribution. If the "shape" <= 2, the distribution has infinite variance; if the "shape" <= 1, then the distribution has infinite mean. The closer the "shape" is to 1, the more bursty the traffic is.

3.1.4 TCP and UDP Traffic Feature

TCP traffic feature

The TCP traffic is examined in the form of FTP application to simulate bulk data transfer. Reno TCP [16] [30] is employed, which incorporates slow start, congestion avoidance, fast retransmit, and fast recovery algorithms.

In contrast to the basic Tahoe TCP, fast recovery is integrated into the fast retransmit in Reno TCP. After fast retransmit sends what appears to be the missing packet, the Fast recovery algorithm governs the transmission of new packet until a non-duplicate ACK arrives, but slow start is not performed. This mechanism prevents the communication pipe from going empty after fast retransmit is invoked. Based on incoming duplicate ACKs, the TCP sender uses the fast retransmit algorithm to detect and recover packet loss. If three or more duplicate ACKs are received, that strongly indicates that a packet has been lost. Therefore, the fast retransmit algorithm uses the arrival of three duplicate ACKs as an indication of packet loss. The arrival of three duplicate ACKs means four identical ACKs without the arrival of any other intervening ACKs are received by the TCP sender. After receiving three duplicate ACKs, TCP retransmit the packet, which is indicated by duplicate ACKs as a dropped packet, without waiting for a retransmission timer to expire. Sequentially, the fast recovery takes over the transmission of new packets. The reason for performing fast recovery is that the receipt of duplicate ACKs not only indicates that a packet has been lost, but also that each duplicate ACK received represents a single packet having left the network. Therefore, the TCP doesn't want to reduce the congestion window (cwnd) size abruptly by going into slow start. During fast recovery the TCP sender increments the congestion window according to the number of received duplicate ACKs until a non-duplicate ACK is received where the TCP sender exits fast recovery and continues in congestion avoidance. The detailed operation of fast retransmit/recovery pair is as follows:

1. Once three duplicate ACKs are received, the TCP sender retransmits the lost packet and reduces its slow start threshold (ssthresh) to half current congestion window. Since three duplicate ACKs are received, the cwnd is set to ssthresh plus 3 times the packet size in order to reflect the three additional packets that have left the network.

2. For each additional duplicate ACK received, increment cwnd by the packet size. This increment of cwnd reflects the additional packet that has left the network. With the inflating of cwnd, when a new packet, which has not been transmitted before, is put in the congestion window, the TCP sender sends the packet. Since then, the TCP sender sends a new packet for each additional duplicate ACK that is received. If no any other packet is dropped since the lost packet, the TCP sender stops transmitting new packet since the retransmission of the lost packet until half a window of packets are acknowledged. This window is that at the receipt of the first duplicate ACK. Thus during the fast recovery, half the window minus one of packets are sent. Minus one is due to the retransmission of the lost packet.

3. Upon receipt of a non-duplicate ACK, the TCP sender exits fast recovery. TCP deflates cwnd to the ssthresh in step1 and performs congestion avoidance. This ACK is the acknowledge of the retransmission of the lost packet in step 1.

The simulations use one-way TCP sender and the TCP receiver that sends an ACK for every packet rather than for every two packets that it receives, the later is typical implementation for real-world. There is no SYN/FIN connection establishment/tear-down. Instead of dynamic window advertisement, a max bound on window size is introduced. The maximum window size for broadband mobile user is 64kbytes while is 8kbytes for narrowband mobile user. These two max bounds ensure no performance degradation due to excessive small window size. When no background traffic is introduced, the congestion window of 64 and 8kbytes can make the maximum usage of bottleneck capacity on routing path for broadband and narrowband mobile user respectively. Given that almost 100 percent of the packets are 1500 bytes (characteristic of Ethernet-attached hosts) or smaller while general GPRS packets are 500-1000 bytes, the TCP packet size is set to 1000 bytes [26]. That is, the maximum window size is 64 and 8 packets for broadband and narrowband mobile user respectively. The TCP connection time is 100s. The duration of 100s is intended to be an intermediate choice from a wide range in average flow duration for FTP from 20-500s, which is reported by [26].

UDP traffic feature

No application layer protocol such as Real-Time Streaming Protocol (RTSP) or transport layer protocol such as Real-time Transport Protocol is introduced in the simulations. The UDP [17] stream is transmitted at Constant Bit Rate (CBR). The bit rate is 28.8kb/s, which is usually used by real video stream. UDP packet size is set to 180bytes. The transmission time is 100s.

3.2 Statistical Methods

The simulations concentrate on studying the performance difference between the indirect and direct routing. It is important to verify if the difference of simulation results between the two routings is statistically significant. The following describes the statistical method used in analyzing the simulation results presented in Chapter 4.

T test

We introduce t test to compare the means of two populations (i.e. w_1 and w_2 shown below) respectively from direct and indirect routing and assume that either the populations are normally distributed or the sample sizes for both populations are large enough so that the Central Limit Theorem (CLT) is effective. If we use $E(\overline{x})$ to denote the expected value of the difference between the two populations, then we test the Null Hypothesis

$$H_0: \quad E(\overline{x}) = 0$$

against the Alternative Hypothesis

$$H_1: \quad E(\overline{x}) > 0.$$

Given that the $E(\overline{x})$ is the difference of two unpaired populations (the individuals in one population are not 'matched' or 'paired' with those in the other population) and the two populations are unlikely to have the same variance, the solution suggested by Welch (1938) [43] is adopted to determine the confidence interval in this case. The suggested methods for the computing of the estimated mean ($\overline{x}$), variance (s^2) and number of degrees of freedom ($\hat{f}$) of the difference are presented as follows.

From the simulation, we get:

w_1: the samples of measured parameter from direct routing (e.g. goodput, sample mean and variance of packet transfer delay, retransmission rate, etc. from individual simulation runs)

w_2: the samples of measured parameter from indirect routing

n_1: sample size of w_1

n_2: sample size of w_2

$\overline{w}_1(n_1)$: the sample mean of $w_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} w_1(i)$

$\overline{w}_2(n_2)$: the sample mean of w_2 $= \frac{1}{n_2}\sum_{i=1}^{n_2} w_2(i)$

$s_1^2(n_1)$: the sample variance of w_1 $= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}[w_1(i) - \overline{w}_1(n_1)]^2$

$s_2^2(n_2)$: the sample variance of w_2 $= \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}[w_2(i) - \overline{w}_2(n_2)]^2$.

Then we focus on the difference between the two routings:

$\overline{x}$: the estimated mean of the difference between $\overline{w}_1(n_1)$ and $\overline{w}_2(n_2)$, and is expressed as*

$$\overline{x} = \overline{w}_1(n_1) - \overline{w}_2(n_2), \quad \textit{if} \quad \overline{w}_1(n_1) > \overline{w}_2(n_2)$$

$$\overline{x} = \overline{w}_2(n_2) - \overline{w}_1(n_1), \quad \textit{if} \quad \overline{w}_1(n_1) < \overline{w}_2(n_2).$$

s^2: the estimated variance of $\overline{x}$ and is expressed as

$$s^2 = \frac{s_1^2(n_1)}{n_1} + \frac{s_2^2(n_2)}{n_2}.$$

$\hat{f}$: the estimated number of degrees of freedom (df) of $\overline{x}$ and is expressed as

$$\hat{f} = \frac{[s_1^2(n_1)/n_1 + s_2^2(n_2)/n_2]^2}{[s_1^2(n_1)/n_1]^2/(n_1 - 1) + [s_2^2(n_2)/n_2]^2/(n_2 - 1)}.$$

Thus, a *100(1-α)%* confidence interval for $\overline{x}$ is given by

$$\left(\overline{x} - t_{\hat{f},1-\alpha/2} \cdot \sqrt{s^2},\ \overline{x} + t_{\hat{f},1-\alpha/2} \cdot \sqrt{s^2}\right),$$

and Margin of Error (MoE) is presented by

$$\textit{Margin of Error (MoE)} = t_{\hat{f},1-\alpha/2} \cdot \sqrt{s^2} \,,$$

where the t value is defined by $\hat{f}$ and $1-\alpha/2$ from the t table [44]. $\alpha = 0.05$ is adopted in Chapter 4 for hypothesis testing.

The relative difference, θ, between the indirect and direct routing is defined as follows:*

$$\theta = \frac{\overline{x}}{\min[\overline{w}_1(n_1), \overline{w}_2(n_2)]} \times 100\% \,.$$

The normalised confidence interval for the relative difference is presented by*

$$\left(\frac{\overline{x} - t_{\hat{f},1-\alpha/2} \cdot \sqrt{s^2}}{\min[\overline{w}_1(n_1), \overline{w}_2(n_2)]}, \frac{\overline{x} + t_{\hat{f},1-\alpha/2} \cdot \sqrt{s^2}}{\min[\overline{w}_1(n_1), \overline{w}_2(n_2)]} \right) \times 100\% \,.$$

*(Note: The comparison of $\overline{w}_1(n_1)$ and $\overline{w}_2(n_2)$ is only accurate to a certain extent: it is mainly based on the simulation results from Topology 2-4 *excluding* Topology 1. If the results show that in most cases the $\overline{w}_1(n_1)$ is larger than $\overline{w}_2(n_2)$, then we assume $\overline{w}_1(n_1) > \overline{w}_2(n_2)$. Therefore in most cases, $\overline{x}$ is a positive value.)

Chapter 4

Simulation Results

This chapter discusses simulation results based on traffic performance, and examines the effects of indirect and direct routing on traffic performance. This chapter consists of two major parts - TCP and UDP performance. Each part is divided into traffic performance for broadband and narrowband mobile user.

4.1 UDP Performance

In this section, we present detailed discussion of simulation results based on UDP performance. For real-time applications, staple and short packet transfer delay and low packet drop rate are desired. According to the simulation environment described in Chapter 3, it can be expected that in a long-term view the packet drop rate is proportional to the length of routing path. Therefore, we only take packet transfer delay and its standard deviation as the performance measures. Packet transfer delay is the time between a packet is put into queue at source node and the packet is completely received at destination node.

4.1.1 Performance for Broadband Mobile User

The simulations focus on measuring the performance of a flow of data stream with constant bit rate - 28.8kb/s on the four studied topologies. We perform 30 simulation runs for each background traffic load through 0.1 to 0.7, with an interval of 0.1. If let i denote background traffic load and j be the jth simulation under the background traffic load i, then we have:

$$\overline{D_i} = \frac{1}{n}\sum_{j=1}^{n} D_{ij} \text{ and } \overline{S_i^2} = \frac{1}{n}\sum_{j=1}^{n} S_{ij}^2$$

where D_{ij} and S_{ij} are respectively the sample mean and standard deviation of packet transfer delay from the jth simulation for background traffic load i. Here, n = 30. Therefore $\overline{D}_i$ and $\overline{S_i^2}$ respectively denote the pooled sample mean and pooled sample variance of packet transfer delay for background traffic load i. From a single simulation run we can obtain the packet transfer delay for each packet. And then a sample mean (D_{ij}) and variance (S_{ij}^2) are figured out from this group of packet transfer delays on behalf of this simulation. Thus, 30 sample means (D_{ij}) and 30 sample variances (S_{ij}^2) of packet transfer delay are available for each background traffic load. We then focus on these 30 sample means and variances respectively. A pooled sample mean ($\overline{D}_i$) and sample standard deviation are achieved from the 30 sample means while a pooled sample variance (sample mean) ($\overline{S_i^2}$) and sample standard deviation are available from the 30 sample standard deviations. We can obtain a pooled sample standard deviation $\overline{S}_i$ from the following formula

$$\overline{S}_i = \sqrt{\overline{S_i^2}} .$$

In convenience of description, the pooled sample mean ($\overline{D}_i$), standard deviation ($\overline{S}_i$) and variance ($\overline{S_i^2}$) are called henceforth as estimated mean of packet transfer delay, estimated mean of standard deviation of packet transfer delay and estimated variance of packet transfer delay respectively. $\overline{D}_i$ and $\overline{S}_i$ ($\overline{S_i^2}$) are respectively discussed in the following two sections.

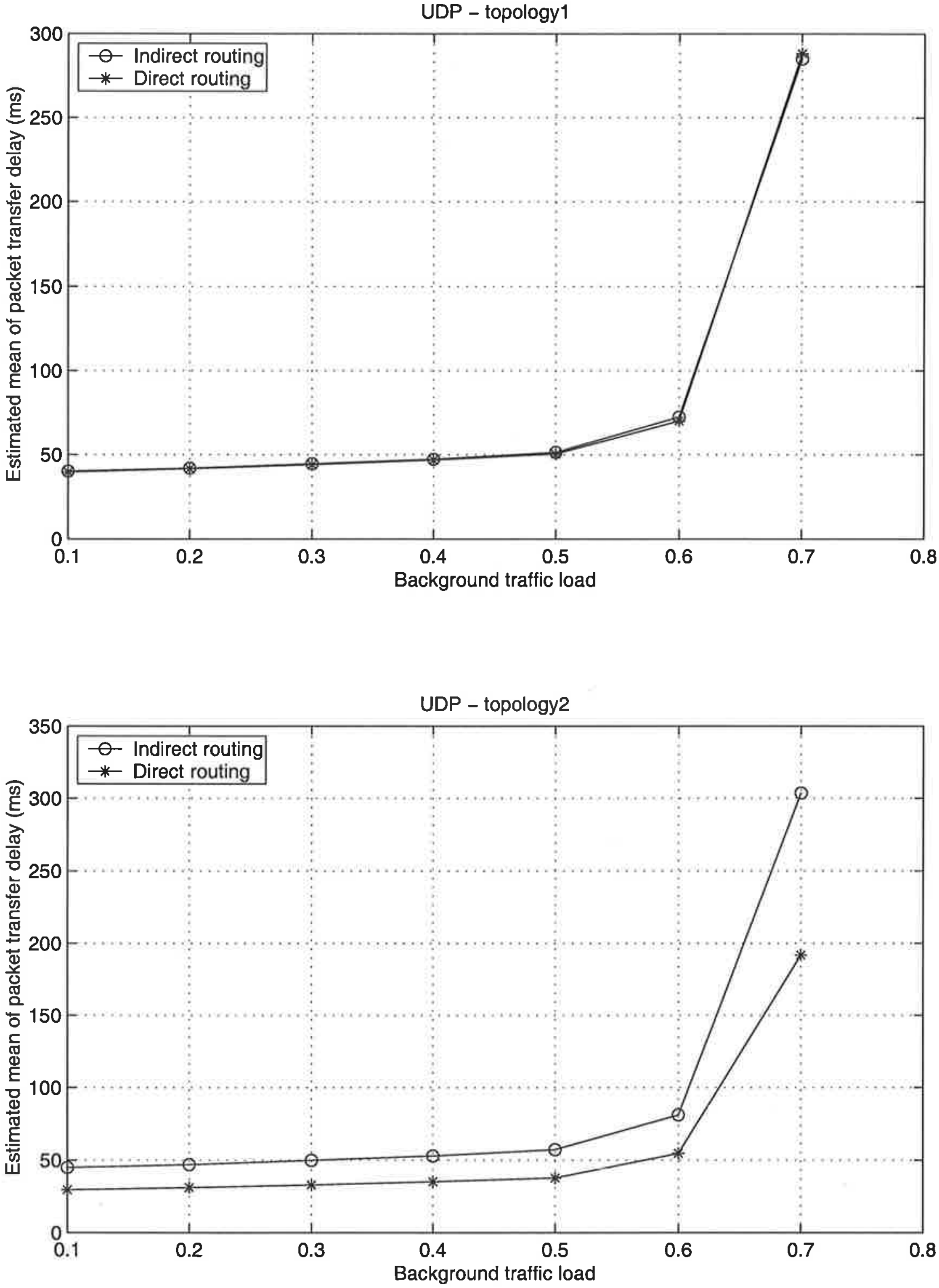
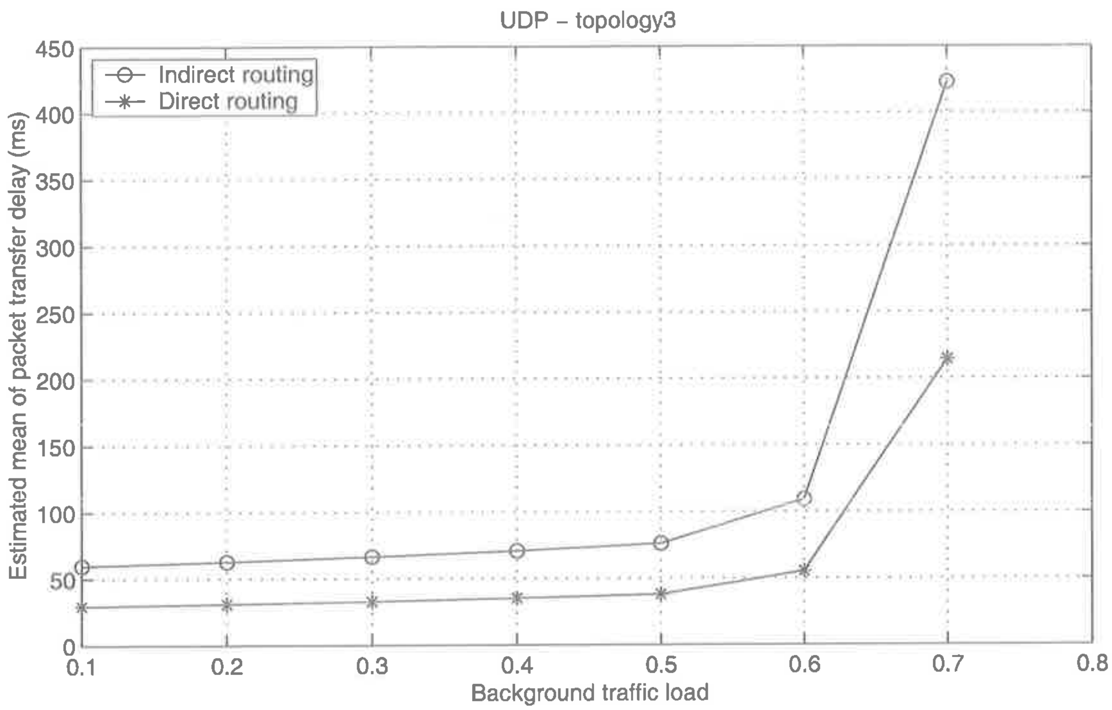


Figure 4.1.1-1: Packet transfer delay on Topology 1 and 2



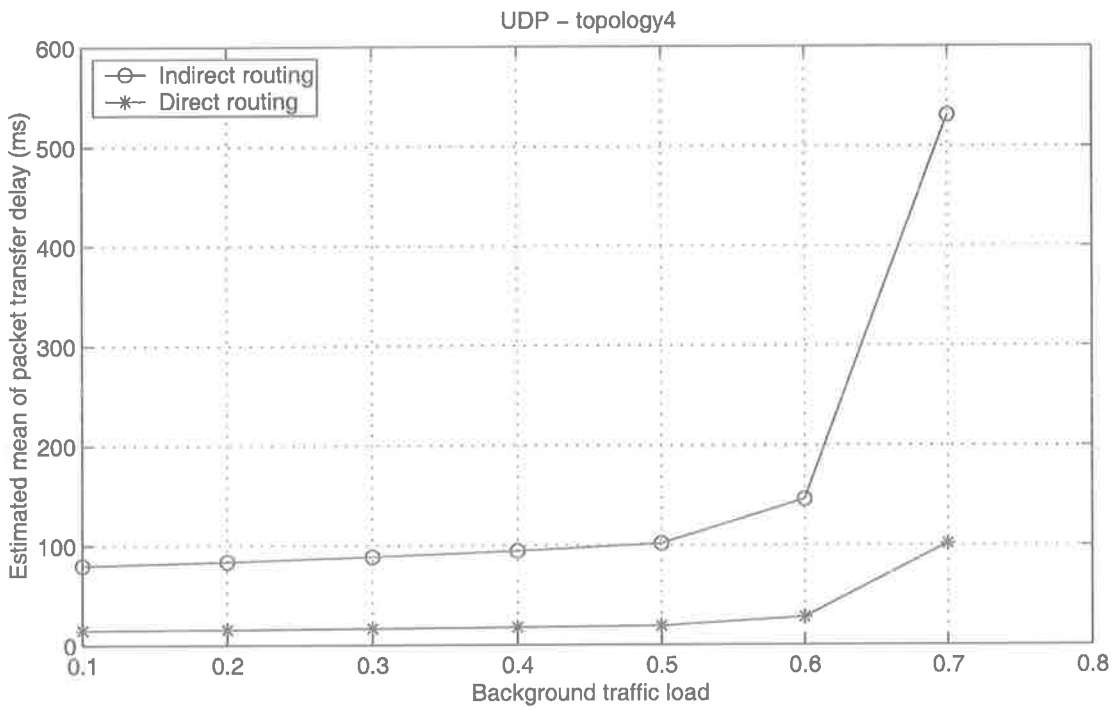


Figure 4.1.1-2: Packet transfer delay on Topology 3 and 4

Packet transfer delay

Table A.1-1 and A.1-2 (Appendix A) list the estimated mean and standard deviation of packet transfer delay under indirect and direct routing respectively. Figure 4.1.1-1 and 4.1.1-2 plot packet transfer delay as a function of background traffic load on the four studied topologies. It is shown that below 0.5 background traffic load there is slow and linear growth of packet transfer delay with the increasing of background traffic load. From 0.5 to 0.7 background traffic load, the growth of packet transfer delay is dramatic and exponential. Similarly, the difference between the indirect and direct routing increases slowly below 0.5 background traffic load while dramatically from 0.5 to 0.7 background traffic load. It is also shown in Table 4.1.1-1. Therefore, as to the relative difference there is no clear difference between background traffic loads. In fact, we cannot find a statistically significant difference as shown later. Figure 4.1.1-3 supports this point clearly. Table A.1-3 (Appendix A) lists the relative difference between the two routings in packet transfer delay. The relative difference θ_d is presented by

$$\theta_d = \frac{d_{ind} - d_d}{d_d} \times 100\% ,$$

where d_{ind} and d_d denote packet transfer delay under indirect and direct routing respectively.

In the following, we explain why the packet transfer delay increases slowly below 0.5 background traffic load and dramatically above the load. The peak background traffic load in the simulation is equal to the double of the average load. That is, below 0.5 average load the practical load is always within the link capacity, therefore there is almost no overload on each link. In this case the queuing delay is not significant although the delay is increasing with the background traffic load. As to 0.6 and 0.7 average load, the peak background load is 1.2 and 1.4 times of the link capacity. In this case, the queuing delay becomes significant and takes more weight in the packet transfer delay. However, this dramatic change at the 0.5 background traffic load doesn't create any significant effect on the relative difference. From Table A.1-3, the relative difference for the later three topologies is within the range of 48.60-58.36%, 97.32-102.25% and 423.78-434.56% respectively. That basically responds the relative difference between the two routings in routing path length, which is 50%, 100% and 433% respectively. In the term of packet transfer delay, it is clear that the UDP performance is proportional to the length of routing path.

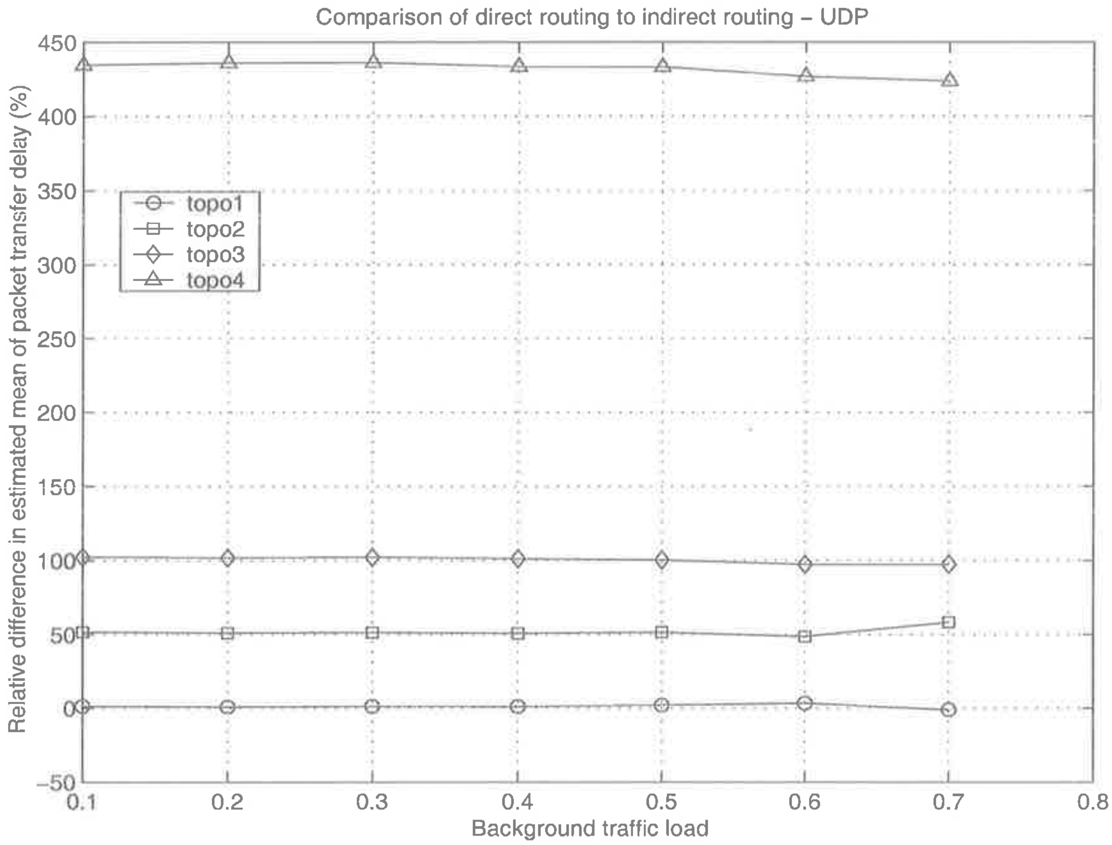


Figure 4.1.1-3: Relative difference between indirect and direct routing in packet transfer delay

We put the difference into the hypothesis testing which is described in chapter 3. Table 4.1.1-1 and 4.1.1-2 list a 95% confidence interval for the average difference and normalised confidence interval for relative difference respectively. It is indicated that in most cases the null hypothesis can be rejected at a significance level of 2.5% except for topology 1 under 0.6 and 0.7 background traffic load. The encapsulation overhead still has little effect on traffic performance, although it has significant effect on transmission delay, which is one of packet transfer delay's three elements - queuing delay, transmission delay and propagation delay. The encapsulation overhead is 48 bytes per packet. The original packet size is 180. Therefore, in the routing path after encapsulation there is 26.67% degradation of transmission delay, however it is hidden by the queuing and propagation delay. The margin of error becomes larger with the increasing of background traffic load, but overall the relative difference keeps constant for the later three topologies and responds to the relative difference in routing path length. We note that under low background traffic load the down limit of the confidence interval of the relative difference (Table 4.1.1-2) is higher than the relative difference between the two routings in the routing path length for the later three topologies, which is 50%, 100% and 433% respectively. We attribute this situation to the encapsulation overhead.

Table 4.1.1-1: 95% confidence interval for the difference between indirect and direct routing in packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	0.53	0.35	0.53	0.51	0.99	2.34	-3.12
	Topo2	15.30	15.86	16.93	17.85	19.45	26.55	111.88
	Topo3	30.22	31.69	33.58	35.57	38.10	53.83	208.50
	Topo4	64.74	67.99	72.01	76.62	82.54	118.35	429.76
Std (ms)	Topo1	0.088	0.15	0.11	0.19	0.22	2.07	22.18
	Topo2	0.054	0.12	0.11	0.22	0.18	2.50	27.20
	Topo3	0.058	0.17	0.11	0.22	0.19	3.33	27.19
	Topo4	0.074	0.16	0.11	0.20	0.25	3.14	27.87
df	Topo1	57.0	57.7	55.8	55.1	58.0	57.7	56.1
	Topo2	55.9	58.0	54.9	57.4	50.3	51.3	53.7
	Topo3	40.4	47.4	52.3	56.8	53.6	46.7	38.8
	Topo4	49.7	45.3	41.8	50.3	42.9	50.6	40.0

(Continued)

Table 4.1.1-1 (continued)

MoE (ms)	Topo1		0.18	0.30	0.22	0.37	0.45	4.15	44.46
	Topo2		0.11	0.23	0.22	0.45	0.36	5.02	54.51
	Topo3		0.12	0.35	0.22	0.43	0.39	6.70	54.95
	Topo4		0.15	0.32	0.23	0.40	0.51	6.30	56.33
Confidence interval (ms)	Top1	Up	0.71	0.65	0.75	0.88	1.44	6.48	41.33
		Down	0.35	0.056	0.31	0.13	0.54	-1.81	-47.58
	Top2	Up	15.40	16.09	17.14	18.30	19.81	31.57	166.39
		Down	15.19	15.63	16.71	17.40	19.10	21.53	57.38
	Top3	Up	30.33	32.04	33.80	36.01	38.49	60.53	263.45
		Down	30.10	31.34	33.36	35.14	37.71	47.13	153.56
	Top4	Up	64.89	68.32	72.24	77.03	83.05	124.65	486.09
		Down	64.59	67.67	71.79	76.22	82.03	112.05	373.43

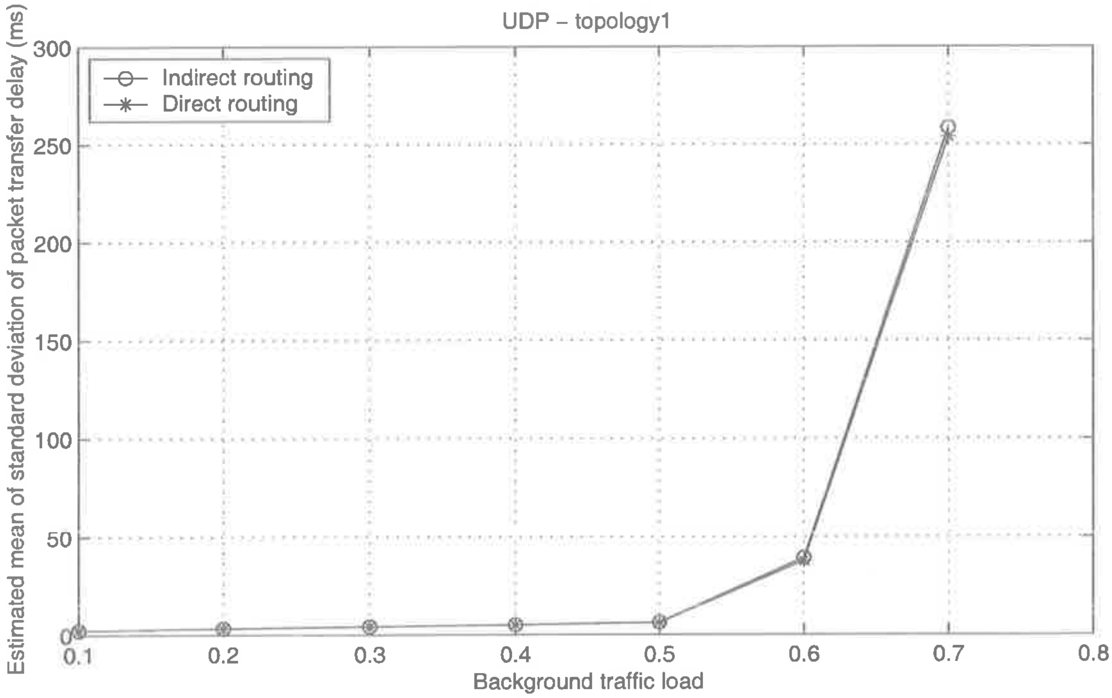
Table 4.1.1-2: Normalised confidence interval for relative difference between indirect and direct routing in packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Confidence interval (%)	Topo	Up	1.78	1.55	1.70	1.88	2.86	9.25	14.34
		Down	0.89	0.13	0.70	0.29	1.08	-2.59	-16.51
	Topo	Up	52.08	51.73	52.07	52.04	52.43	57.79	86.80
		Down	51.34	50.25	50.77	49.49	50.54	39.42	29.93
	Topo	Up	102.65	102.92	102.84	102.47	101.46	109.55	122.97
		Down	101.86	100.66	101.50	100.01	99.41	85.29	71.67
	Topo	Up	435.56	438.09	437.57	435.74	435.97	449.64	479.33
		Down	433.57	433.95	434.81	431.16	430.60	404.18	368.23

Standard deviation of packet transfer delay

We have obtained 30 sample variances of packet transfer delay (S_{ij}^2) under each background traffic load from 30 simulation runs. Then, an estimated mean ($\overline{S_i^2}$) and standard deviation are figured out for these 30 sample variances (S_{ij}^2). Sequentially, an estimated mean of standard deviation of packet transfer delay ($\overline{S}_i$) is derived from $\overline{S}_i = \sqrt{\overline{S_i^2}}$. In this section we compare the $\overline{S}_i$ and $\overline{S_i^2}$ between the indirect and direct routing.

Table A.1-4 and A.1-5 (Appendix A) list $\overline{S}_i$ under indirect and direct routing respectively. Figure 4.1.1-4 and 4.1.1-5 plot $\overline{S}_i$ as a function of background traffic load. The curves go in the similar way with packet transfer delay in the last section. The $\overline{S}_i$ holds a slow and linear growth below 0.5 background traffic load and a dramatic growth after the load. The dramatic change beginning at 0.5 background traffic load is due to the overload created by 0.6 and 0.7 background traffic load. The difference between the two routings goes in the same way. Consequently, there is no significant difference between various background traffic loads in the relative difference shown in Table A.1-6 (Appendix A) and Figure 4.1.1-6 for the individual topologies. That is the same as we find in the packet transfer delay, but with a difference in the value of relative difference. The relative difference in packet transfer delay is about 50%, 100% and 433% for topology 2, 3 and 4 respectively. It is equal to the relative difference in routing path length. The relative difference in $\overline{S}_i$ is about 22%, 42% and 130% for these three topologies. We cannot see its direct relationship to the relative difference in routing path length.



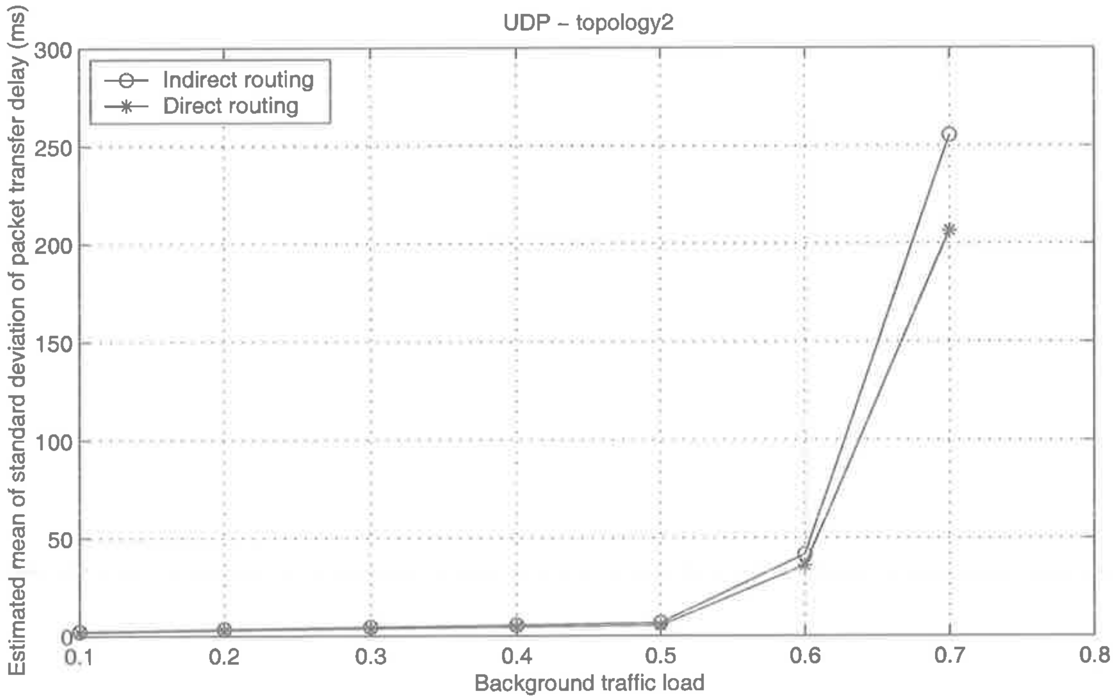
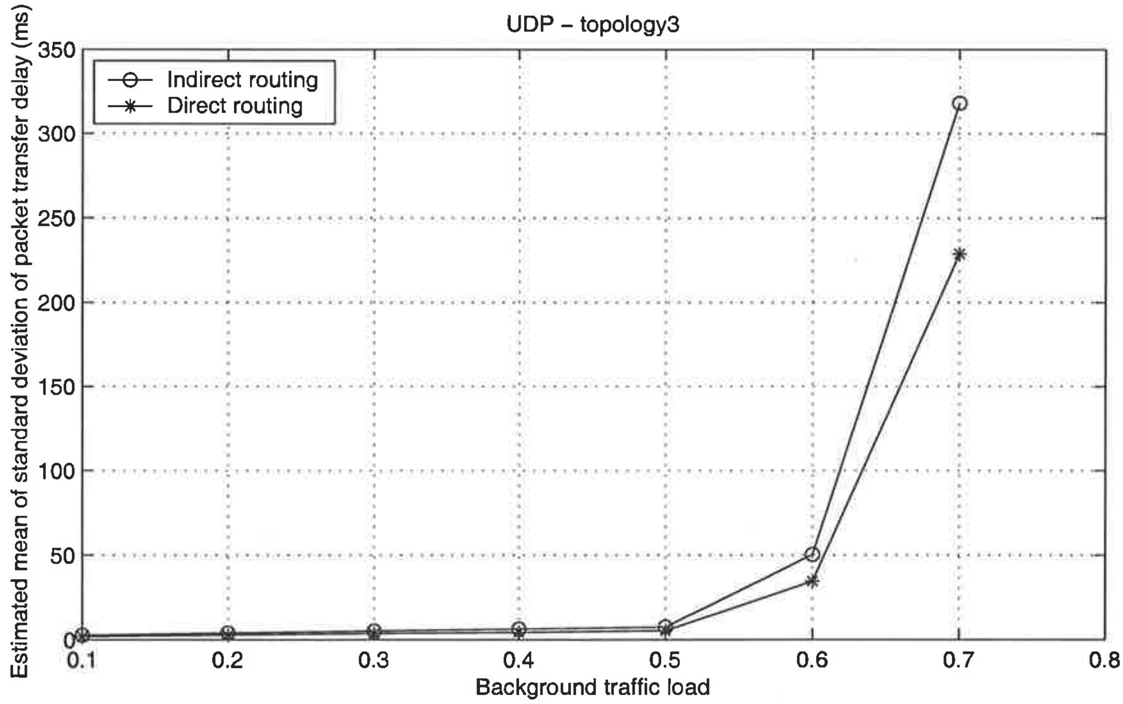


Figure 4.1.1-4: Standard deviation of packet transfer delay on Topology 1 and 2



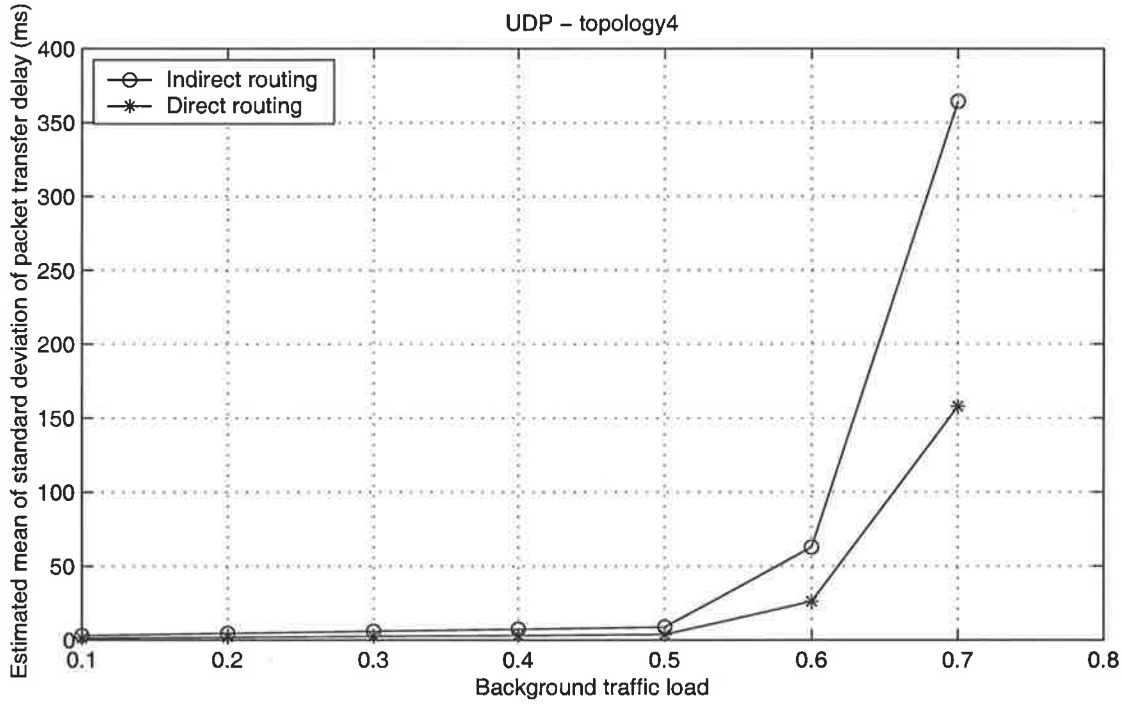


Figure 4.1.1-5: Standard deviation of packet transfer delay on Topology 1 and 2

To explain the difference, we introduce "variance". Figure 4.1.1-7 shows the relative difference between the two routings in the variance of packet transfer delay as a function of background traffic load. It is shown that the relative difference in variance is proportional to the relative difference in the length of routing path between the two routings. Table A.1-7 (Appendix A) lists the relative difference between the two routings in variance of packet transfer delay. We use V(X) and E(X) to denote the variance and expected value of the random variable X respectively. If X and Y are uncorrelated, then we have

$$V(X + Y) = E((X + Y)^2) - (E(X + Y))^2 = V(X) + V(Y).$$

We have known whether indirect or direct routing path consists of a number of hops and these hops hold the same features and create their own background traffic independently. Then the packet transfer delay is equal to the total of delay created on each hop. If *Dd* and *Di* denote packet transfer delay under direct and indirect routing respectively, and Dh_i denotes the delay on the *i*th hop on the routing path, then we have

$$Dd = \sum_{i=1}^{n} Dh_i$$

$$Di = \sum_{i=1}^{m} Dh_i.$$

Here, *m* and *n* denote the number of hops under indirect and direct routing respectively. Using $V(X+Y) = V(X) + V(Y)$, this becomes

$$V(Dd) = \sum_{i=1}^{n} V(Dh_i)$$

$$V(Di) = \sum_{i=1}^{m} V(Dh_i)$$

Given that background traffic is created based on each individual hop (link), in term of statistics the $E(Dh_i)$ and $V(Dh_i)$ can be considered being independent of the value of *i*. Therefore, if *Dh* denotes the random variable of packet transfer delay on hop, then the expected mean and variance of *Dh* are *E(Dh)* and *V(Dh)*. Thus, we have

$$V(Dd) = n \cdot V(Dh)$$
$$V(Di) = m \cdot V(Dh).$$

In the same way, we have

$$E(Dd) = n \cdot E(Dh)$$
$$E(Di) = m \cdot E(Dh). \quad (4.1\text{-}1)$$

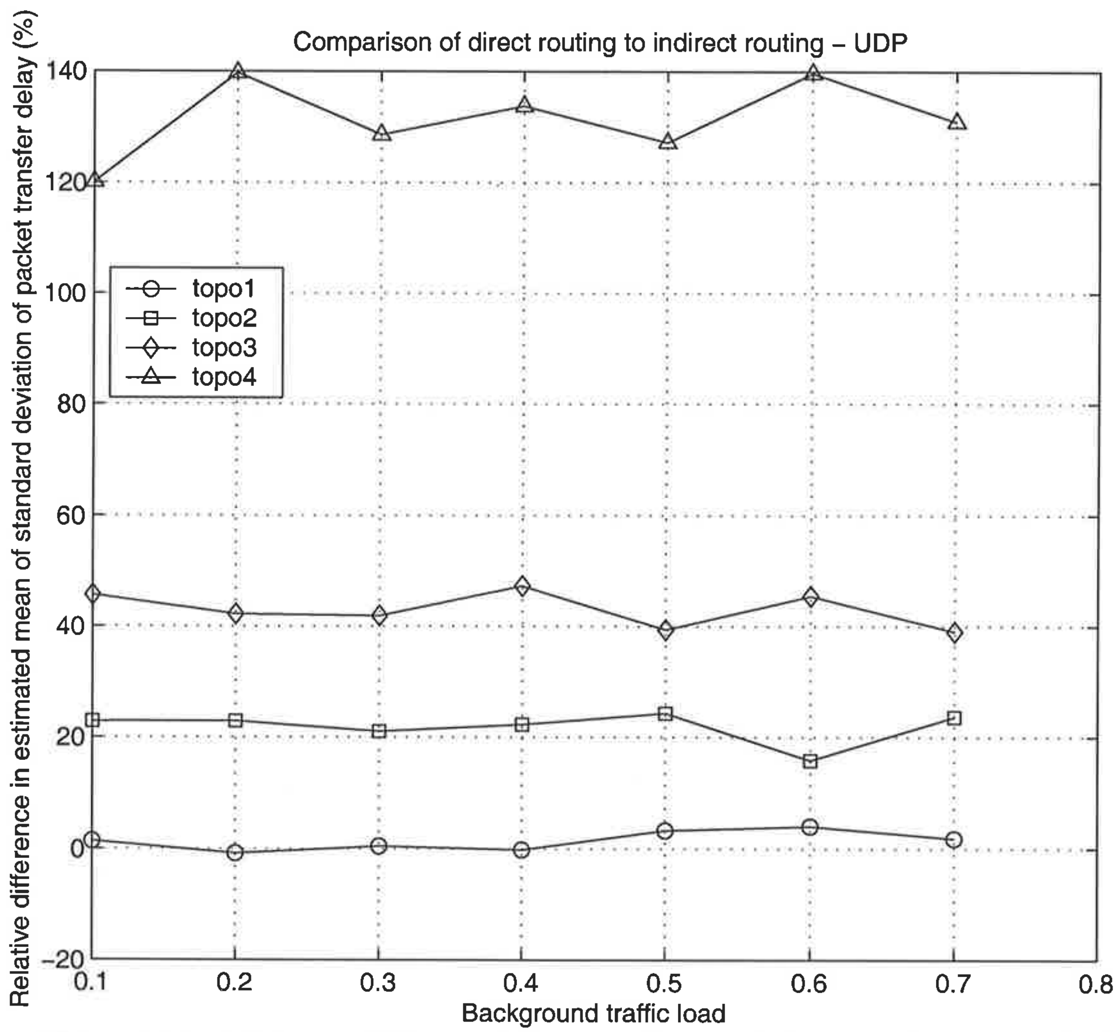


Figure 4.1.1-6: Relative difference between indirect and direct routing in standard deviation of packet transfer delay

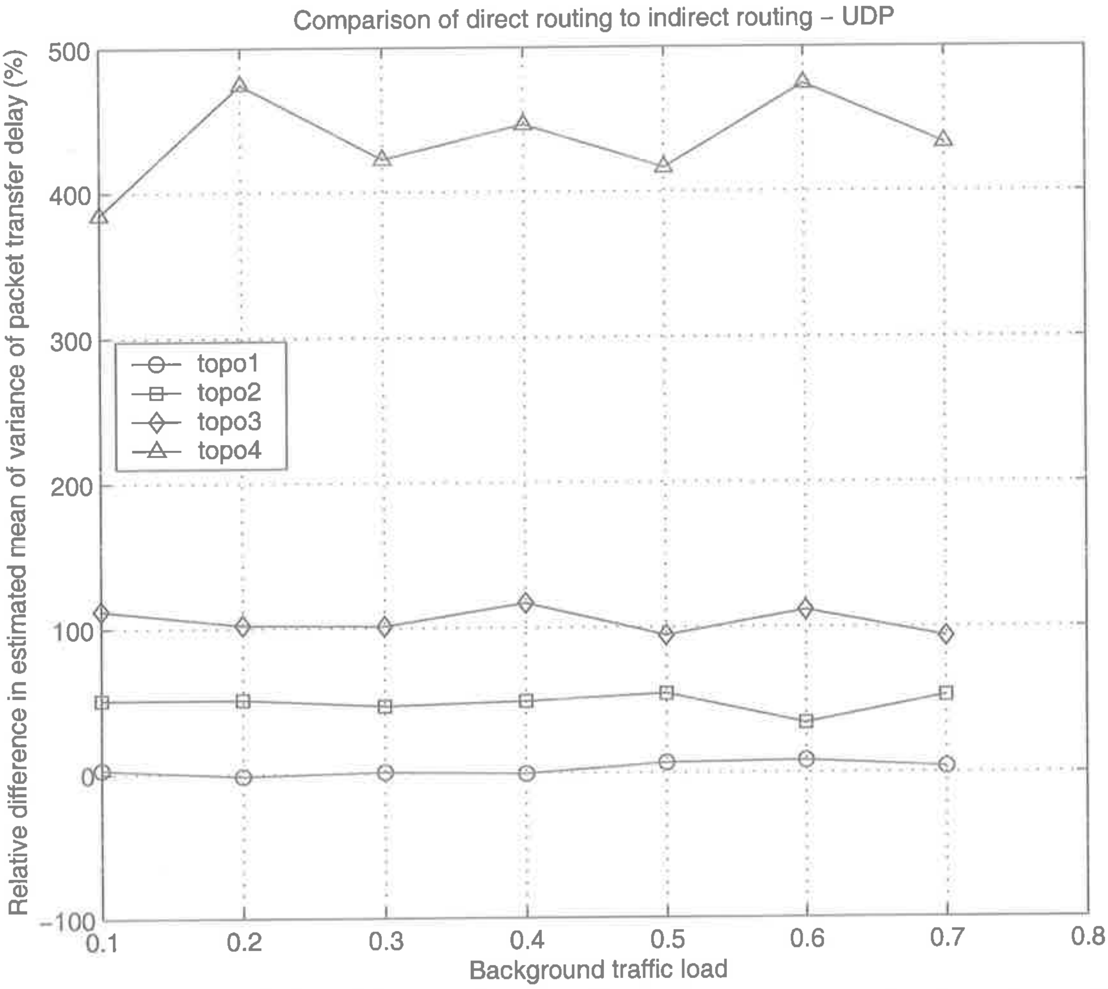


Figure 4.1.1-7: Relative difference between indirect and direct routing in variance of packet transfer delay

It is clear that both the expected mean and variance of packet transfer delay is proportional to the length of routing path. That confirms the observation from Figure 4.1.1-3 and 4.1.1-7. The relative difference θ_s in standard deviation is thus figured out:

$$\theta_s = \frac{s_{ind} - s_d}{s_d} = \frac{\sqrt{V(Di)} - \sqrt{V(Dd)}}{\sqrt{V(Dd)}} = (\sqrt{\frac{m}{n}} - 1) \times 100\%. \qquad (4.1\text{-}2)$$

Here, s_{ind} and s_d denote the standard deviation of packet transfer delay under indirect and direct routing respectively. The computed value of θ_s is 22.47%, 41.42% and

130.94% for topology 2,3 and 4 respectively. The simulation results approximately confirm the computed values of θ_s.

Table 4.1.1-3: 95% confidence interval for difference between indirect and direct routing in variance of packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms^2)	Topo1		0.13	-0.17	0.17	-0.09	2.57	117.6	2350
	Topo2		1.78	4.05	6.47	10.01	15.49	441.6	22560
	Topo3		3.88	8.13	13.67	22.07	27.63	1340.1	48830
	Topo4		7.38	17.87	29.08	44.42	63.25	3254.7	107880
Std (ms^2)	Topo1		0.18	0.52	0.45	0.80	0.89	419.17	10830
	Topo2		0.15	0.36	0.51	0.91	0.83	401.99	15930
	Topo3		0.19	0.52	0.47	1.27	0.93	615.43	15780
	Topo4		0.28	0.68	0.50	1.43	1.37	671.02	21230
df	Topo1		57.4	49.8	54.4	51.6	56.8	56.5	51.5
	Topo2		51.2	55.6	50.7	55.4	46.6	54.9	58.0
	Topo3		40.8	47.7	56.2	37.2	50.2	41.8	37.6
	Topo4		46.6	34.4	41.8	31.1	34.8	47.3	35.8
MoE (ms^2)	Topo1		0.36	1.05	0.91	1.60	1.79	840.0	21750
	Topo2		0.31	0.73	1.03	1.82	1.68	805.6	31860
	Topo3		0.39	1.05	0.94	2.57	1.86	1243.8	31880
	Topo4		0.57	1.39	1.00	2.92	2.78	1351.4	43100
Confidence interval (ms^2)	Top1	Up	0.49	0.88	1.08	1.51	4.36	957.6	24100
		Down	-0.23	-1.22	-0.74	-1.69	0.77	-722	-19410
	Top2	Up	2.09	4.78	7.50	11.82	17.17	1247.1	54430
		Down	1.47	3.32	5.45	8.19	13.81	-364.0	-9300
	Top3	Up	4.26	9.18	14.60	24.64	29.49	2583.9	80710
		Down	3.49	7.09	12.73	19.50	25.76	96.3	16950
	Top4	Up	7.95	19.26	30.08	47.34	66.04	4606.1	150980
		Down	6.81	16.48	28.07	41.49	60.47	1903.3	64780

Table 4.1.1-4: Normalised confidence interval for relative difference between indirect and direct routing in variance of packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Confidence interval (%)	Topo1	Up	10.38	8.33	5.96	5.76	11.40	66.98	37.39
		Down	-4.83	-11.62	-4.08	-6.44	2.02	-50.53	-30.11
	Topo2	Up	59.74	60.27	53.89	58.57	60.46	96.45	127.53
		Down	42.13	41.91	39.12	40.57	48.63	-28.15	-21.79
	Topo3	Up	123.49	115.35	108.27	130.60	100.67	215.15	154.35
		Down	101.09	89.05	94.38	103.36	87.94	8.02	32.41
	Topo4	Up	414.39	511.29	437.12	475.89	434.66	671.33	606.32
		Down	354.67	437.62	407.96	417.13	398.03	277.39	260.14

Table 4.1.1-3 and 4.1.1-4 list a 95% confidence interval for the difference and normalised confidence interval for the relative difference between the two routings in standard deviation of packet transfer delay. It is shown that the difference on topology 1 is unclear and we cannot reject the null hypothesis at a significance level of 2.5%. That indicates that the effect from encapsulation overhead is negligible, although the overhead takes 26.67% space of packet size. Therefore, its effect cannot be shown on the other three topologies. From topology 2 to 4, the difference becomes more and more clear. Because of stronger effect from high load of background traffic, we also cannot reject the null hypothesis at the significance level, for topology 2 under 0.6 - 0.7 background traffic load.

4.1.2 Performance for Narrowband Mobile User

In this section we use GPRS functionality as an example. The same UDP data stream used for the broadband mobile user is put on the studied topologies. We assume that enough timeslots are available for the transmission. That is, 28.8kb/s data stream asks for at least 3 timeslots (39.9kb/s) to be available for the mobile user. The traffic performance is measured in two ways: the first one is to keep the GPRS radio capacity fixed for the end user while changing the background traffic load from 0.1 to 0.8, the other one is to keep the background traffic load constant while changing the number of timeslots available from 3 to 8. Similar to the analysis in 4.1.1, we focus on packet transfer delay ($\overline{D_i}$) and its standard deviation ($\overline{S}_i$) respectively, which are described in 4.1.1 with a difference that in the cases of fixed background traffic load, the *i* denotes number of timeslots rather than background traffic load.

Packet transfer delay

A. Fixed GPRS radio link capacity

Table A.2-1 and A.2-2 (Appendix A) list the estimated mean ($\overline{D_i}$) and standard deviation of packet transfer delay respectively under indirect and direct routing. Figure 4.1.2-1 and 4.1.2-2 plot the packet transfer delay as a function of background traffic load. Under low background traffic load, there is a slow and approximately linear growth. When the background traffic load is after 0.5, the growth becomes dramatic and exponential. As described in section 4.1.1, this change is due to the overload which only happens when the background traffic load is larger than 0.5. Similarly, the difference between the two routings is small below 0.5 background load, and large after the load. However, different from the performance for broadband mobile user where the relative difference keeps constant between various background traffic loads, Table A.2-3 (Appendix A) and Figure 4.1.2-3 show the relative difference increases with the increment of background traffic load. The difference is due to the introduction of the GPRS radio link, which is of limited capacity. Here, we use Dg to denote the transfer delay produced by the radio link. Integrated with equation (4.1-1), we have

$$E(Dd) = n \cdot E(Dh) + Dg$$
$$E(Di) = m \cdot E(Dh) + Dg.$$

Therefore, the relative difference θ_d can be expressed as follows:

$$\theta_d = \frac{E(Di) - E(Dd)}{E(Dd)}$$
$$= \frac{(m-n) \cdot E(Dh)}{n \cdot E(Dh) + Dg} \times 100\%.$$

Considering $E(Dh) \to \infty$ and $\to 0$, this becomes

$$\lim_{E(Dh) \to \infty} \theta_d(E(Dh)) = \frac{m-n}{n} \times 100\%$$
$$\lim_{E(Dh) \to 0} \theta_d(E(Dh)) = 0 \times 100\%.$$

Therefore,

$$\theta_d \in (0 \times 100\%, \frac{m-n}{n} \times 100\%).$$

We have known the relative difference between n and m is 50%, 100% and 433% respectively for topology 2, 3 and 4. Therefore with the increasing of background traffic load, the relative difference θ_d grows within $(0 \times 100\%, \frac{m-n}{n} \times 100\%)$. The simulation results show that the growth is clearer after 0.5 background traffic load.

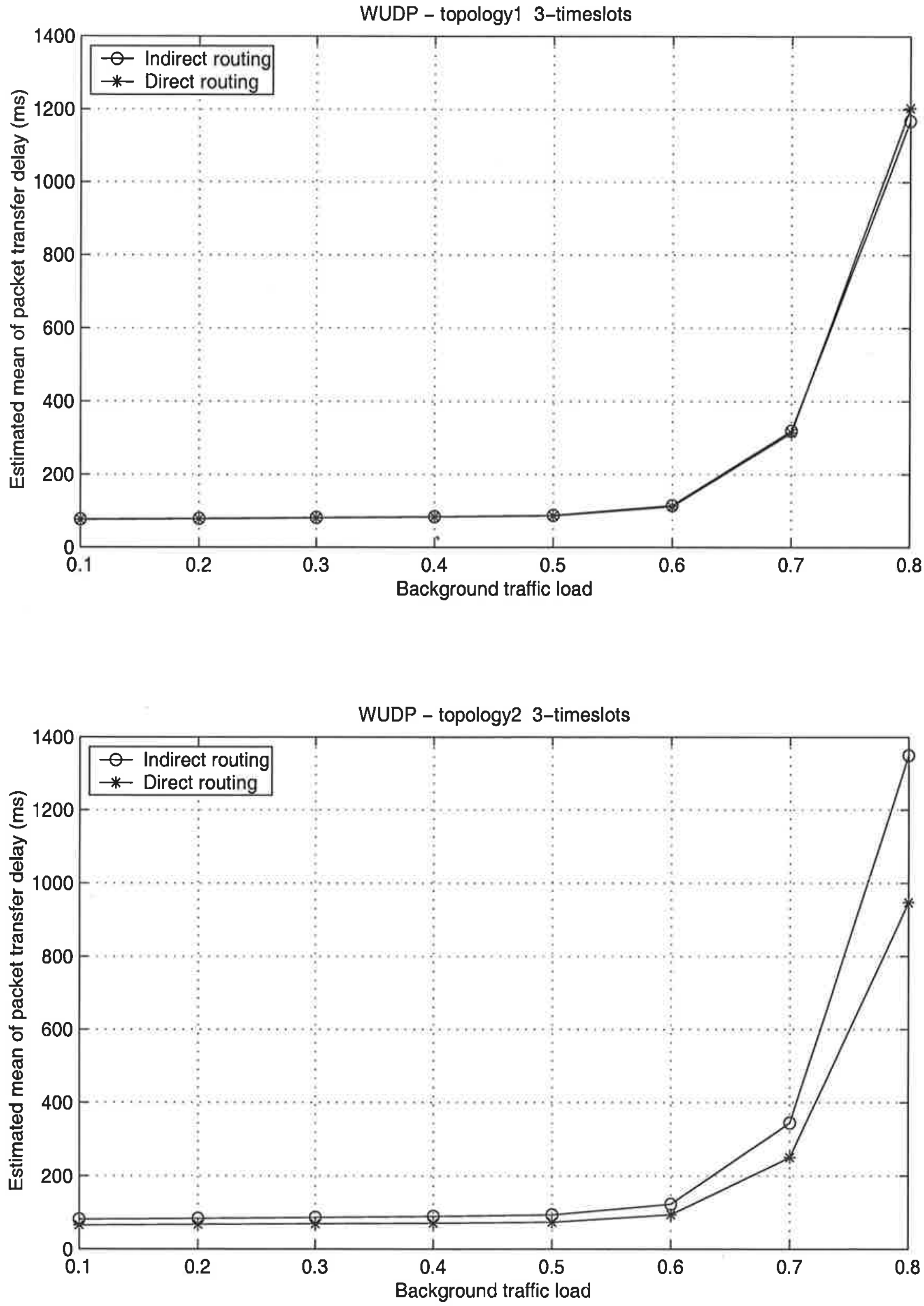
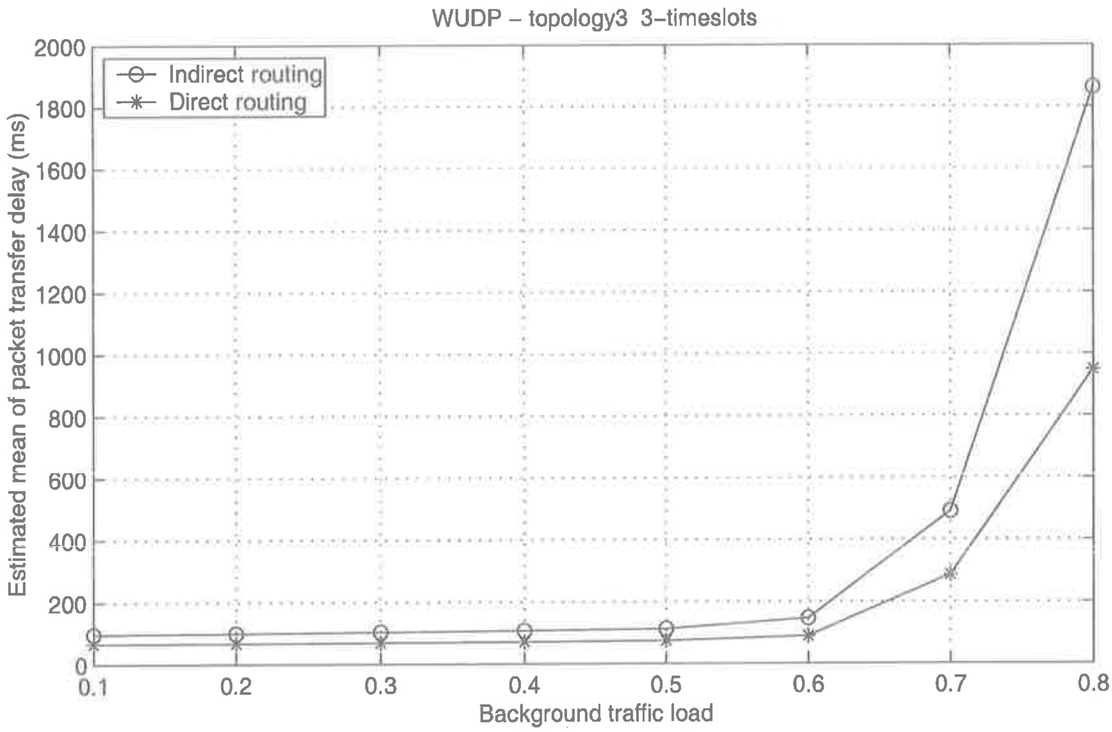


Figure 4.1.2-1: Packet transfer delay on Topology 1 and 2



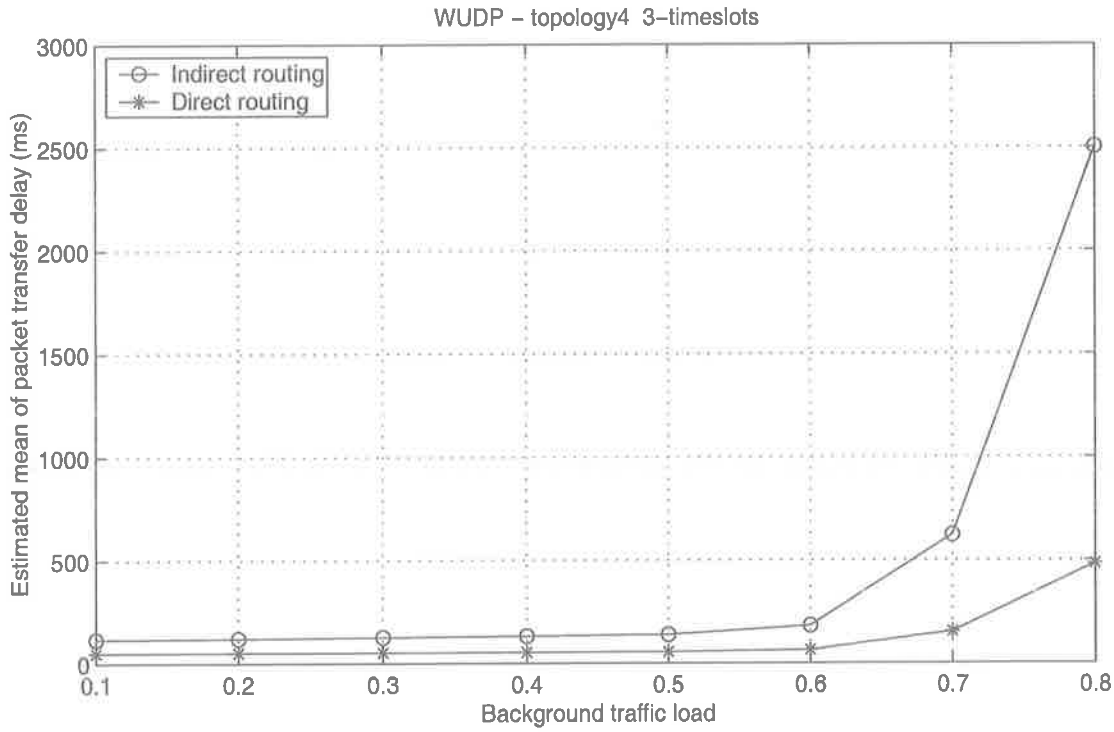


Figure 4.1.2-2: Packet transfer delay on Topology 3 and 4

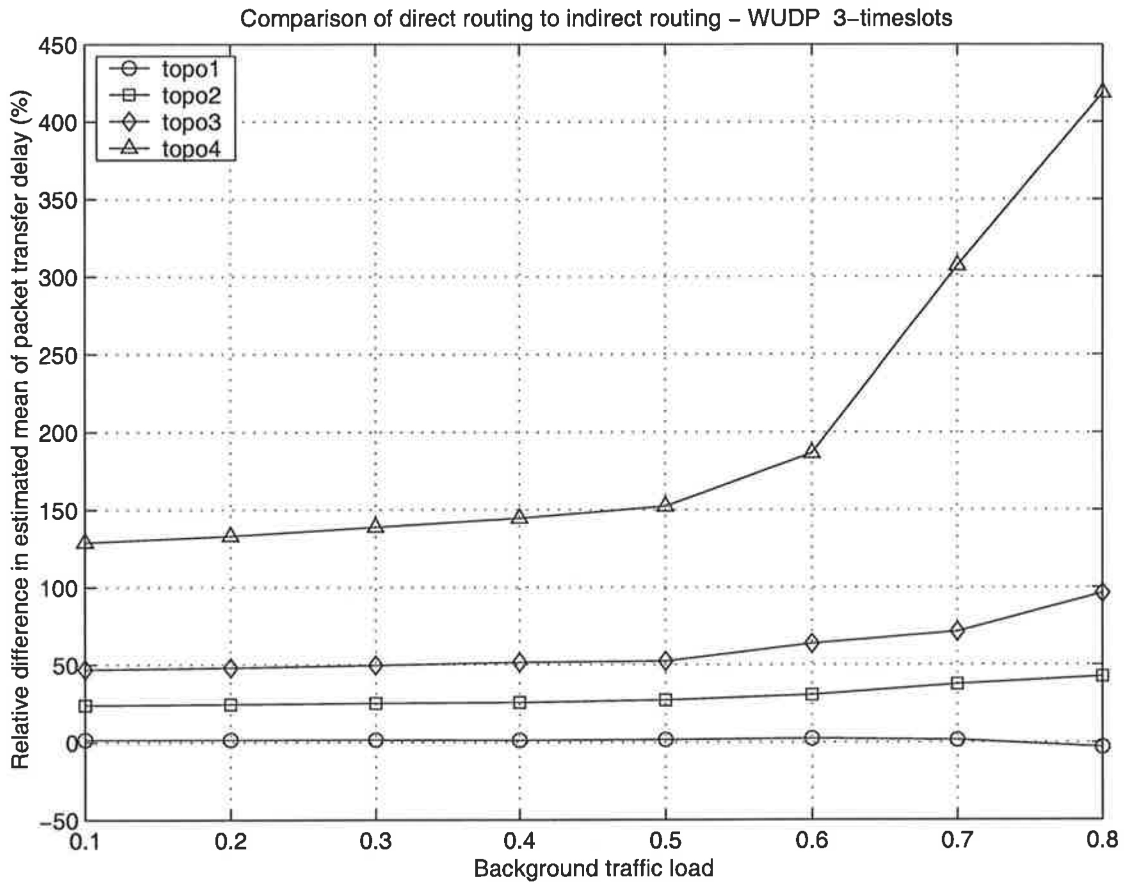


Figure 4.1.2-3: Relative difference between indirect and direct routing in packet transfer delay

Table 4.1.2-1 and 4.1.2-2 list the 95% confidence interval for the difference and normalised confidence interval for the relative difference between the two routings. The difference on topology 1 is trivial and we cannot reject the null hypothesis at a significance level of 2.5% under high background traffic load. We can reject the null hypothesis for the other three topologies.

Table 4.1.2-1: 95% confidence interval for difference between indirect and direct routing in packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean (ms)	Topo1		1.2	1.2	1.3	1.1	1.5	2.7	5.0	-35.6
	Topo2		15.7	16.5	17.5	18.4	20.1	28.8	94.1	402.9
	Topo3		30.8	32.3	34.2	36.6	38.7	57.0	204.7	913.2
	Topo4		65.7	68.9	73.1	77.7	83.8	118.8	469.7	2021.2
Std (ms)	Topo1		0.075	0.14	0.13	0.19	0.19	4.52	19.60	82.30
	Topo2		0.064	0.17	0.11	0.19	0.20	4.90	23.05	84.31
	Topo3		0.068	0.14	0.11	0.20	0.19	2.38	34.15	82.03
	Topo4		0.064	0.18	0.13	0.23	0.21	2.09	23.67	94.12
df	Topo1		56.0	55.3	55.4	48.1	46.9	52.8	57.8	52.6
	Topo2		57.8	57.8	47.2	57.4	41.1	45.7	56.4	55.9
	Topo3		50.1	57.7	54.4	42.4	51.5	52.4	54.9	51.1
	Topo4		41.4	41.0	41.3	44.0	42.4	57.9	52.4	42.2
MoE (ms)	Topo1		0.15	0.29	0.26	0.39	0.38	9.07	39.20	164.93
	Topo2		0.13	0.33	0.21	0.39	0.40	9.88	46.19	168.96
	Topo3		0.14	0.28	0.22	0.41	0.39	4.78	68.43	164.71
	Topo4		0.13	0.35	0.26	0.46	0.43	4.17	47.52	190.22
Confidence interval (ms)	Top1	Up	1.4	1.5	1.5	1.5	1.9	11.8	44.2	129.4
		Down	1.1	0.9	1.0	0.7	1.1	-6.3	-34.2	-200.5
	Top2	Up	15.8	16.8	17.7	18.8	20.5	38.7	140.3	571.9
		Down	15.6	16.2	17.3	18.0	19.7	19.0	47.9	233.9
	Top3	Up	31.0	32.6	34.4	37.0	39.1	61.8	273.1	1077.9
		Down	30.7	32.0	34.0	36.2	38.3	52.2	136.2	748.5
	Top4	Up	65.8	69.3	73.3	78.2	84.3	123.0	517.2	2211.4
		Down	65.6	68.6	72.8	77.3	83.4	114.6	422.2	1831.0

Table 4.1.2-2: Normalised confidence interval for relative difference between indirect and direct routing in packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Confidence interval (%)	Top1	Up	1.84	1.95	1.90	1.79	2.20	10.54	14.07	10.75
		Down	1.44	1.21	1.25	0.85	1.32	-5.65	-10.87	-16.66
	Top2	Up	24.05	24.98	25.60	26.35	27.80	41.12	56.00	60.40
		Down	23.66	24.00	24.98	25.26	26.73	20.13	19.11	24.71
	Top3	Up	47.13	48.37	49.90	52.04	52.76	68.96	95.27	113.6
		Down	46.71	47.54	49.25	50.89	51.71	58.28	47.52	78.89
	Top4	Up	129.3	133.8	139.4	145.5	153.0	193.1	338.2	458.2
		Down	128.8	132.5	138.4	143.8	151.5	180.0	276.0	379.4

B. fixed background traffic load

Given that there is an apparent change of packet transfer delay at 0.5 background traffic load, we run the simulations under 0.2 and 0.6 background traffic load respectively. Table A.2-4 (Appendix A) lists the estimated mean ($\overline{D_i}$) and standard deviation of packet transfer delay under indirect and direct routing where the background traffic load is 0.2. Table A.2-6 (Appendix A) lists the data under direct routing where the background traffic load is 0.6. Figure 4.1.2-4 and -5 plot the packet transfer delay as a function of number of GPRS timeslots under 0.2 background traffic load while Figure 4.1.2-7 and -8 show the relationship under 0.6 background traffic load. It is shown that the packet transfer delay decreases slowly from 3- to 8-timeslot operation. That can be explained by the following formulas. From equation (4.1-1), we have

$$E(Dd) = n \cdot E(Dh) + Dg(ts)$$
$$E(Di) = m \cdot E(Dh) + Dg(ts).$$

Here, *ts* denotes number of timeslots. *Dg(ts)* is in inverse proportion to ts while *E(Dh)* keeps constant under the same background traffic load. Using the above formulas, the relative difference θ_d can be derived from

$$\begin{aligned}\theta_d &= \frac{E(Di) - E(Dd)}{E(Dd)} \\ &= \frac{(m-n)\cdot E(Dh)}{n\cdot E(Dh) + Dg(ts)} \times 100\%.\end{aligned}$$

Considering $Dg(ts) \to 0$ and $\to \infty$, this becomes

$$\lim_{Dg(ts)\to 0} \theta_d(Dg(ts)) = \frac{m-n}{n} \times 100\%$$
$$\lim_{Dg(ts)\to \infty} \theta_d(Dg(ts)) = 0 \times 100\%.$$

Therefore,

$$\theta_d \in (0\times 100\%, \frac{m-n}{n} \times 100\%).$$

Given that there is a dramatic growth of *E(Dh)* from 0.2 to 0.6 background traffic load and the *ts* is within 3 - 8, the range of θ_d under 0.6 background traffic load (in case of topology 4, 186.1% - 283.8%) holds an up-offset over the range under 0.2 background traffic load (in case of topology 4, 133.6% - 237.8%). However, the growth trend of θ_d under 0.6 background traffic load is similar to that under 0.2 background traffic load. The simulation results shown in Table A.2-5, -7 (Appendix A) and Figure 4.1.2-6, -9 confirm the above deduction.

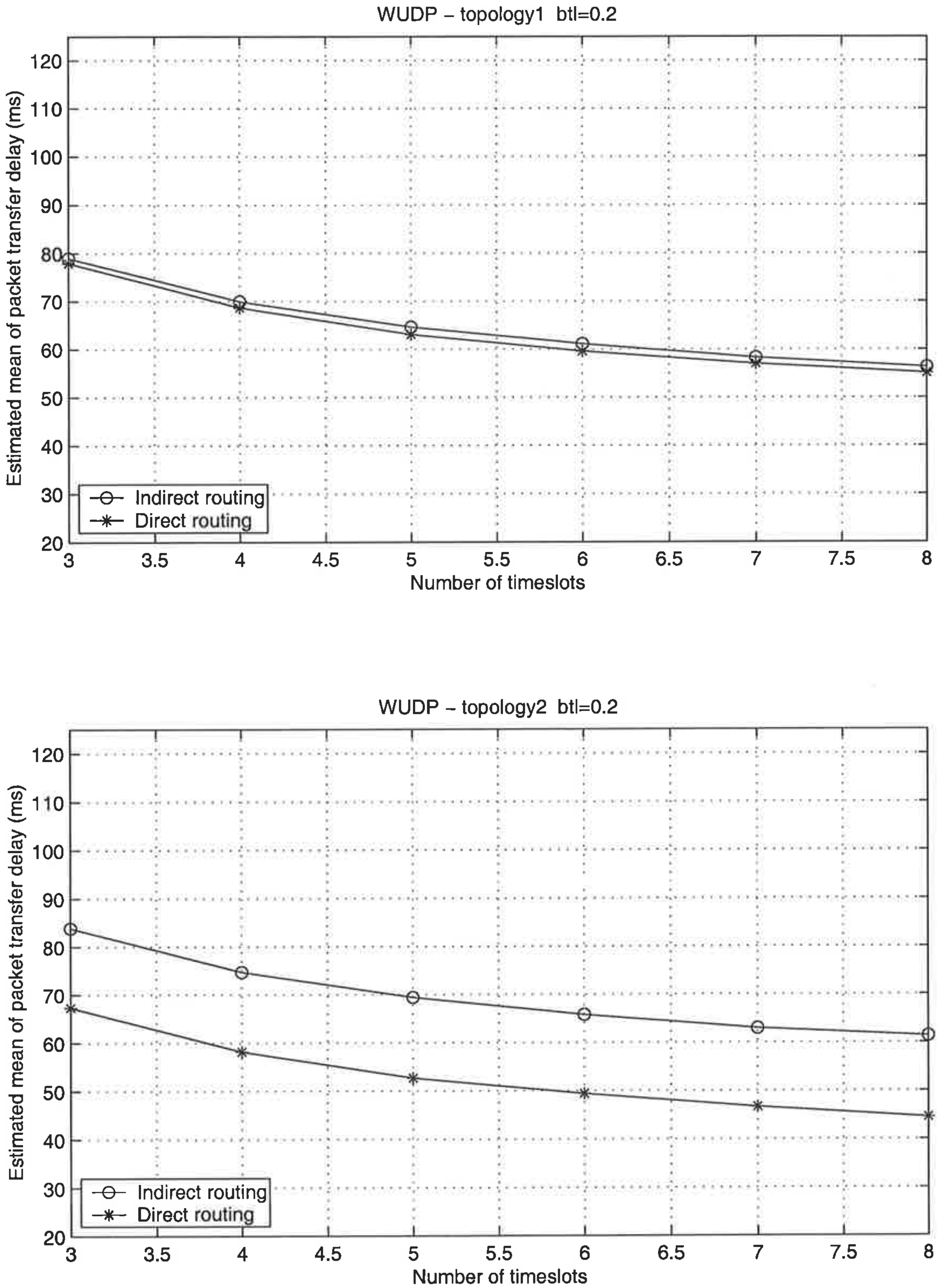
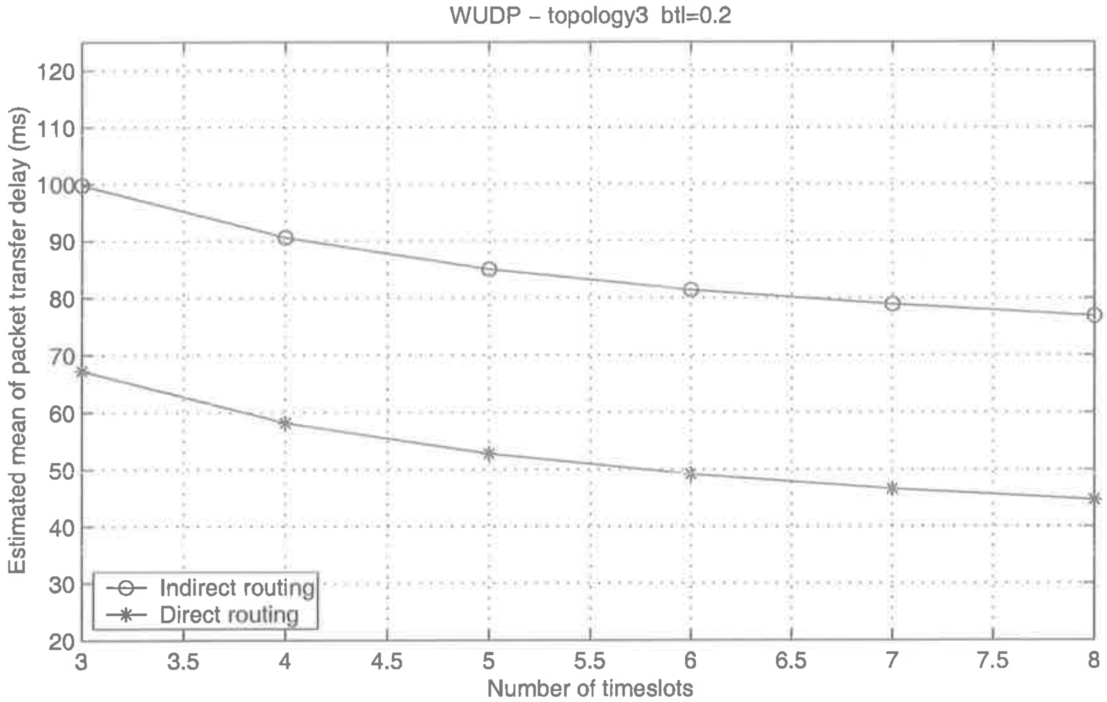


Figure 4.1.2-4: Packet transfer delay on Topology 1 and 2 (background traffic load = 0.2)



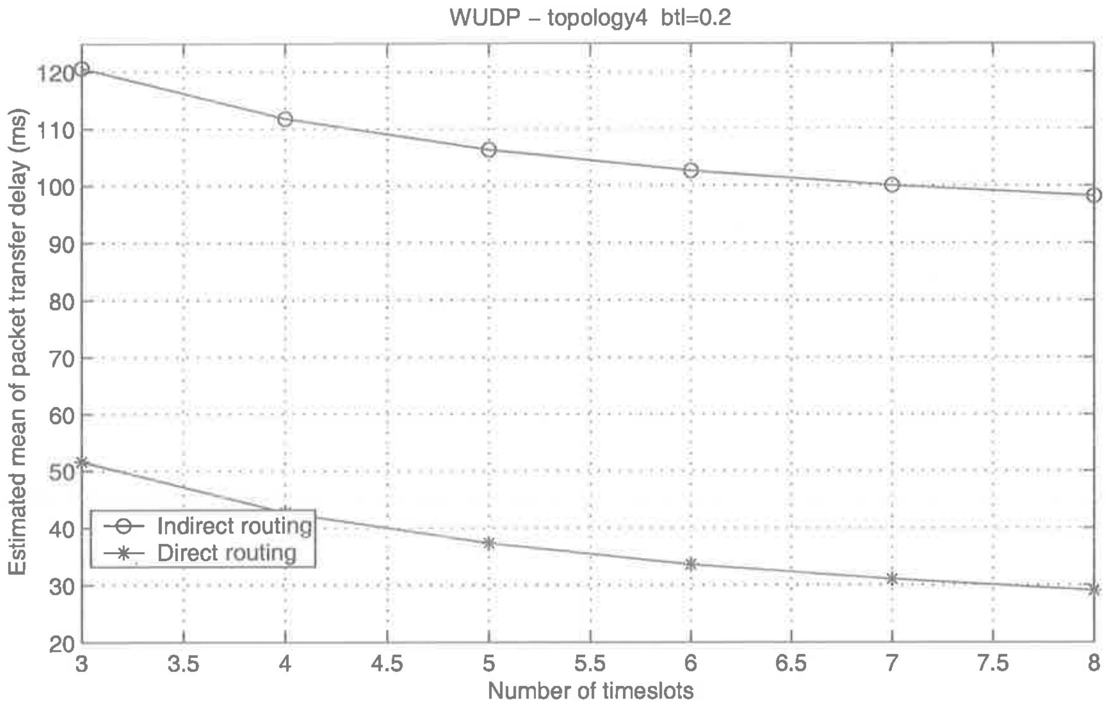


Figure 4.1.2-5: Packet transfer delay on Topology 3 and 4 (background traffic load = 0.2)

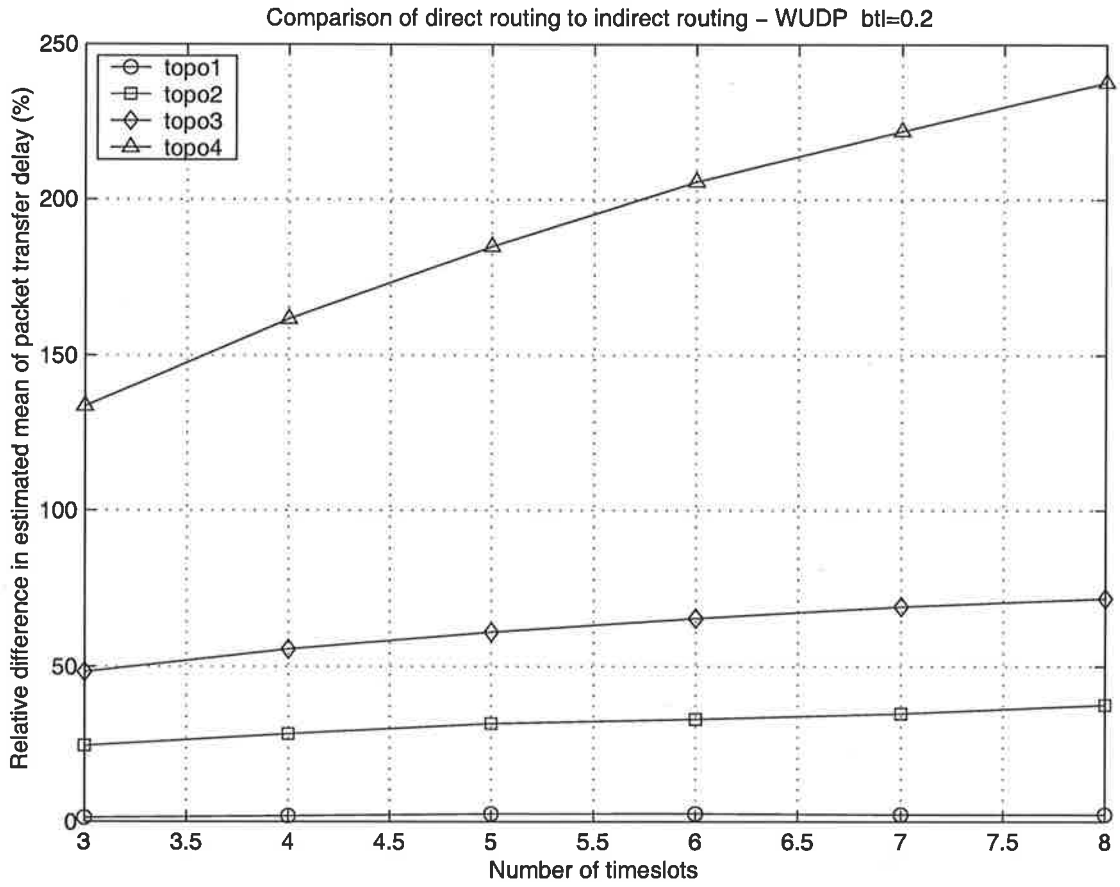
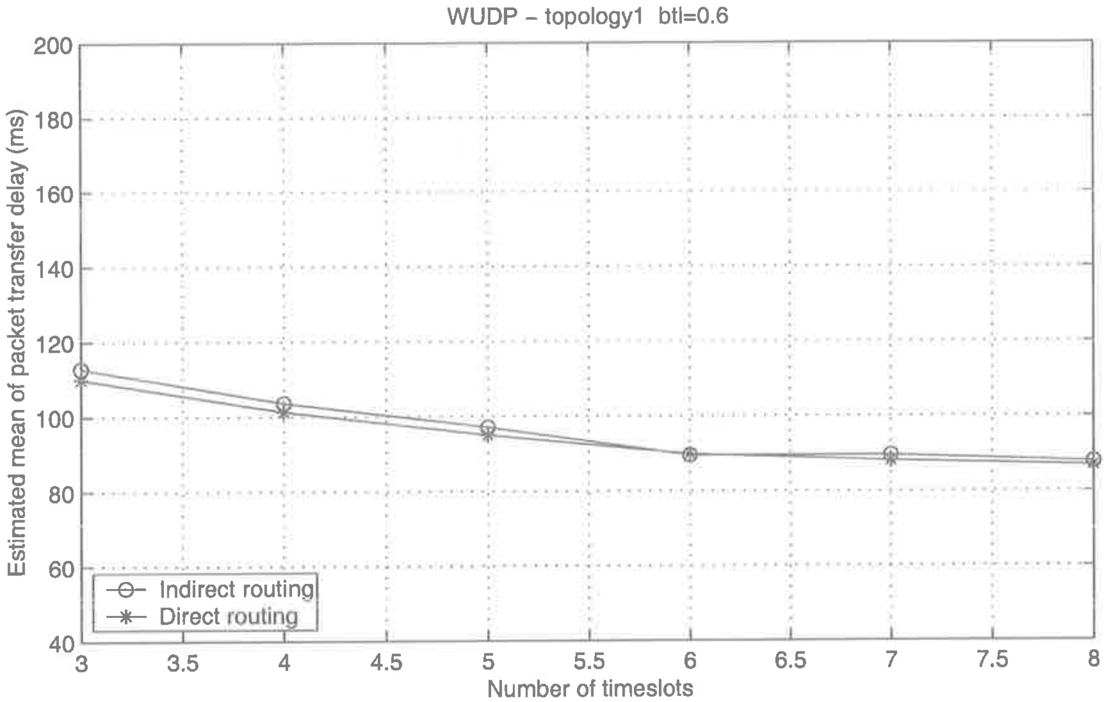


Figure 4.1.2-6: Relative difference between indirect and direct routing in packet transfer delay (background traffic load = 0.2)



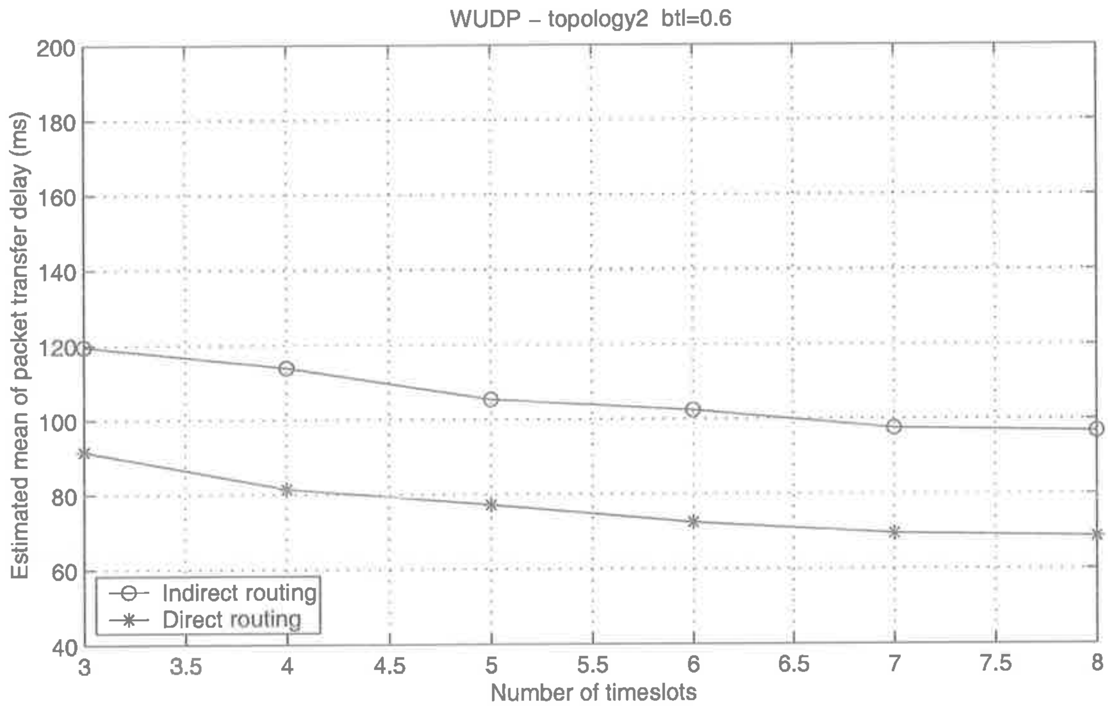
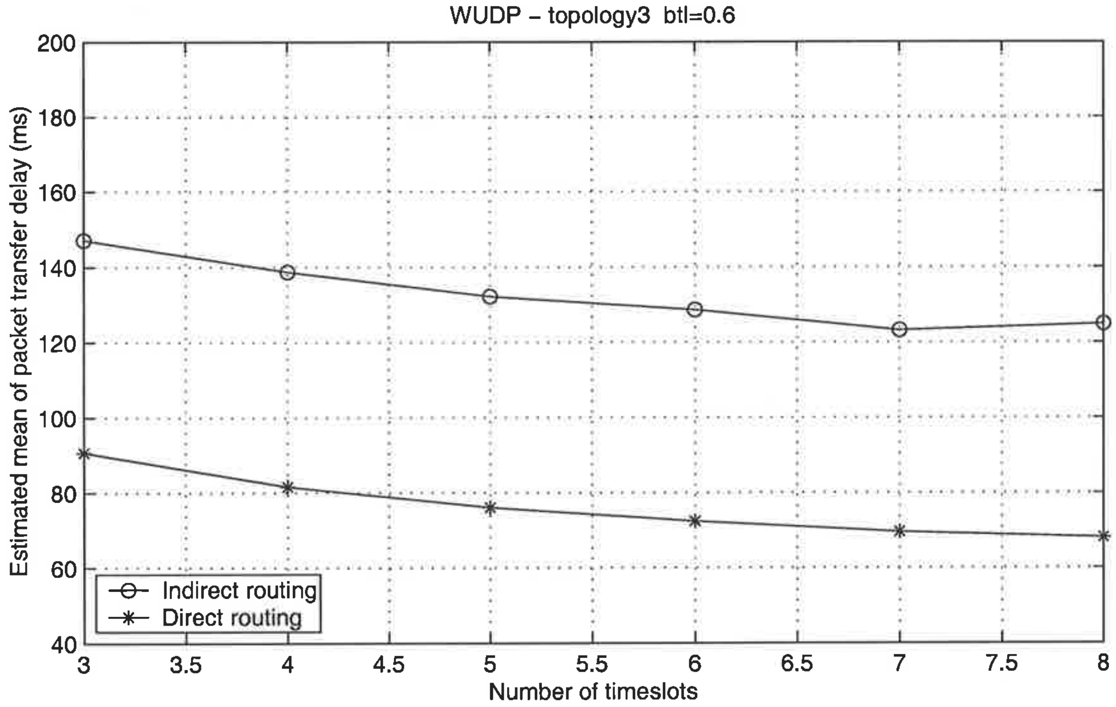


Figure 4.1.2-7: Packet transfer delay on Topology 1 and 2 (background traffic load = 0.6)



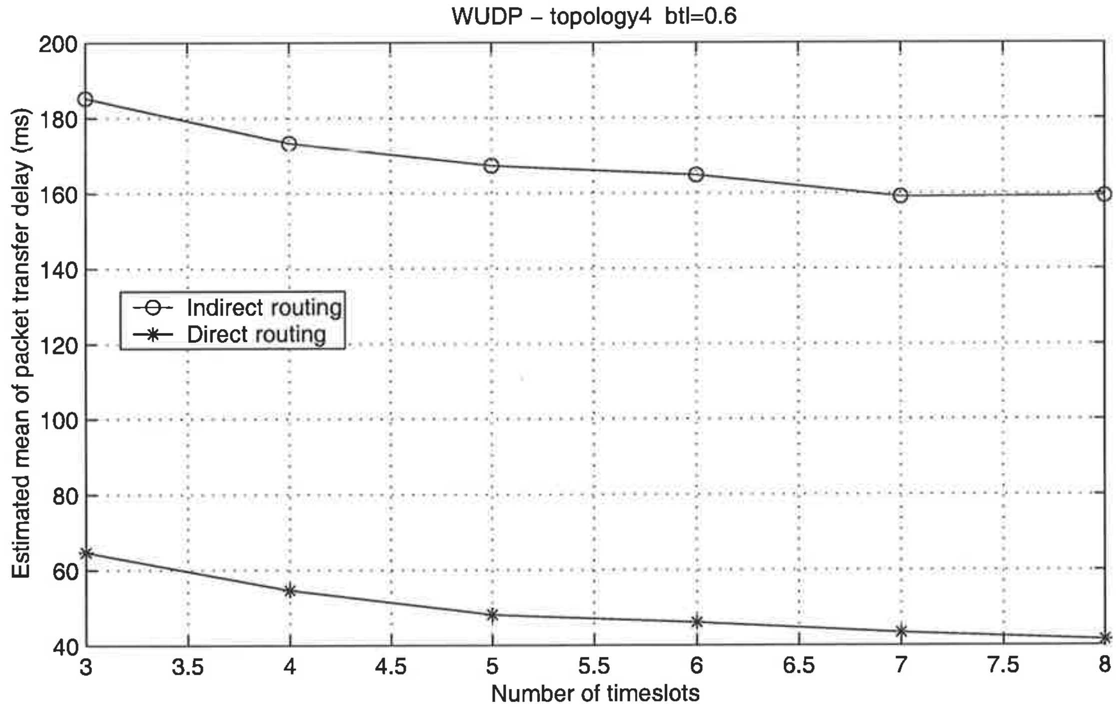


Figure 4.1.2-8: Packet transfer delay on Topology 3 and 4 (background traffic load = 0.6)

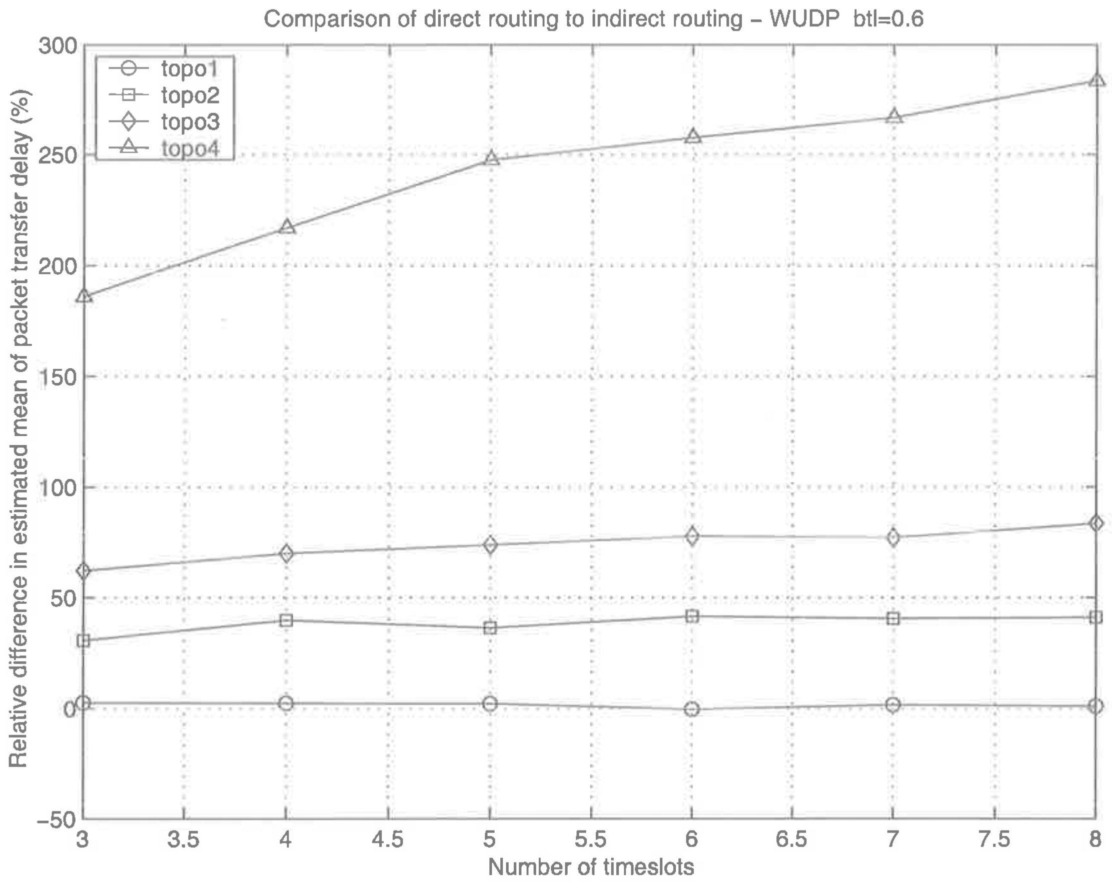


Figure 4.1.2-9: Relative difference between indirect and direct routing in packet transfer delay (background traffic load = 0.6)

Standard deviation of packet transfer delay

A. Fixed GPRS radio link capacity

Theoretically, there is no difference between the performance for broadband and narrowband mobile user in this measure, because we just introduce a GPRS radio link with fixed transfer delay. Therefore, the relative difference θ_s can be derived from equation (4.1-2)

$$\theta_s = (\sqrt{\frac{m}{n}} - 1) \times 100\%.$$

Please refer to the description of standard deviation of packet transfer delay in the previous section.

B. Fixed background traffic load

When a certain number of timeslots are assigned for data transmission on GPRS radio link, the radio link just introduces a constant transfer delay. Therefore, it doesn't affect standard deviation of packet transfer delay. That is, under a certain background traffic load the estimated mean of standard deviation of packet transfer delay basically keeps constant for each topology as the available timeslots increase. Since GPRS radio link doesn't contribute to the standard deviation, the relative difference θ_s between the indirect and direct routing can still use the following formula equation (4.1-2):

$$\theta_s = (\sqrt{\frac{m}{n}} - 1) \times 100\%.$$

Please refer to the description of standard deviation of packet transfer delay in the previous section.

4.2 TCP Performance

In this section we focus on TCP performance and take Packet Transfer Delay, Goodput and Retransmission Rate as performance measures. They are defined as follows:

- Throughput: the number of packets delivered successfully per unit of time (100 seconds, unless otherwise specified) from source to destination node.
- Goodput (number of packet transferred): the number of packets delivered successfully *excluding duplicate packets* per unit of time (100 seconds, unless otherwise specified) from source to destination node.
- Retransmission: the transmission that is triggered by TCP congestion control, either by the reception at TCP sender of "triple-duplicate" acknowledgements or timeouts.
- Retransmission rate: dividing retransmission times by total transmission times (whether the packet is received or not).
- Packet transfer delay: the time between when a packet is put into queue at source node and when the packet is completely received at destination node. In our simulations it consists of the total of queuing delay, transmission delay and propagation delay on each link. Queuing delay is between the time the packet is assigned to a queue for transmission and the time it starts being transmitted. Transmission delay is between the time the first and last bits of the packet are transmitted. Propagation delay is between the time the last bit of the packet is transmitted at the head node of the link and the time the bit is received at the tail node of the link.

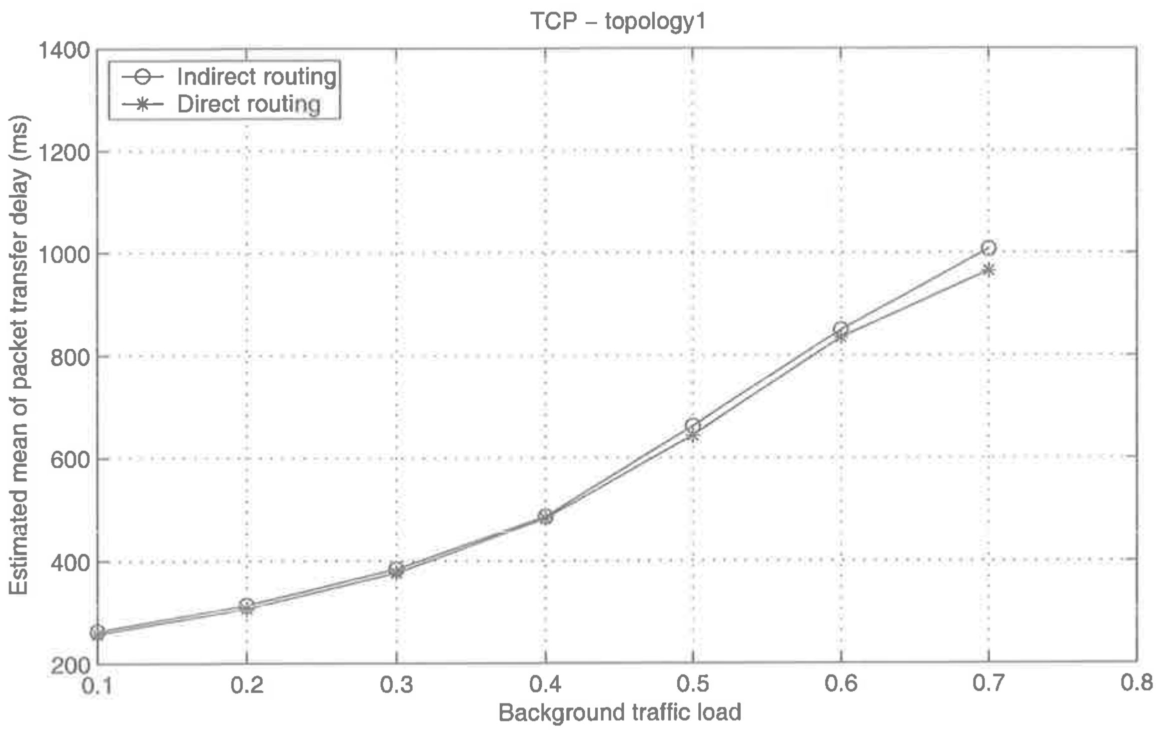
We run simulations on the four topologies described in Chapter 3 under indirect and direct routing respectively.

4.2.1 Performance for Broadband Mobile User

The simulation is performed under various background traffic loads from 0.1 to 0.7 with an interval of 0.1. We make 30 simulation runs for each background traffic load. From a single simulation run we can collect the transfer delay for each packet. Then a mean for packet transfer delay is available for each simulation run. A value of Goodput is obtained for each simulation run. Under high background traffic load, retransmission rate is measured to examine performance. The TCP connection time is 100s. If let i denote background traffic load and j be the jth simulation under the background traffic load i, then we have:

$$\overline{D_i} = \frac{1}{n}\sum_{j=1}^{n} D_{ij} \text{ , } \overline{G_i} = \frac{1}{n}\sum_{j=1}^{n} G_{ij} \text{ and } \overline{R_i} = \frac{1}{n}\sum_{j=1}^{n} R_{ij}$$

where D_{ij}, G_{ij} and R_{ij} are respectively the sample mean of packet transfer delay, goodput and retransmission rate from the jth simulation for background traffic load i. Here, $n = 30$. Therefore $\overline{D_i}$, $\overline{G_i}$ and $\overline{R_i}$ respectively denote the pooled sample mean of packet transfer delay, sample mean of goodput and sample mean of retransmission rate for background traffic load i. **In convenience of description, the pooled sample mean ($\overline{D_i}$) of packet transfer delay, sample mean ($\overline{G_i}$) of goodput (number of packets transferred) and sample mean ($\overline{R_i}$) of retransmission rate are called henceforth as estimated mean of packet transfer delay, estimated mean of goodput (number of packets transferred) and estimated mean of retransmission rate.** $\overline{D_i}$, $\overline{G_i}$ and $\overline{R_i}$ are respectively discussed in the following three sections.



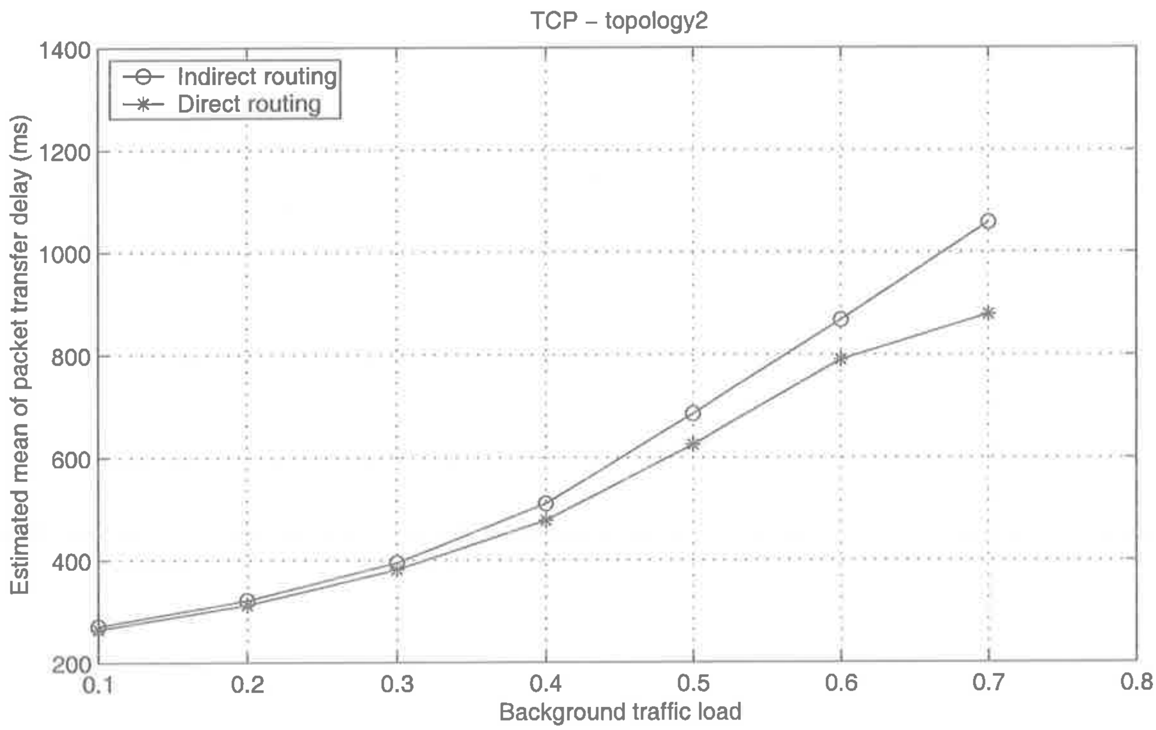
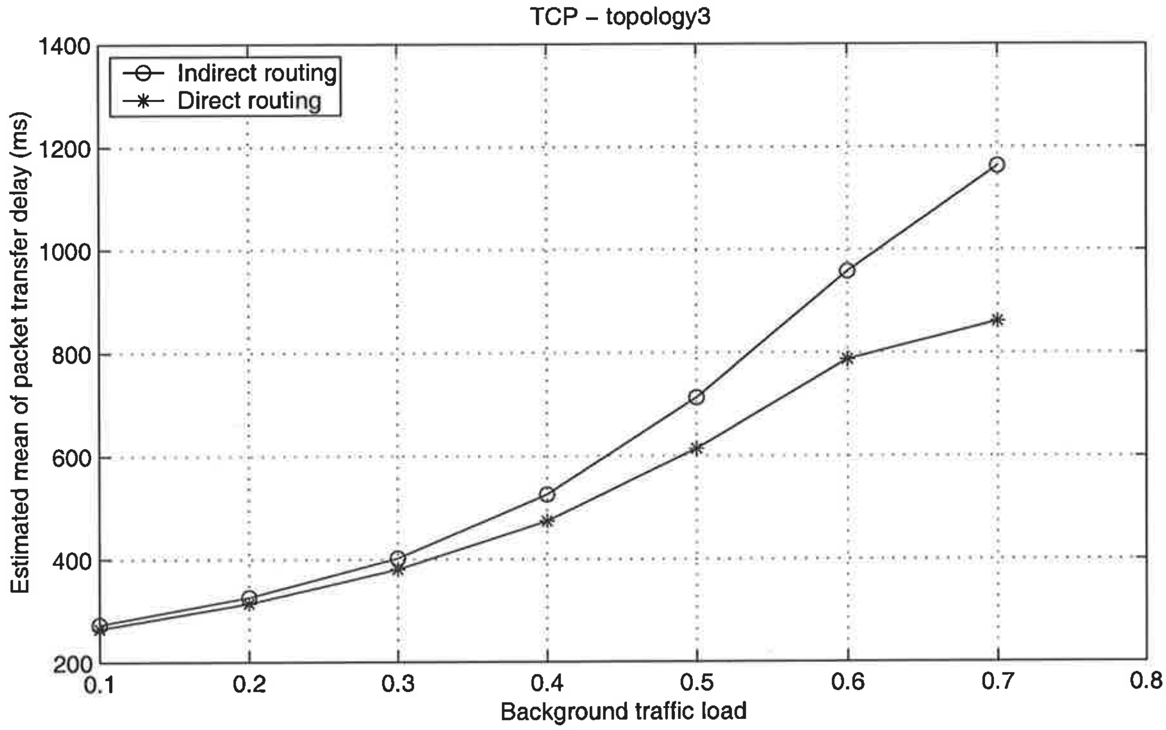


Figure 4.2.1-1: Packet transfer delay on Topology 1 and 2



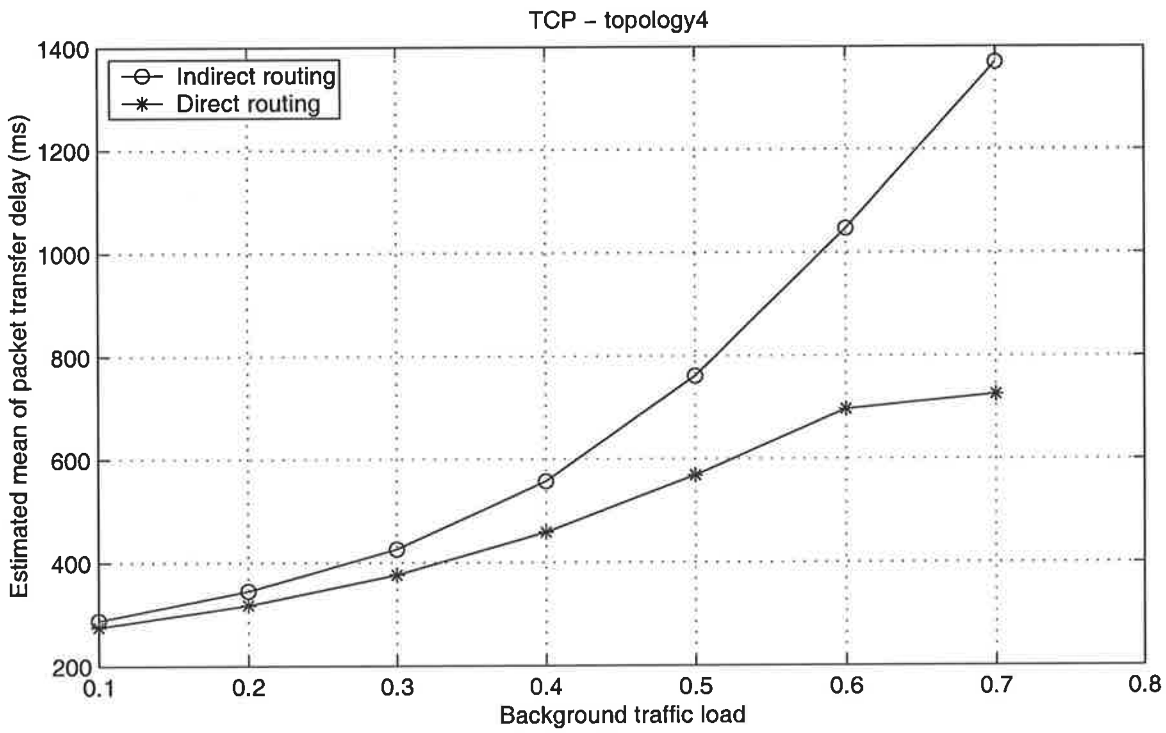


Figure 4.2.1-2: Packet transfer delay on Topology 3 and 4

Packet transfer delay

Sample means (D_{ij}) are derived from 30 simulation runs. A pooled sample mean ($\overline{D_i}$) and a sample standard deviation can be figured out from these 30 sample means (D_{ij}). We take these 30 sample means with the same significant level, although the number of elements for each of them is different. For example, we get 22,000 values of packet transfer delay from the first simulation run and 20,000 values from the second simulation run. Two sample means of packet transfer delay can be obtained from these two simulation runs, say, 0.2354s and 0.2503s. We take these two sample means with the same significant level since each simulation is independently run with the same given simulation time. That is, each simulation run holds the same significance level.

Table B.1-1 and B.1-2 (Appendix B) list the estimated mean ($\overline{D_i}$) and standard deviation of packet transfer delay for the studied topologies under indirect and direct routing. Figure 4.2.1-1 and 4.2.1-2 plot packet transfer delay as a function of background traffic load under indirect and direct routing respectively. It is shown that packet transfer delay increases exponentially with the increasing of background traffic load and we also find that under direct routing, from a background load of 0.6 and upwards the growth of packet transfer delay deviates significantly from exponential (see page 92 for explanation). It accords with common sense. However under low background traffic loads, the difference between indirect and direct routing is not apparent. With the increasing of background traffic load, the difference is becoming larger and larger. Figure 4.2.1-1, -2 and Table 4.2.1-1 show this point clearly.

On the studied topologies, if no studied TCP traffic is introduced, the packet transfer delay on the routing path is the total of queuing delays created by background traffic, propagation delays and transmission delays on all hops. The packet transfer delay depends on the length of routing path. Therefore, packet transfer delay in the case of no TCP traffic introduced is of significant relative difference between indirect and direct routing (except for topology 1) as shown in the UDP performance. However, when the studied TCP traffic is introduced we don't see such a significant difference between the two routings under low background traffic load. There must be a relatively long delay between source node and destination node under whether indirect or direct routing so that the difference becomes not significant. This long delay is produced by TCP traffic itself. Because of TCP congestion avoidance mechanism, packets at TCP sender are transmitted in the amount of a congestion window (cwnd) of packets. Once all packets falling within the window have been sent in the back to back transmission, no other packets are transmitted until the ACK for the first transmitted packet of these packets is received if no timeout or duplicate ACK loss indication happens. This procedure begins with the transmission of the first packet in the window and ends with the reception of its ACK. We refer to the procedure as a "round" [31]. The reception of this ACK marks the end of this round and the beginning of the next round. Apparently, the duration of a round is equal to the round-trip-time (RTT). The round duration is longer or equal to the time needed to send all the packets in the congestion window. When it is longer the round duration can be divided into two parts: burst and idle duration. The burst duration is the time needed to send all the packets in the congestion window. In burst duration, TCP sender is engaged in sending packets. The idle duration is between the last bit in the congestion window is transmitted and the ACK of the first packet in this round is received. In idle duration, TCP sender is waiting for the ACK and no other packets are transmitted. We refer to a round duration as T, burst part as T_{burst} and idle part as T_{idle}. Here is

$$T = T_{burst} + T_{idle}.$$

In the stage of TCP warm-up before the maximum allowed window size is reached, which is defined by the minimum window size of congestion window imposed by TCP sender and advertised window imposed by TCP receiver, the duration of T_{burst} increases while T_{idle} decreases from round to round. When the maximum window size is reached the duration of T_{burst} and T_{idle} is relatively stable. If there is no idle duration, that is, T_{idle} equals to zero, then T is equal to T_{burst}. The TCP sender is always engaged in sending packets and the TCP receiver is engaged in receiving packets, with a relatively stable interval between two continuous packets.

First of all, we analyze traffic performance without background traffic introduced. After TCP warm-up stage, if $T > T_{burst}$, there is no queue induced by the TCP traffic between the sender and receiver. In this case, the packet transfer delay depends on the delays created on the intermediate links between the TCP sender and receiver. That is, the relative difference between indirect and direct routing should be significant if the difference between the indirect and direct routing is significant in path length. On the other hand, if there is no idle duration, that is, T is equal to T_{burst}, the link with minimum capacity is fully used. There are two situations: one is no queue on the link (or in other words, the queue size is equal to zero), the other is that there is a constant queue on the link. In the former case, there is no queue between the TCP sender and receiver as $T > T_{burst}$. In the later case, the queue on that link would contribute to the packet transfer delay. How much the contribution is depends on the length of the queue, which depends on the maximum allowed window size for TCP congestion window for a given topology scenario. Increasing the window size increases the queue size, otherwise reduces the queue size. Since the maximum allowed window size is reached, this queue keeps constant until the end of transmission. In contrast to the $T > T_{burst}$ scenario, an additional queuing delay is imposed. In this case, when the maximum window size is reached queuing delay in the nearest bottleneck to TCP sender plays an important role in packet transfer delay.

Returning to the studied topologies, we still consider the situation without background traffic introduced. The TCP sender and receiver are integrated into Correspondent Gateway and Visited Gateway respectively. And all links have the same link capacity, 2Mbits/s. Therefore, both the TCP sender and receiver have the same 2Mbits/s capacity as other links. During TCP warm-up stage, congestion window is incremented by one packet every time an ACK is received. That is, the reception of one ACK asks two packets be put in congestion window for transmission. Since the capacity at the TCP sender is the same but not twice as the minimum capacity within the routing path, queue is inevitably created at the TCP sender in each round during TCP warm-up. When all the ACKs for the precious round of packets are received, the queue doesn't grow further. And since then, because packets stored in the queue are still released one by one the queue size decreases until the next round of ACKs are received. It is possible that the queue is empty or not at that time. When the maximum window size is reached, the queue is stable. If $T_{idle} > 0$ there is no queue. If $T_{idle} = 0$ there is a constant queue (the queue size may be equal to zero). If the maximum allowed window size is the same for both indirect and direct routing, the difference in T_{idle} between the indirect and direct routing decides the difference in packet transfer delay. The larger the difference in T_{idle} is, the larger the difference in packet transfer delay is. Under the situation without encapsulation overhead, if no T_{idle} exists for the two routings there is no difference in packet transfer delay except TCP warm-up stage. That is, in this case encapsulation overhead is the only difference contributor. Since the encapsulation overhead is introduced at home network gateway, the outgoing link to the visited gateway has to transmit the encapsulation overhead. Consequently, the bandwidth available to transmit the original FTP packets is less than 2Mbits/s. The link is the nearest bottleneck to the TCP sender. Therefore, when TCP warm-up finishes there are two queues in the routing path: one is at the TCP sender (the queue may be equal to zero), the other is on the outgoing link. As time increases, the queue at the TCP sender diminishes and vanishes while the queue on the outgoing link accumulates.

Under low background traffic load, the queuing delay created by background traffic is trivial while the TCP traffic is the major contributor of the total traffic. Therefore, we can't see the apparent difference caused by background traffic. However, the situation becomes different when high background traffic load is introduced. In this case, background traffic takes a more important role. As we see in Figure 4.2.1-1 and 4.2.1-2, the difference becomes larger and larger with the increasing of background traffic load. The relative difference shown in Table B.1-3 (Appendix B) and Figure 4.2.1-3 goes in the same way. The relative difference is derived from the following formula:

$$\theta_d = \frac{d_{ind} - d_d}{d_d} \times 100\%$$

Here, θ_d denotes relative difference between indirect and direct routing in packet transfer delay. And d_{ind} and d_d denote the estimated mean of packet transfer delay under indirect and direct routing respectively.

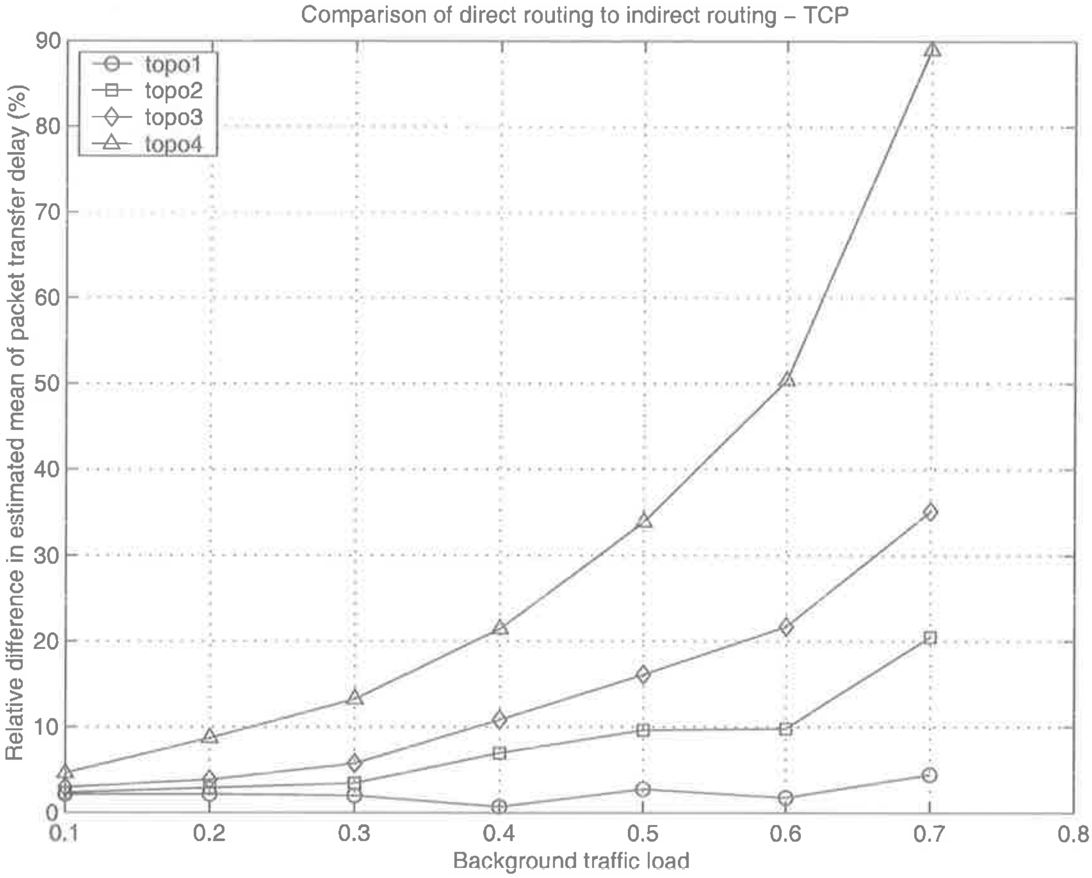


Figure 4.2.1-3: Relative difference between indirect and direct routing in packet transfer delay

Figure 4.2.1-1 and 4.2.1-2 show that the increment of packet transfer delay has a structural change under direct routing from 0.6 to 0.7 background traffic load: the growth of packet transfer delay is not so clear as expected. However, no such a change exists or the change is not clear under indirect routing. Let's take a look at retransmission rate under 0.6 and 0.7 background traffic load in Table 4.2.1-5 and -6. From 0.6 to 0.7 background traffic load there is relatively dramatic increment in retransmission rate. Retransmission means that the TCP sender judges that congestion happens due to timeout or duplicate ACK loss indication. Before the TCP sender makes the judgement, it waits for a number of duplicate ACKs to see whether a duplicate ACK is caused by a lost packet or just a reordering of packets. If three or more duplicate ACKs are received, the TCP sender considers that the packet is lost, then fast retransmit and fast recovery processes are invoked. During this period, the TCP throughput decreases. Sequentially, the total traffic load decreases on the routing path and thus packet transfer delay decreases too. On the other hand, the TCP throughput is higher under direct routing than under indirect routing (The discussion about throughput is presented in the next section). Lower throughput under indirect routing indicates that its weight within the total traffic is lighter than TCP traffic's under direct routing. That is why retransmission has more effect on packet transfer delay under direct routing than indirect routing, even though there is a higher retransmission rate under indirect routing.

In order to prove that the chance explanation is implausible for the difference between the packet transfer delays under indirect and direct routing, we introduce T test described in Chapter 3 for hypothesis testing. Please refer the detailed computing steps in Chapter 3. A 95% confidence interval is chosen to define *t* score in *t* table. Table 4.2.1-1 lists estimated mean and standard deviation of the difference, estimated number of degrees of freedom, margin of error and confidence interval.

Table 4.2.1-1: 95% confidence interval for difference between indirect and direct routing in packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	5.59	6.62	7.53	3.42	17.61	14.58	42.77
	Topo2	6.19	9.05	13.11	33.15	60.03	77.38	179.77
	Topo3	7.87	12.09	21.80	51.33	98.82	170.69	302.21
	Topo4	12.70	27.65	49.75	98.29	192.64	350.42	644.97
Std (ms)	Topo1	0.43	1.06	2.47	3.92	10.68	23.57	34.93
	Topo2	0.45	1.08	2.10	5.23	9.87	17.21	37.59
	Topo3	0.43	0.95	3.09	5.48	100.29	19.08	30.59
	Topo4	0.40	1.30	2.59	4.91	9.63	16.54	31.61
df	Topo1	56.78	54.35	57.89	57.58	57.44	57.78	57.21
	Topo2	57.74	56.72	56.22	56.03	52.70	56.08	57.65
	Topo3	54.27	48.33	53.57	52.86	57.37	56.18	55.87
	Topo4	45.25	50.10	52.10	50.70	53.89	52.61	38.16

(Continued)

Table 4.2.1-1 (continued)

MoE (ms)	Topo1		0.86	2.12	4.94	7.84	21.40	47.15	69.99
	Topo2		0.91	2.17	4.22	10.48	19.78	34.48	75.19
	Topo3		0.87	1.91	6.20	10.98	20.61	38.24	61.30
	Topo4		0.81	2.61	5.20	9.86	19.30	33.15	63.88
Confidence interval (ms)	Top1	Up	6.45	8.74	12.47	11.26	39.02	61.73	112.76
		Down	4.73	4.50	2.60	-4.43	-3.79	-32.56	-27.22
	Top2	Up	7.10	11.22	17.32	43.63	79.81	111.86	254.94
		Down	5.28	6.89	8.89	22.67	40.24	42.90	104.58
	Top3	Up	8.74	14.01	28.00	62.31	119.43	208.93	363.51
		Down	7.01	10.18	15.60	40.35	78.21	132.44	240.91
	Top4	Up	13.50	30.26	54.95	108.15	211.93	383.58	708.84
		Down	11.89	25.04	44.54	88.43	173.34	317.27	581.09

Table 4.2.1-2: Normalised confidence interval for relative difference between indirect and direct routing in packet transfer delay

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Confidence interval (%)	Top1	Up	2.50	2.85	3.31	2.34	6.06	7.40	11.69
		Down	1.84	1.47	0.69	-0.92	-0.59	-3.90	-2.82
	Top2	Up	2.68	3.60	4.54	9.15	12.77	14.16	29.03
		Down	1.99	2.21	2.33	4.75	6.44	5.43	11.91
	Top3	Up	3.30	4.46	7.36	13.14	19.45	26.55	42.26
		Down	2.65	3.24	4.10	8.51	12.74	16.83	28.01
	Top4	Up	4.91	9.53	14.61	23.58	37.29	55.10	97.80
		Down	4.32	7.89	11.84	19.28	30.50	45.58	80.18

From Table B.1-3 and Figure 4.2.1-3, θ_d keeps low through all background loads on topology 1. That indicates that encapsulation overhead doesn't make significant effect on packet transfer delay. Furthermore, according to Table 4.2.1-1 and -2 we fail to reject the null hypothesis at a significance level of 2.5% from 0.4 through 0.7 background traffic load while we can reject the null hypothesis at the significant level from 0.1 through 0.3 background traffic load. That is, with the increasing of background traffic load the effect from encapsulation overhead is becoming more and more trivial. Through topology 2 to 4, the relative difference is larger and larger with the increasing of background traffic load, and we can reject the null hypothesis at a significance level of 2.5%. In the view of topology scenario, there is an approximately logarithmic growth of the relative difference through topology 1 to 4.

Number of packets (Goodput)

We obtain 30 values (G_{ij}) of goodput from 30 simulation runs for each background traffic load. Table B.1-4 and B.1-5 (Appendix B) respectively list the estimated mean ($\overline{G_i}$) and standard deviation of goodput under indirect and direct routing through 0.1 to 0.7 background traffic load. Figure 4.2.1-4 and 4.2.1-5 plot goodput as a function of background traffic load under indirect and direct routing respectively and show that there is a linear decrease of goodput with the increasing of background traffic load. However except for topology 1, with the increasing of background traffic load there is an approximate increase of the difference between these two routings in goodput on the rest three topologies. That is also shown in Table 4.2.1-3 (Mean). These two points make that the relative difference increases exponentially with the increasing of background traffic load. That is shown on Figure 4.2.1-6. The relative difference is derived from the following formula:

$$\theta_g = \frac{g_d - g_{ind}}{g_{ind}} \times 100\%$$

Here, θ_g denotes the relative difference. g_{ind} and g_d denote the estimated mean of goodput under indirect and direct routing respectively. The figure shows that under low background traffic load goodput under direct routing has no significant improvement over indirect routing while the improvement is becoming apparent with the increasing of background traffic load. To explain this phenomenon, please recall the concept of "round" mentioned in "packet transfer delay" in this section. When RTT is larger than the time needed to send all the packets in the congestion window, the round duration T can be divided into two parts: burst part T_{burst} plus idle part T_{idle}. The burst part is the duration to send packets in the congestion window and equals to the time needed to send all the packets in the congestion window. The idle part is the duration between the last bit in the congestion window is transmitted and the ACK of the first packet in this round is received.

In the stage of TCP warm-up before the maximum allowed window size is reached, the duration of T_{burst} increases while T_{idle} decreases from round to round. When the maximum window size is reached the duration of T_{burst} and T_{idle} is relatively stable. When no background traffic is introduced and both indirect and direct routing have the same maximum allowed window size, the duration of T_{burst} is the same for the two routings. Accordingly, the T_{idle} value would define the difference between the two routings in throughput. Since indirect routing is of a longer routing path, it has a larger T_{idle} value. Therefore, it is reasonable to expect that the indirect routing would degrade throughput in some extent. The extent of the degradation depends on the relative difference between the two routings in the T value. That further depends on the T_{idle}. Because of the existence of the T_{idle}, the T value is proportional to the length of the routing path. Therefore, the relative difference between the two routings in the T value depends on the relative difference between them in the length of routing path. When background traffic is introduced, if the T_{idle} exists the performance difference between the two routings should reflect the difference between them in the length of routing path. However, the simulation results (Figure 4.2.1-6) show that the relative difference between the two routings in goodput is not clear under low background traffic load. Some readers may note that the same situation happens to packet transfer delay. Packet transfer delay takes an important role in RTT and is approximately proportional to RTT

- thus approximately proportional to T. That is, the T values for indirect and direct routing are similar with each other. We can accordingly deduce that is why there is no clear difference between indirect and direct routing in goodput. However, before drawing this judgement we should consider if the two routings hold the same congestion window size. For instance, if under the same conditions except that we just cut the maximum allowed window size to half under indirect routing, the packet transfer delay would be apparently smaller under indirect routing since the queuing delay in the nearest bottleneck to the TCP sender is dramatically reduced under indirect routing. Therefore, T value is much smaller under indirect routing. However, that cannot ensure a higher throughput under indirect routing. Adversely, that may aggravate the performance of indirect routing because smaller window size may make insufficient usage of link capacity. That is why in many cases above we assume that both these routings have the same maximum allowed window size.

We note that the above expectation of direct routing' improvement over indirect routing in traffic performance is based on an assumption that T is always larger than T_{burst}. In particular, it is applicable to routing paths with high bandwidth, high delay or both while the TCP window size is small. Oppositely, the T value may be equal to the T_{burst}. In our simulations, the T value is equal to the T_{burst} (see Figure 4.2.1-7). In other words, the TCP sender and receiver are always engaged in sending and receiving packets with an interval between two continuous packets and no T_{idle} exists (see page 90-1). Under this situation the TCP throughput depends on the average interval between two continuous packets. The larger the interval is, the lower the throughput is. Since $T = T_{burst}$, denoting the average interval by dt, then we have:

$$T = T_{burst} = N\,dt.$$

Here, N is the size of TCP congestion window in number of packets. dt denotes the average interval between two continuous packets. If both indirect and direct routings hold a same N value, the T value would define the difference between the two routings in throughput. Certainly, that is under the situation that the TCP is always in the steady state (Congestion window reaches the maximum allowed window size) during the transmission. Apparently, if packet dropping happens the balance is broken. In our simulation when background traffic load is under 0.5 there is almost no packet dropping and the maximum allowed window size is the same to the two routings. Therefore below 0.5 background traffic load, the trend of the relative difference in goodput is similar to the trend of the relative difference in packet transfer delay. Please refer to Table B.1-3, B.1-6 and Figure 4.2.1-3, 4.2.1-6. When background traffic load increases to 0.6 and 0.7 packet dropping happens which invokes fast retransmit and fast recovery. Given that the direct routing holds a shorter RTT, it completes this process more quickly. On the other hand, the retransmission rate is higher under indirect routing since more hops are met in its routing path. Table 4.2.1-5 supports this point. These two points indicate that retransmission has more negative effect on indirect routing. That is why the relative difference in goodput becomes more apparent than in packet transfer delay under 0.6 and 0.7 background traffic load.

We use the statistical method described in Chapter 3 to achieve a 95% confidence interval for the difference between indirect and direct routing in goodput. Table 4.2.1-3 lists the relevant data about the confidence interval. NoP stands for Number of Packets.

Table 4.2.1-3: 95% confidence interval for difference in goodput between indirect and direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (NoP)	Topo1	404.0	344.4	270.4	75.0	156.1	-55.7	176.7
	Topo2	477.8	504.6	499.6	771.1	711.5	665.3	889.9
	Topo3	627.7	679.7	821.3	1196.0	1431.9	1337.8	1065.9
	Topo4	1037.9	1568.9	1912.1	2371.7	2763.8	3024.3	2943.2
Std (NoP)	Topo1	30.91	54.75	85.85	88.24	151.53	177.18	167.95
	Topo2	33.02	58.14	75.12	114.63	134.79	149.64	154.78
	Topo3	31.76	50.50	105.49	124.66	159.80	182.27	183.68
	Topo4	29.67	71.46	99.44	124.39	160.76	221.57	200.00

(Continued)

Table 4.2.1-3 (continued)

df	Topo1		56.21	54.94	57.90	57.80	53.39	57.52	57.24
	Topo2		57.37	55.97	55.12	57.63	57.96	54.61	57.92
	Topo3		52.53	46.37	52.15	47.74	57.10	52.15	57.70
	Topo4		42.90	45.53	44.94	40.69	40.80	42.52	41.36
MoE (NoP)	Topo1		61.95	109.71	171.70	176.48	303.67	354.35	336.57
	Topo2		66.16	116.52	150.53	229.26	269.57	299.87	309.56
	Topo3		63.64	101.70	211.82	250.32	320.24	366.01	367.37
	Topo4		59.75	143.92	200.28	251.40	324.90	446.24	404.21
Confidence interval (NoP)	Top1	Up	466.0	454.1	442.1	251.4	459.8	298.6	513.3
		Down	342.1	234.7	98.7	-101.5	-147.5	-410.1	-159.8
	Top2	Up	544.0	621.1	650.2	1000.3	981.0	965.2	1199.4
		Down	411.7	388.1	349.1	541.8	441.9	365.4	580.3
	Top3	Up	691.3	781.4	1033.1	1446.4	1752.2	1703.8	1433.3
		Down	564.1	578.0	609.5	945.7	1111.7	971.8	698.6
	Top4	Up	1097.6	1712.9	2112.4	2623.1	3088.7	3470.5	3347.4
		Down	978.1	1425.0	1711.9	2120.3	2438.9	2578.1	2539.0

Table 4.2.1-4: Normalised confidence interval for relative difference in goodput between indirect and direct routing

Load			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Confidence interval (%)	Top1	Up	2.20	2.52	2.97	2.10	5.14	4.93	17.36
		Down	1.62	1.30	0.66	-0.85	-1.65	-6.77	-5.41
	Top2	Up	2.56	3.43	4.37	8.58	11.15	16.13	42.89
		Down	1.94	2.15	2.35	4.65	5.02	6.11	20.75
	Top3	Up	3.28	4.39	7.08	12.78	21.06	30.95	59.35
		Down	2.68	3.25	4.17	8.36	13.36	17.65	28.93
	Top4	Up	5.25	9.77	14.79	23.85	38.23	71.37	155.41
		Down	4.68	8.13	11.98	19.28	30.19	53.01	117.88

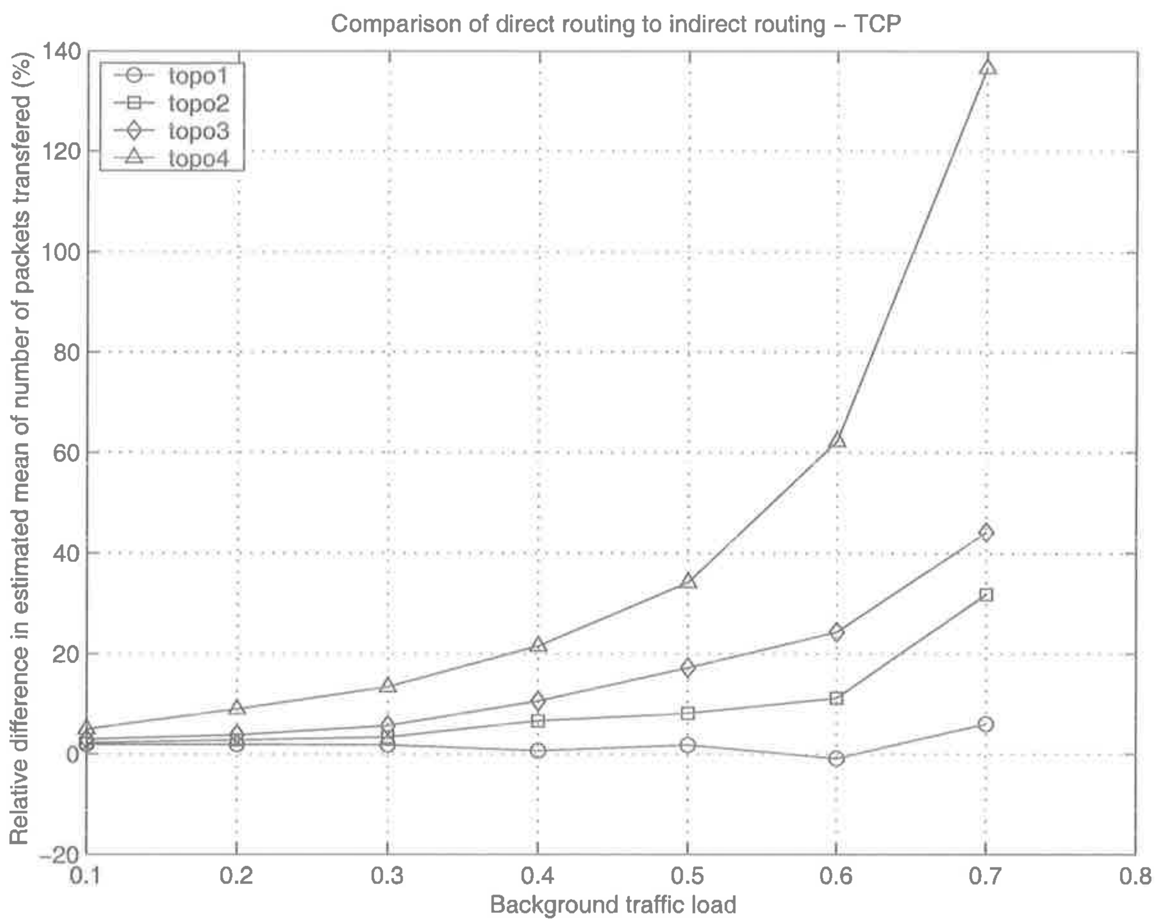


Figure 4.2.1-6: Relative difference between indirect and direct routing in goodput

Table B.1-6 and Figure 4.2.1-6 show that for topology 1 there is no apparent difference between the two routing through 0.1 to 0.7 background traffic load while for the other three topologies the relative difference becomes more and more clear with the increasing of background traffic load. From Table 4.2.1-3 and -4 for the later three topologies we can reject the null hypothesis at a significance level of 2.5%. The data for topology 1 indicates that the influence of tunnelling (encapsulation) overhead on traffic performance is insignificant, and with the increment of background traffic load, we cannot reject the null hypothesis through 0.4 to 0.7 background traffic load. That indicates that the influence of background traffic has hidden the influence from encapsulation overhead. In the view of topology scenario, we can see that there is an approximately logarithmic growth of the relative difference through topology1 to 4.

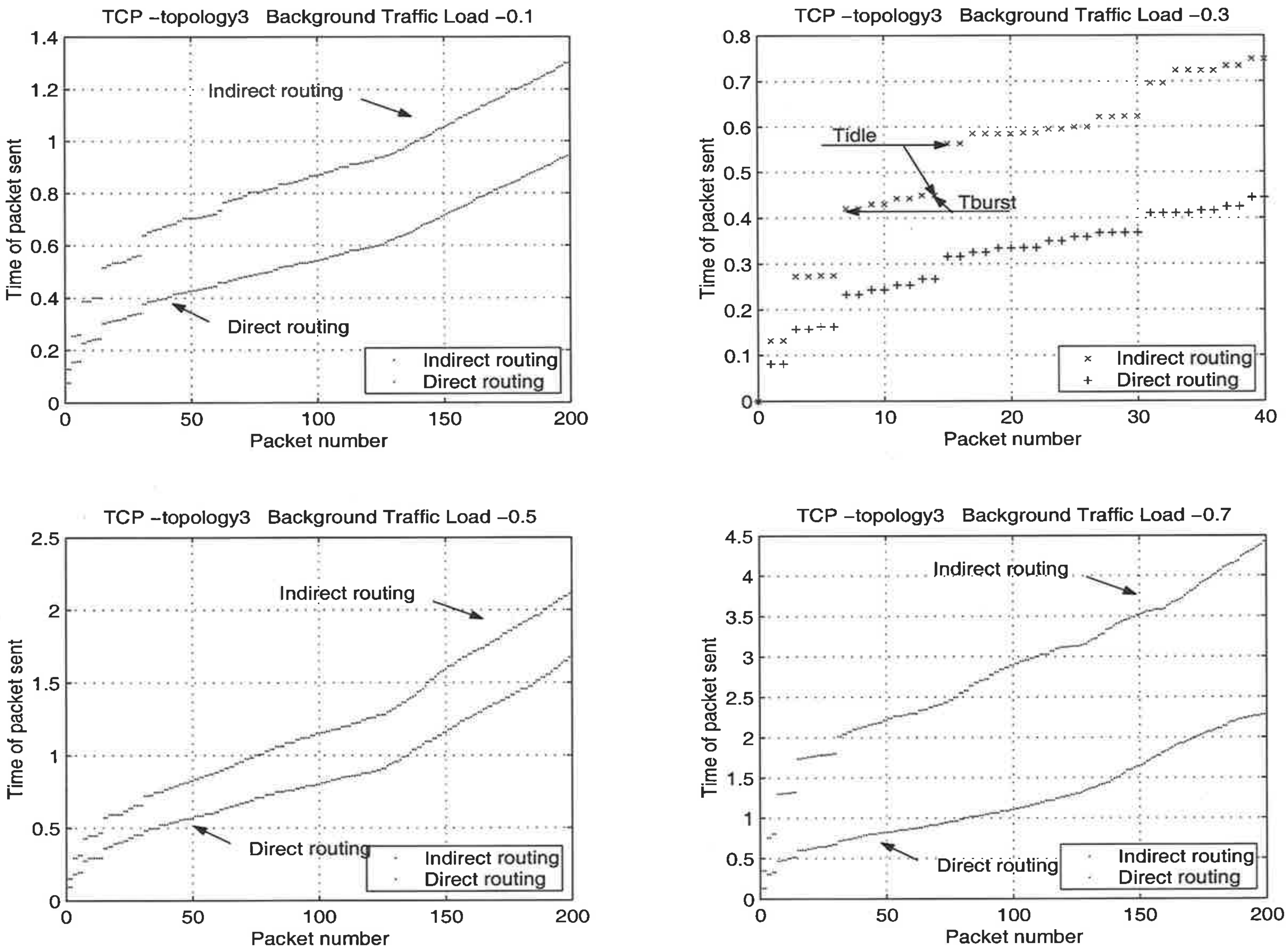


Figure 4.2.1-7: No clear T_{idle} after TCP warm-up

Retransmission rate

Table 4.2.1-5 lists retransmission rate under 0.6 and 0.7 background traffic load. Packets are dropped occasionally under 0.6 background traffic load. When the load is up to 0.7, packet dropping happens regularly. In the following we just analyze packet loss under 0.7 background traffic load. We obtain a 95% confidence interval under 0.7 background traffic load in Table 4.2.1-6. At a significance level of 2.5% we can reject the null hypothesis for topology 3, 4, and approximately for topology 2. The retransmission rate is not strictly proportional to the length of routing path, although indirect routing would have more chances to meet congestion since its longer routing path. From the previous two sections, we know that the TCP traffic throughput is lower under indirect routing, thus in a long-term view the link load is lower under indirect routing. That is helpful to reduce retransmission rate. Therefore, the retransmission rate is low in proportion to its path length under indirect routing. In the simulations, the dropping rate is low, however the packet losses result in more serious degradation under indirect routing than direct routing. It has been shown in the previous two sections by the situations under high background traffic load. It is expected that if dropping rate increases direct routing would have better performance than indirect routing.

Table 4.2.1-5: Retransmission rate under 0.6 and 0.7 background traffic load

Topology			Topo1		Topo2		Topo3		Topo4	
			Mean	Std	Mean	Std	Mean	Std	Mean	Std
Rtr rate (‰)	0.6	Indirect	0.335	0.301	0.345	0.313	0.356	0.453	0.450	0.428
		Direct	0.403	0.375	0.291	0.278	0.258	0.270	0.287	0.336
	0.7	Indirect	1.83	0.97	1.92	1.15	2.25	1.30	2.40	1.21
		Direct	1.92	1.16	1.44	0.69	1.59	0.85	1.18	0.71

Table 4.2.1-6: Difference between indirect and direct routing in retransmission rate under 0.7 background traffic load

Topology		Topo1	Topo2	Topo3	Topo4
Mean		-0.092	0.480	0.663	1.228
Std		0.276	0.244	0.284	0.257
df		56.15	47.59	49.85	47.02
Margin of error		0.553	0.491	0.571	0.517
Confidence interval	Up limit	0.460	0.971	1.234	1.745
	Down limit	-0.645	-0.010	0.092	0.712

4.2.2 Performance for Narrowband Mobile User

The simulation is performed based on GPRS features. The encapsulation overhead is 48 bytes rather than 20 bytes, which is used for broadband mobile user. In contrast to the traditional GSM system, a major feature of GPRS is the dynamic allocation of timeslots [13][14][33]. We introduce this feature into the simulation. Because of the limited capacity of GPRS radio link we can't see significant performance difference between the indirect and direct routing under low background traffic load. Therefore, we run all the simulations under 0.7 background traffic load with 1 to 8 timeslots assigned. We assume that the wireless loss can be perfectly recovered by local link layer retransmission mechanism [32]. That is, all packet losses are due to congestion. Although it doesn't accord with the reality, we believe that if time-out rate due to local recovery is trivial, depending on chance, the influence of the time-out is not significant. However, if the time-out rate is high enough we expect that the influence would be significant - reducing the performance difference between indirect and direct routing, given that the RTT is longer under indirect routing thus the TCP sender is more patient to wait for ACKs.

Similar to the analysis in section 4.2.1, we focus on packet transfer delay ($\overline{D_i}$) and goodput ($\overline{G_i}$) with a difference that the i here denotes the number of timeslots rather than background traffic load. We respectively make 50 simulation runs through one - to eight - timeslots operation. From a single simulation run we can collect the transfer delay for each packet. Then a sample mean (D_{ij}) for packet transfer delay is available for each simulation run. From 50 simulation runs, 50 sample means of packet transfer delay (D_{ij}) are available. An estimated mean ($\overline{D_i}$) and standard deviation are thus figured out for these 50 sample means (D_{ij}). A value of goodput (number of packets) (G_{ij}) is obtained for each simulation run. From 50 simulation runs, 50 values of goodput (G_{ij}) are available. Then an estimated mean ($\overline{G_i}$) and standard deviation are figured out for these 50 sample means (G_{ij}). Because of the low TCP traffic throughput due to the GPRS radio link, the packet-dropping rate is extremely low. Therefore, we don't use retransmission rate as a performance measure in this section. The TCP connection time is 100s.

Packet transfer delay

In the simulations, we employ GPRS coding scheme 2 to define the capacity of radio link (Um interface). The data rate for one channel (one timeslot) is 13.3 kb/s. correspondingly, 26.6, 39.9, 53.2, 66.5, 79.8, 93.1 and 106.4 kb/s are the capacity of 2-8 timeslots respectively. We collect and process the data in the same way used in 4.2.1 Performance for Broadband Mobile User.

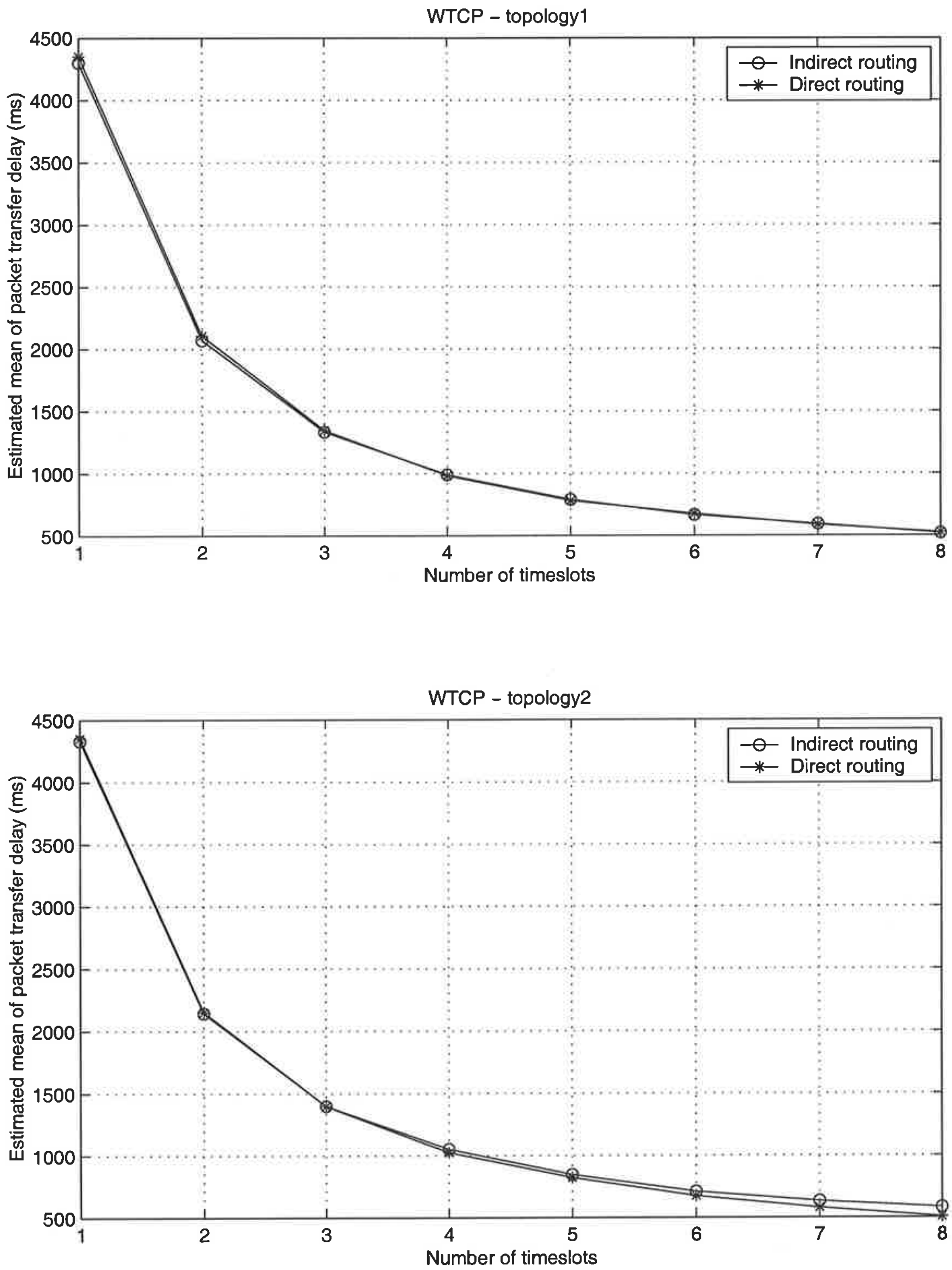
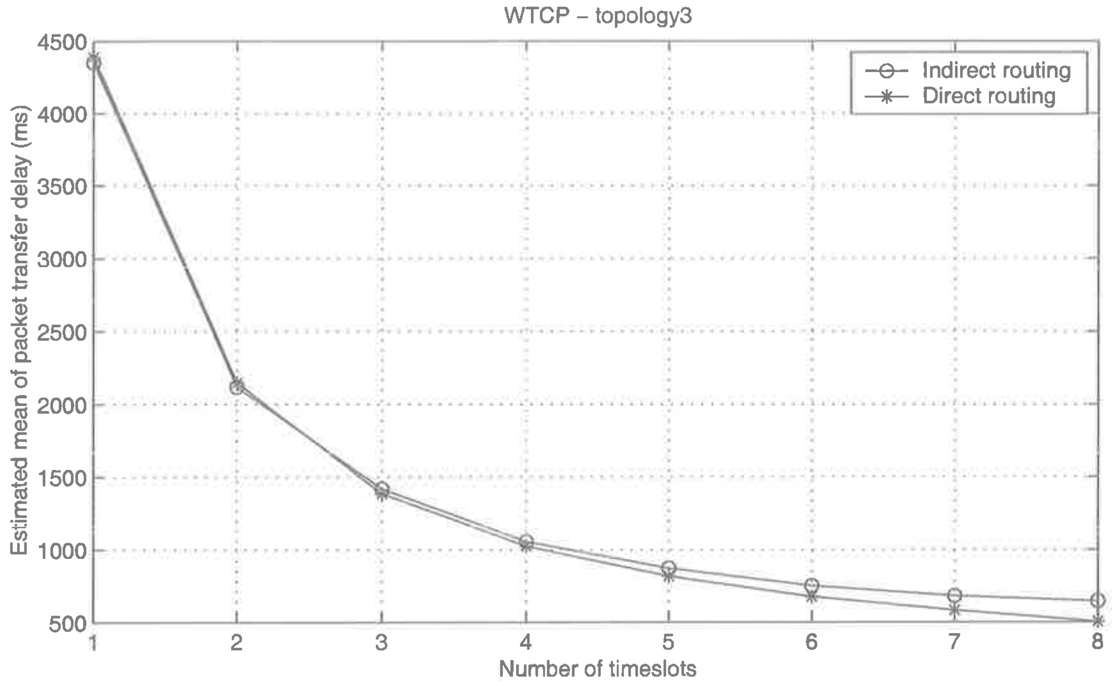


Figure 4.2.2-1: Packet transfer delay on Topology 1 and 2



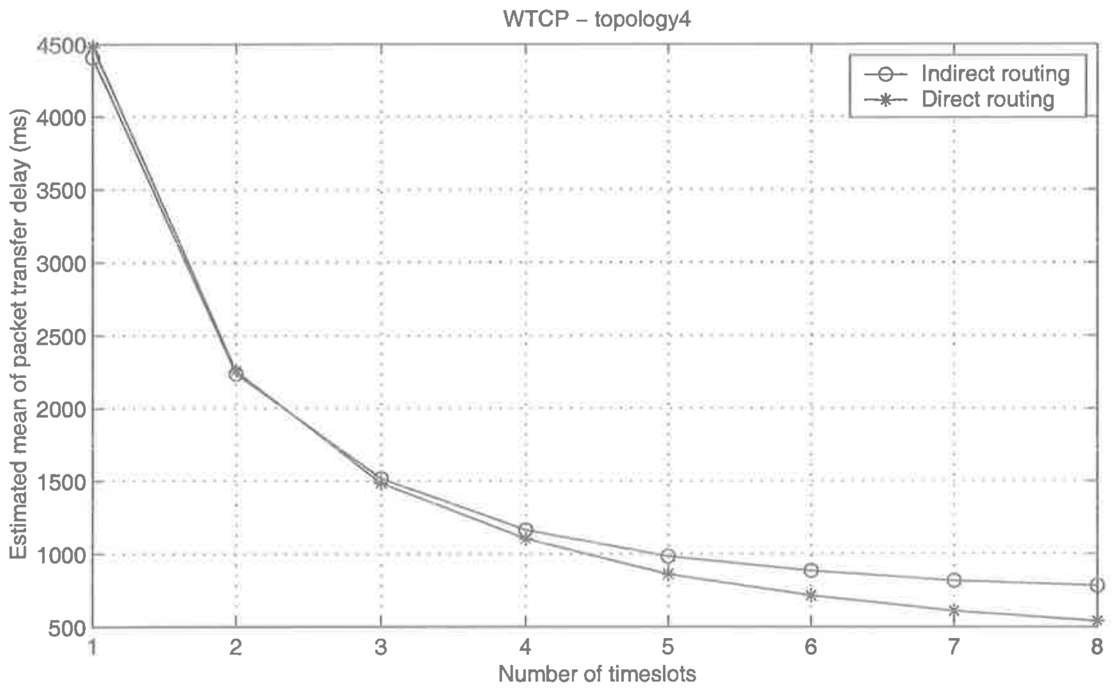


Figure 4.2.2-2: Packet transfer delay on Topology 3 and 4

Table B.2-1 and B.2-2 (Appendix B) list the estimated mean ($\overline{D_i}$) and standard deviation of packet transfer delay derived from D_{ij} for the four studied topologies. Figure 4.2.2-1 and 4.2.2-2 plot the mean packet transfer delay as a function of number of timeslots. It is shown that there is an exponential decrease of packet transfer delay as the number of timeslots increases. And with the growth of number of timeslots, the difference between indirect and direct routing becomes larger and larger. The relative difference holds an approximately exponential growth as the number of timeslots increases. It is shown in Figure 4.2.2-3 and Table B.2-3 (Appendix B). The relative difference is derived from the following formula:

$$\theta_d = \frac{d_{ind} - d_d}{d_d} \times 100\%$$

Here, θ_d denotes relative difference between indirect and direct routing in packet transfer delay. d_{ind} and d_d denote the estimated mean of packet transfer delay under indirect and direct routing respectively.

As we know the GPRS radio link capacity is very limited, it is the bottleneck in the routing path. The queuing delay at the TCP receiver would make significant contribution to the total packet transfer delay. During TCP warm-up stage, slow-start algorithm makes congestion window be increased by one packet per new ACK received. The reception of a new ACK asks for two packets be put in the congestion window. Because of the limited capacity of the TCP receiver, it is very possible that the interval between two continuous ACKs is so long that the two packets due to an ACK have been transmitted when the next ACK is received. If no background traffic introduced, then during the stage of TCP warm-up TCP packets are transmitted at twice speed available to the TCP receiver. Therefore queue is created at the TCP receiver. Certainly, the introduction of background traffic would conciliate this situation.

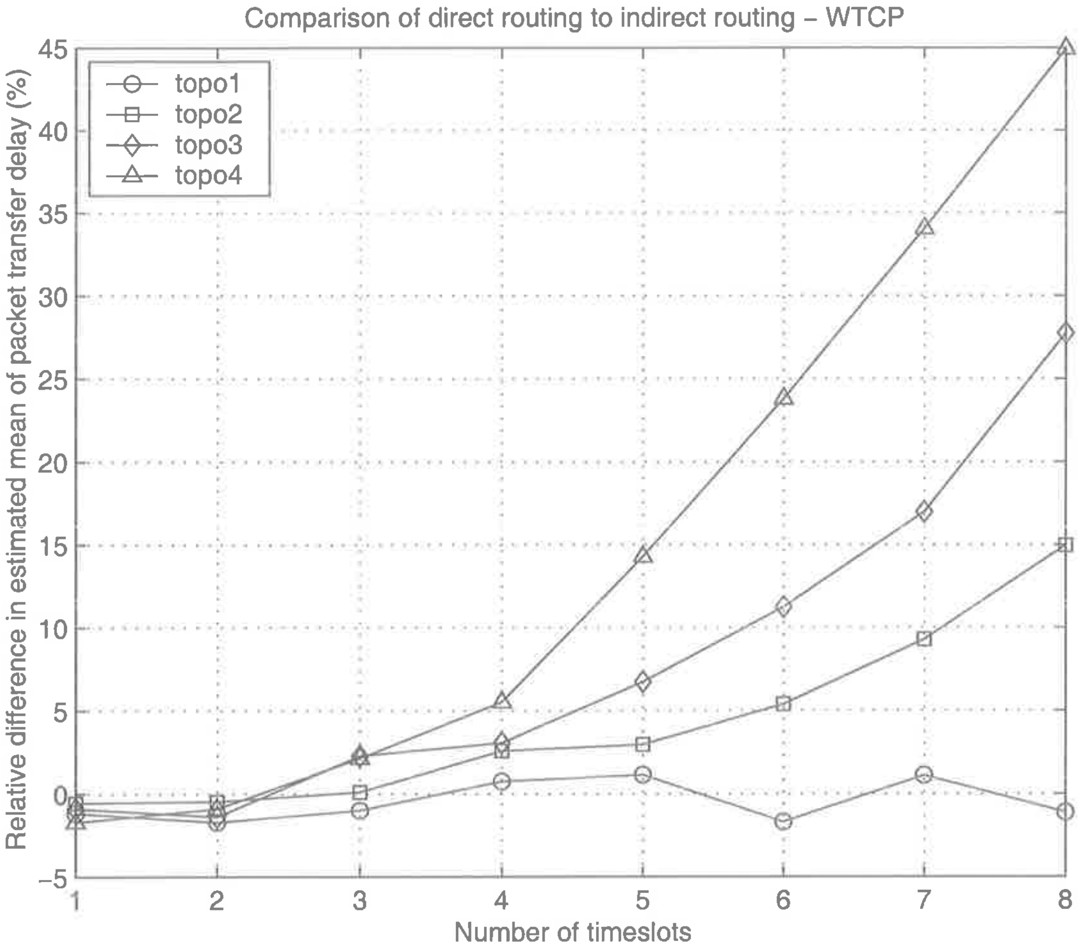


Figure 4.2.2-3: Relative difference between indirect and direct routing in packet transfer delay

With the increasing of the capacity available for the TCP receiver, the queuing and transmission delay created there diminish while the delay produced in the public Internet approximately remains the same (the delay should be slightly lengthened because of the increment of TCP traffic load). That is, the delay created on the radio link decreases while the delay created in the public Internet keeps approximately constant. Thus, in the total delay created on the routing path, the former takes a smaller and smaller proportion while the later take a larger and larger proportion. Therefore, the effect of background traffic on the routing path takes a more and more important role as the capacity on the radio link increases. We have known that background traffic have more serious degradation on packet transfer delay under indirect routing than direct routing, especially under high background traffic load. Therefore, the direct routing could improve traffic performance more and more effectively when more and more timeslots are assigned for the mobile user, as shown in Figure 4.2.2-3 and Table B.2-3. The simulation results show that there is no clear difference between the two routings for topology 1 while from topology 2 to 4 the difference becomes more and more clear. The effect of encapsulation overhead is trivial while the effect of routing path length is significant.

Table 4.2.2-1: 95% confidence interval for difference between indirect and direct routing in packet transfer delay

Timeslots		1	2	3	4	5	6	7	8
Mean (ms)	Topo1	-50.7	-35.9	-13.5	7.5	8.9	-11.3	6.5	-5.8
	Topo2	-23.7	-9.8	1.5	26.5	24.3	36.3	53.9	75.8
	Topo3	-38.8	-29.7	31.7	31.4	55.0	75.8	98.9	140.0
	Topo4	-76.3	-20.3	31.7	60.9	123.0	170.1	207.4	242.6
Std (ms)	Topo1	29.8	20.3	19.0	13.8	13.2	11.5	15.5	11.5
	Topo2	42.6	19.6	13.3	12.1	10.7	11.9	8.8	14.7
	Topo3	33.6	19.4	13.6	12.6	12.3	11.5	12.3	12.3
	Topo4	29.2	11.0	11.3	7.3	12.7	15.3	14.6	12.5
df	Topo1	93.0	90.2	97.5	97.8	97.4	95.7	82.9	95.2
	Topo2	94.9	96.6	97.9	92.6	95.9	98.0	95.1	74.5
	Topo3	80.8	90.4	90.9	89.3	86.1	93.9	88.5	92.5
	Topo4	74.9	92.8	96.2	84.9	96.5	96.1	68.0	65.1
MoE (ms)	Topo1	59.1	40.2	37.7	27.3	26.2	22.8	30.8	22.9
	Topo2	84.6	38.9	26.4	24.1	21.2	23.5	17.4	29.2
	Topo3	66.7	38.6	27.1	25.1	24.4	22.7	24.4	24.5
	Topo4	58.3	21.9	22.3	14.6	25.2	30.4	29.2	25.0

(Continued)

Table 4.2.2-1 (continued)

Confidence interval (ms)	Top1	Up	8.4	4.3	24.2	34.7	35.1	11.4	37.3	17.1
		Down	-109.8	-76.1	-51.2	-19.8	-17.3	-34.1	-24.3	-28.6
	Top2	Up	61.0	29.1	27.9	50.6	45.5	59.9	71.3	105.1
		Down	-108.4	-48.6	-24.9	2.4	3.0	12.8	36.5	46.6
	Top3	Up	27.9	8.9	58.8	56.5	79.4	98.6	123.3	164.5
		Down	-105.6	-68.3	4.6	6.4	30.7	53.1	74.5	115.6
	Top4	Up	-18.0	1.7	54.0	75.5	148.2	200.5	236.6	267.6
		Down	-134.6	-42.2	9.4	46.3	97.8	139.8	178.2	217.6

Table 4.2.2-2: Normalised confidence interval for relative difference between indirect and direct routing in packet transfer delay

Timeslots			1	2	3	4	5	6	7	8
Confidence interval (%)	Top1	Up	0.19	0.21	1.80	3.5	4.5	1.7	6.4	3.3
		Down	-2.5	-3.6	-3.8	-2.0	-2.2	-5.0	-4.2	-5.5
	Top2	Up	1.4	1.4	2.0	4.9	5.5	8.9	12.3	20.7
		Down	-2.5	-2.3	-1.8	0.24	0.37	1.9	6.3	9.2
	Top3	Up	0.64	0.41	4.2	5.5	9.7	14.6	21.2	32.6
		Down	-2.4	-3.2	0.33	0.62	3.8	7.9	12.8	22.9
	Top4	Up	-0.40	0.07	3.6	6.8	17.2	28.1	38.8	49.6
		Down	-3.0	-1.9	0.63	4.2	11.4	19.6	29.2	40.3

We introduce the T test described in Chapter 3 to see the difference between the indirect and direct routing. Table 4.2.2-1 and 4.2.2-2 present the relevant data to compute the 95% confidence interval for the difference and normalised confidence interval for the relative difference in packet transfer delay between the indirect and direct routing. It is shown that when only a few timeslots are available we cannot reject the null hypothesis at a significance level of 2.5% while we can reject the null hypothesis at the same level when enough timeslots are allocated. For whether the up-limit or down-limit of the relative difference (Table 4.2.2-2, except for topology 1), there is a growth of the relative difference as the available timeslots increase. In the view of topology scenarios, there is an approximately logarithmic growth of the relative difference through topology 1 to 4 (Table B.2-3). We cannot reject the null hypothesis under any timeslot allocation scheme for topology 1. That indicates that the encapsulation overhead is not an important factor in degrading traffic performance. Through topology2 to 4, we have more and more chances to reject the null hypothesis.

In contrast to the TCP performance for broadband mobile user, we note that the performance improvement induced by direct routing is diminished because of the effect of bottleneck in the wireless local area.

Number of packets (goodput)

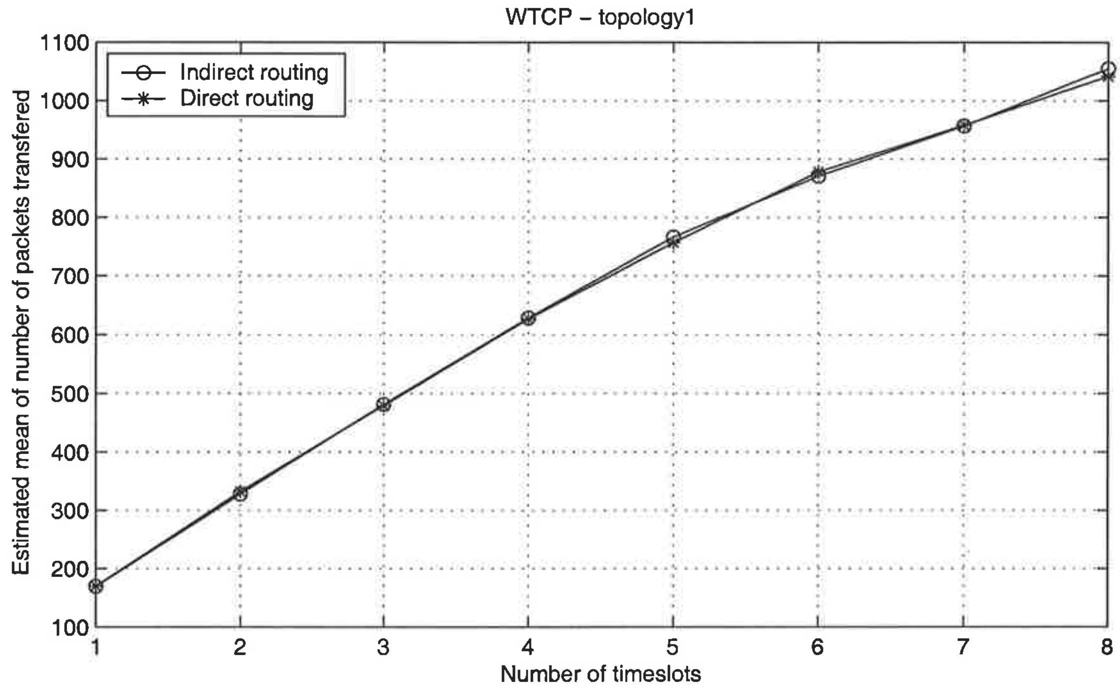
We obtain 50 values of goodput (G_{ij}) from 50 simulation runs for each timeslot allocation scheme. Table B.2-4 and B.2-5 (Appendix B) list the estimated mean ($\overline{G_i}$) and sample standard deviation of goodput for each topology and timeslot allocation scheme. Figure 4.2.2-4 and 4.2.2-5 plot the goodput as a function of the number of timeslots. It is shown that when more and more timeslots are assigned there is a logarithmic growth of goodput and the difference between the two routing becomes more and more clear. The relative difference is shown in Figure 4.2.2-6. It is approximately an exponential growth with the increasing of number of timeslots. We obtain the relative difference from the following formula:

$$\theta_g = \frac{g_d - g_{ind}}{g_{ind}} \times 100\% .$$

Here, θ_g denotes the relative difference. g_{ind} and g_d denote the estimated mean of goodput under indirect and direct routing respectively.

We have known that the GPRS radio link is a major cause of delay. Although we introduce a high background traffic load - 0.7 into the simulation, we still cannot clearly discern the difference between the two routings when only a few timeslots are assigned, for instance, 1-4 timeslots for topology 2 and 1-2 timeslots for topology 4. However, from one-timeslot to two-timeslot operation the radio bandwidth available for the TCP receiver is doubled. It indicates that the effect of bottleneck is tremendous under one-timeslot operation. We find that the simulation without background traffic introduced (experimental results not presented) and the simulation under 0.7 background traffic load (Table B.2-4, -5) basically have the same goodput under one-timeslot operation, with a very slight difference created by the difference during TCP warm-up. In fact, the goodput under 0.7 background traffic load is around 13.3kbps, the capacity of one time-slot. That is, except for the stage of TCP warm-up, 0.7 background traffic load has almost no effect on traffic performance for the two routings if only one timeslot is available. Furthermore, according to Table B.2-4, B.2-5 and Figure 4.2.2-4, 4.2.2-5, the front half part of the curves is almost linear under direct routing. Especially for topology 4, the whole curve is almost linear. That means that the goodput is completely defined by the capacity of the radio link within that range and there almost is no effect on goodput from the background traffic because of the serious effect of the bottleneck on the radio link. As the timeslots assigned to the TCP receiver is increased one by one from one to eight timeslots the capacity of the radio link is linearly increased. On the other hand, the throughput is defined by the bottleneck on the routing path - increasing the capacity of the bottleneck would induce clear improvement of the throughput while the increment of capacity of non-bottleneck may not create any significant improvement of the throughput. The linear growth mentioned above in goodput indicates that the effect of bottleneck on the radio link is so serious in the above mentioned situations that there is always a queue on the radio link (the link is fully used) and thus the throughput is proportional to the radio link capacity. We can imagine that during the data transmission there always is a queue on the radio link while the queue size is changing due to background traffic, large or small from time to time. When more timeslots are available, the linearity becomes unclear. This change is more apparent for the indirect routing. It indicates that the background traffic begins to exert its effect (creates bottleneck on the routing path) and the effect is more powerful under indirect routing because of its longer routing path than the direct routing. There is not always a queue at the TCP receiver. Sometimes there is, sometimes not. The later more frequently happens under indirect routing than direct routing. That is, with the increasing of radio link capacity, the effect due to background traffic in the public

Internet is gradually becoming clear, and therefore the performance difference due to the background traffic between the two routings becomes more and more clear.



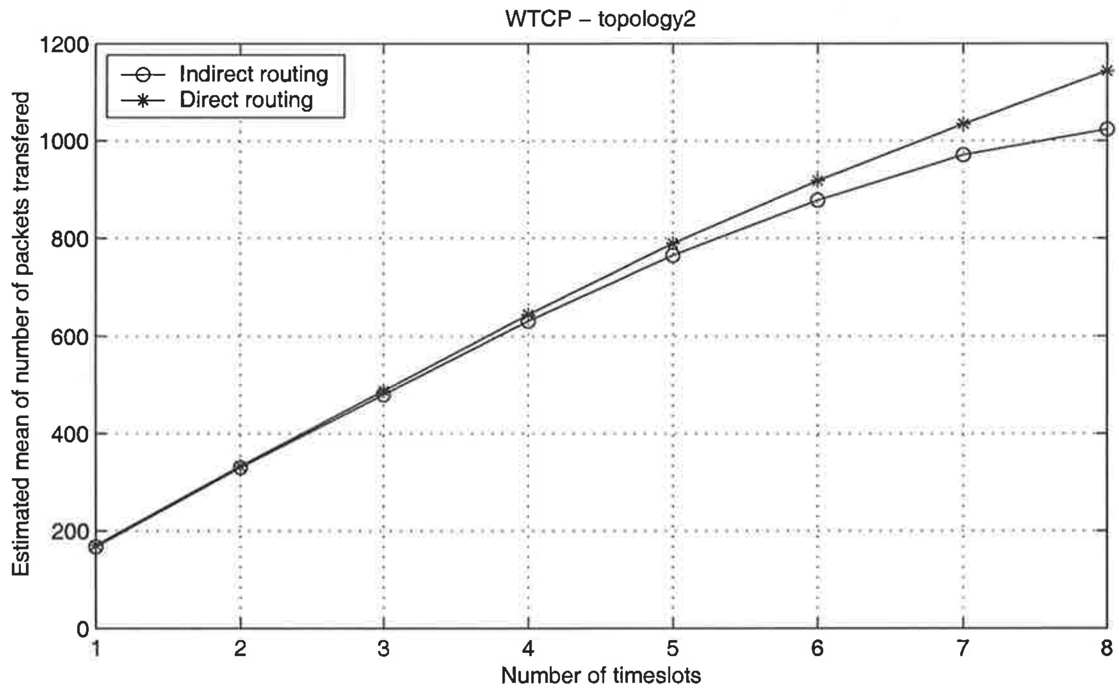


Figure 4.2.2-4: Goodput on Topology 1 and 2

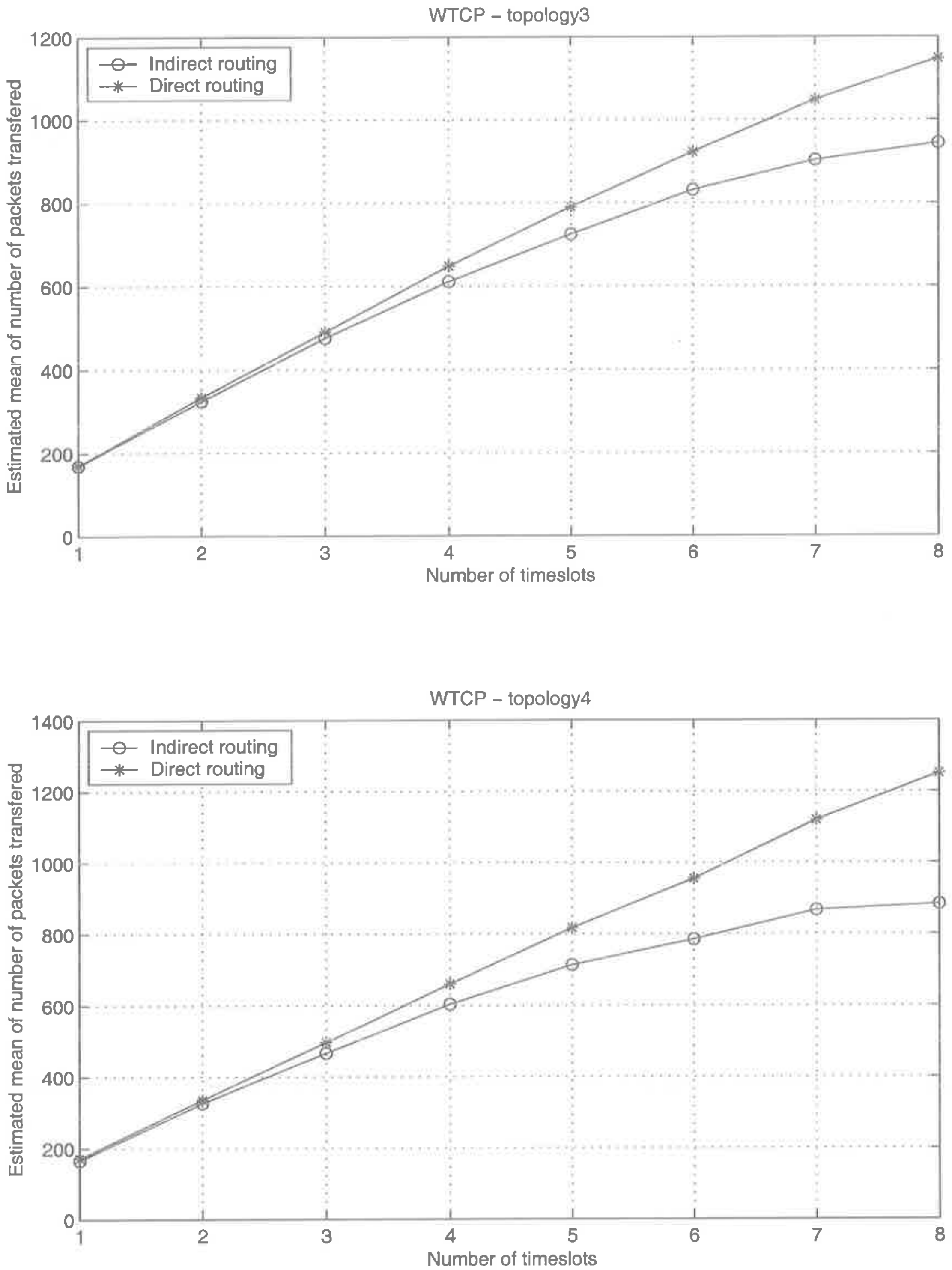


Figure 4.2.2-5: Goodput on Topology 3 and 4

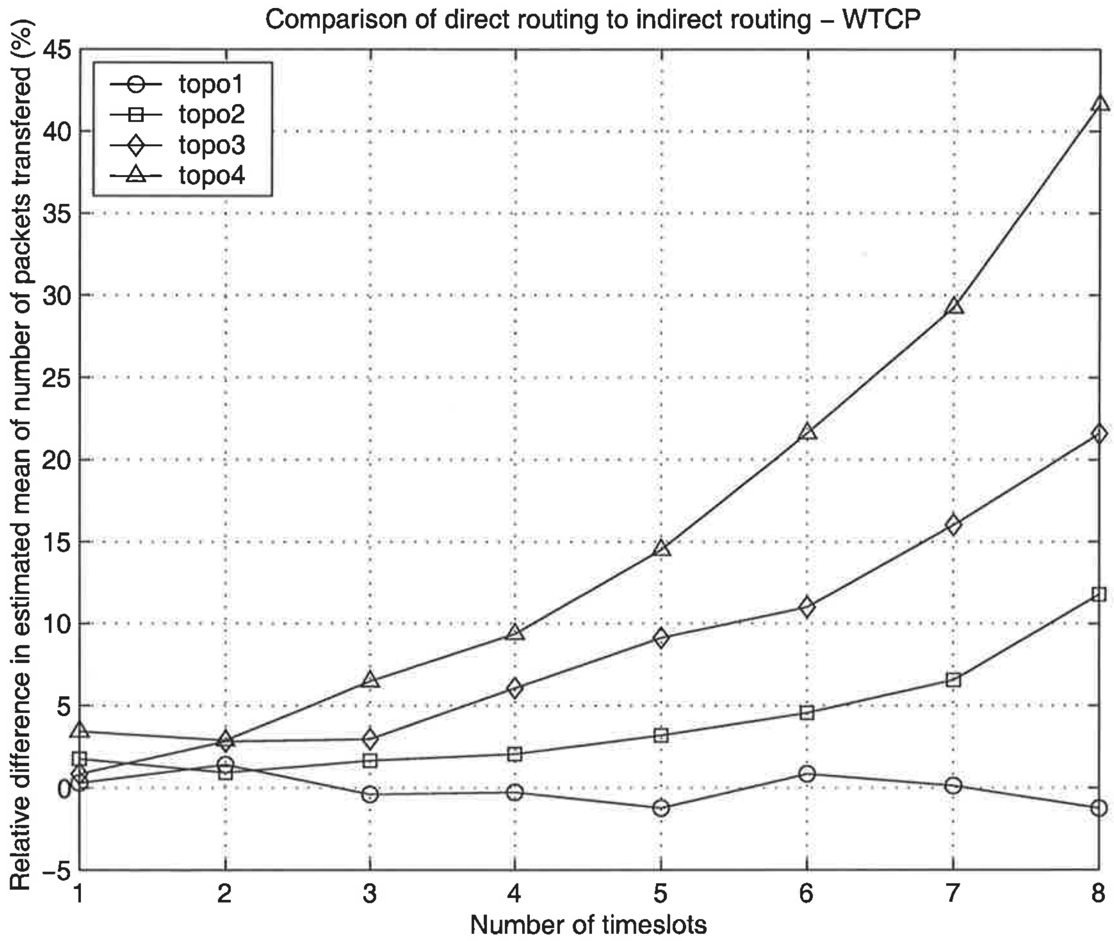


Figure 4.2.2-6: Relative difference between indirect and direct routing in goodput

Table 4.2.2-3: 95% confidence interval for difference between indirect and direct routing in goodput

Timeslots			1	2	3	4	5	6	7	8
Mean (NoP)	Topo1		0.50	4.6	-1.9	-1.7	-9.5	7.4	1.1	-13.1
	Topo2		2.9	3.1	8.0	13.0	24.4	40.0	63.7	120.7
	Topo3		1.4	9.1	14.0	37.0	66.1	91.6	144.7	203.9
	Topo4		5.7	9.4	30.2	56.5	103.4	169.6	253.3	367.8
Std (NoP)	Topo1		0.91	2.2	5.4	6.5	10.8	14.7	24.2	25.8
	Topo2		1.5	2.1	5.3	6.7	9.4	15.2	20.6	27.6
	Topo3		0.91	3.4	5.2	7.4	10.5	13.9	16.6	21.8
	Topo4		1.4	1.9	6.1	7.6	9.6	18.7	18.0	23.0
df	Topo1		98.0	61.5	98.0	97.8	95.3	96.3	86.9	96.8
	Topo2		67.2	85.0	89.6	89.6	91.1	89.2	97.9	74.8
	Topo3		79.9	58.8	80.2	66.1	80.9	87.1	89.5	93.1
	Topo4		49.7	52.5	57.6	52.2	61.6	95.6	62.2	94.9
MoE (NoP)	Topo1		1.8	4.3	10.8	13.0	21.5	29.3	48.4	51.5
	Topo2		3.0	4.3	10.7	13.5	18.7	30.3	41.0	55.1
	Topo3		1.8	6.8	10.4	14.7	20.9	27.6	33.1	43.3
	Topo4		2.8	3.8	12.2	15.2	19.2	37.3	36.0	46.0
Confidence interval (NoP)	Top1	Up	2.3	9.0	8.9	11.3	12.0	36.7	49.6	38.4
		Down	-1.3	0.31	-12.8	-14.7	-31.0	-21.9	-47.3	-64.7
	Top2	Up	6.0	7.3	18.7	26.5	43.1	70.4	104.7	175.8
		Down	-0.08	-1.2	-2.7	-0.46	5.7	9.7	22.7	65.6
	Top3	Up	3.2	16.0	24.4	51.7	87.0	119.2	177.8	247.2
		Down	-0.36	2.3	3.7	22.4	45.1	63.9	111.6	160.6
	Top4	Up	8.5	13.2	42.4	71.8	122.6	206.9	289.4	413.8
		Down	2.9	5.6	18.0	41.3	84.2	132.3	217.3	321.8

Table 4.2.2-4: Normalised confidence interval for relative difference between indirect and direct routing in goodput

Timeslots			1	2	3	4	5	6	7	8
Confidence interval (%)	Top1	Up	1.4	2.7	1.9	1.8	1.6	4.2	5.2	3.6
		Down	-0.77	0.10	-2.7	-2.3	-4.0	-2.5	-4.9	-6.1
	Top2	Up	3.6	2.2	3.9	4.2	5.6	8.0	10.8	17.2
		Down	-0.04	-0.37	-0.57	-0.07	0.75	1.1	2.3	6.4
	Top3	Up	1.9	4.9	5.1	8.5	12.0	14.3	19.7	26.2
		Down	-0.22	0.71	0.77	3.7	6.2	7.7	12.4	17.0
	Top4	Up	5.1	4.1	9.1	11.9	17.2	26.4	33.4	46.9
		Down	1.7	1.7	3.9	6.8	11.8	16.9	25.1	36.4

We use the statistical method described in Chapter 3 to achieve a 95% confidence interval for the difference between indirect and direct routing in goodput (Table 4.2.2-3). Table 4.2.2-4 lists the normalised confidence interval for the relative difference. Table B.2-6 (Appendix B) and Figure 4.2.2-6 show that for topology 1 there is no apparent difference between the two routing through one- to eight-timeslot operations while for the other three topologies the relative difference becomes more and more clear with the increasing of timeslots. In the view of topology scenarios, there is an approximately logarithmic growth of the relative difference through topology 1 to 4. From Table 4.2.2-3 and 4.2.2-4 we cannot reject the null hypothesis for topology 1 under any timeslot allocation schemes at a significance level of 2.5%. For the rest three topologies we can reject the null hypothesis at the significance level when enough timeslots are available and we have more chances to reject the null hypothesis through topology 1 to 4. It further indicates that through topology 1 to 4 the performance difference between the two routings becomes more and more significant.

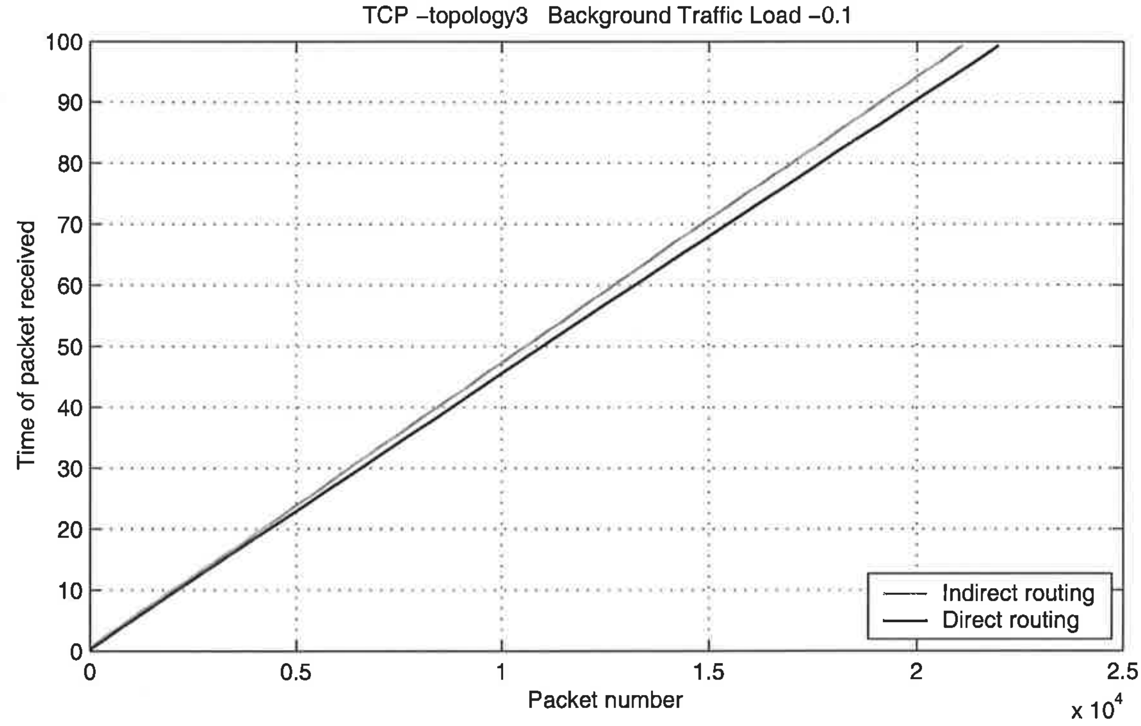
Because of the effect of the bottleneck in the GPRS service area, the performance degradation from indirect routing is not so apparent as it happens for broadband mobile user. Especially when only a few of timeslots are available, there is almost no degradation. On the other hand, if the GPRS retransmission mechanism creates long delay due to the recovery of wireless loss, the TCP sender would consider that time-out happened and therefore invokes fast recovery mechanism. It would degrade the traffic performance. Given that the indirect routing holds longer RTT, it would more possibly avoid invoking fast recovery mechanism due to the time-out created by local retransmission. That would conciliate the effect due to the difference in path length between the two routings.

4.2.3 Typical Traces of Simulations

Figure 4.2.3-1 ~ 4.2.3-8 present the typical measure results from Topology 3. Figure 4.2.3-1 ~ -4 is on behalf of TCP performance for broadband mobile user under 0.1, 0.3, 0.5 and 0.7 background traffic loads, where the communication bottleneck is located in the public Internet. Figure 4.2.3-5 ~ -8 present TCP performance for narrowband mobile user with 2, 4, 6, 8 timeslots operations under 0.7 background traffic load, where the communication bottleneck is located in the wireless local area.

From Figure 4.2.3-1 ~ -4, it is shown that the performance difference between the indirect and direct routing becomes more and more apparent with the increasing of background traffic load. When the background traffic load reaches 0.7 (Figure 4.2.3-4) packet losses due to congestion would further degrade the performance under indirect routing since where packets may meet more congestion. Figure 4.2.3-4 shows that the performance difference between the two routings is dramatic under 0.7 background traffic load. However, the difference is not always so dramatic under other environments.

All the simulations shown by Figure 4.2.3-5 ~ -8 are performed under 0.7 background traffic load. The shown performance difference is not so dramatic as Figure 4.2.3-4. With one timeslot operation, we almost cannot detect the difference. When more timeslots are available, the difference becomes more apparent. Because of the effect of bottleneck within the wireless local area, the performance improvement created by direct routing is consumed by the bottleneck.



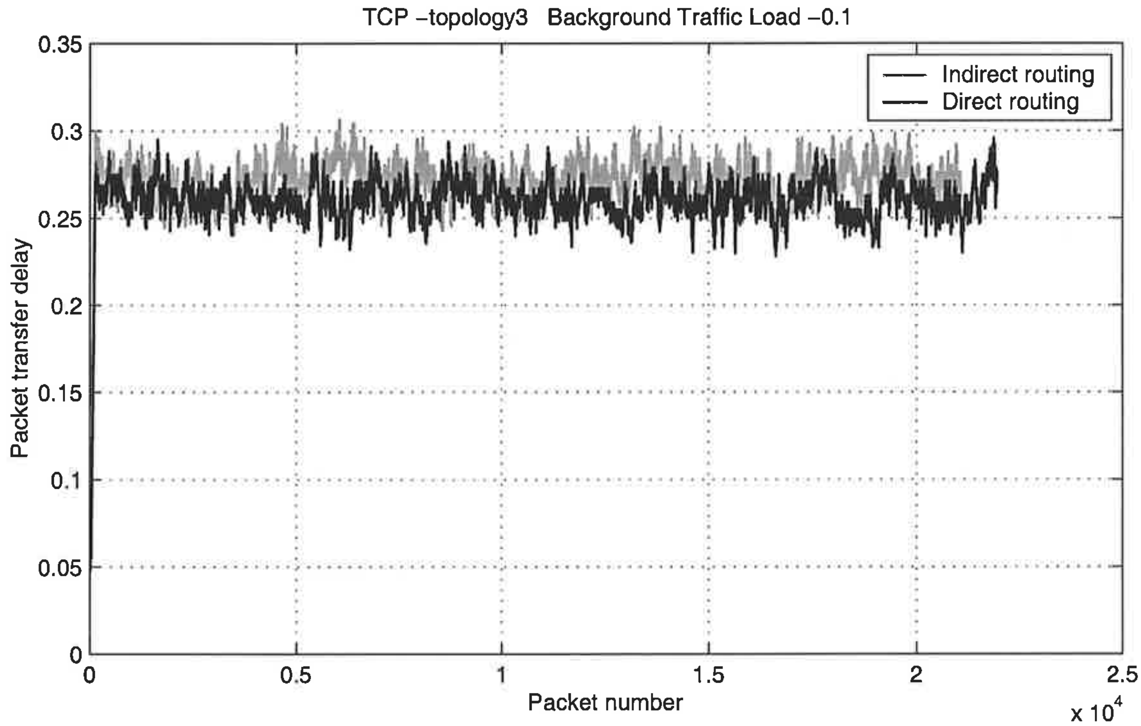
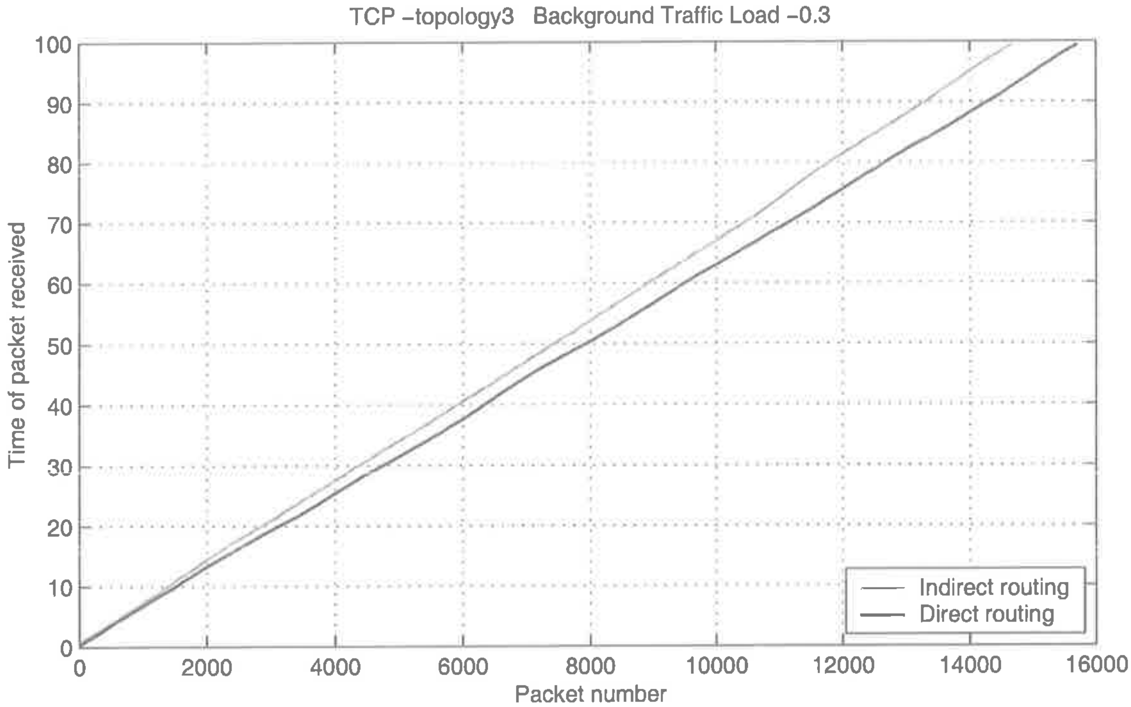


Figure 4.2.3-1: TCP performance under 0.1 background traffic load



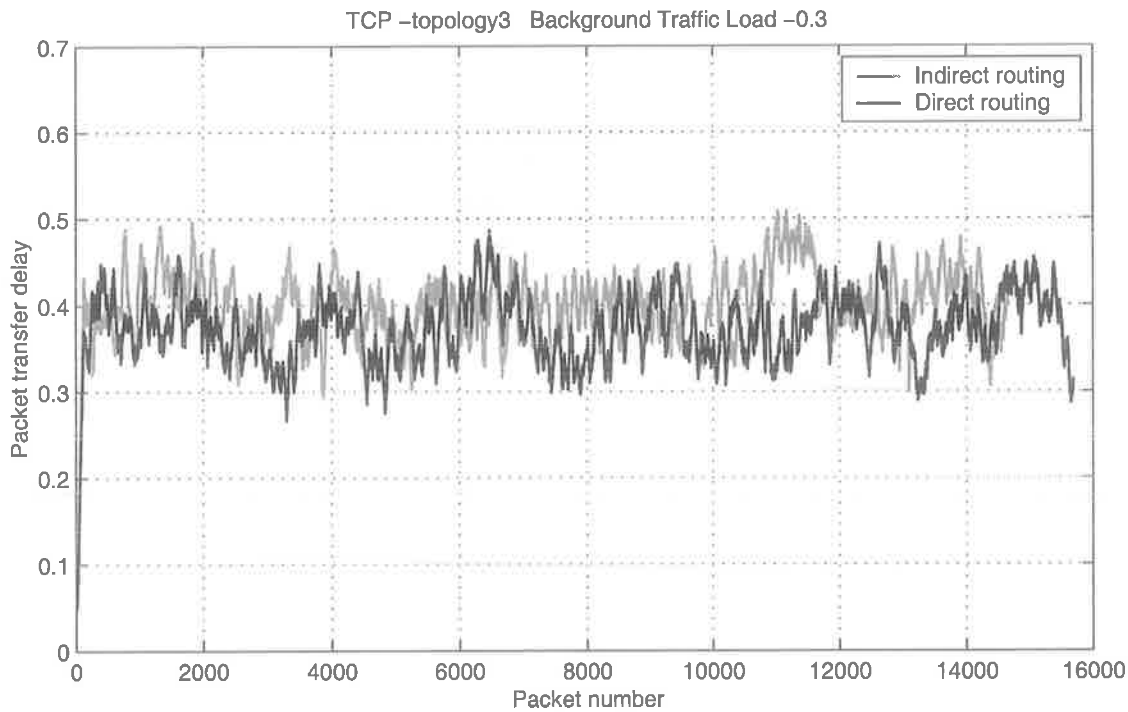
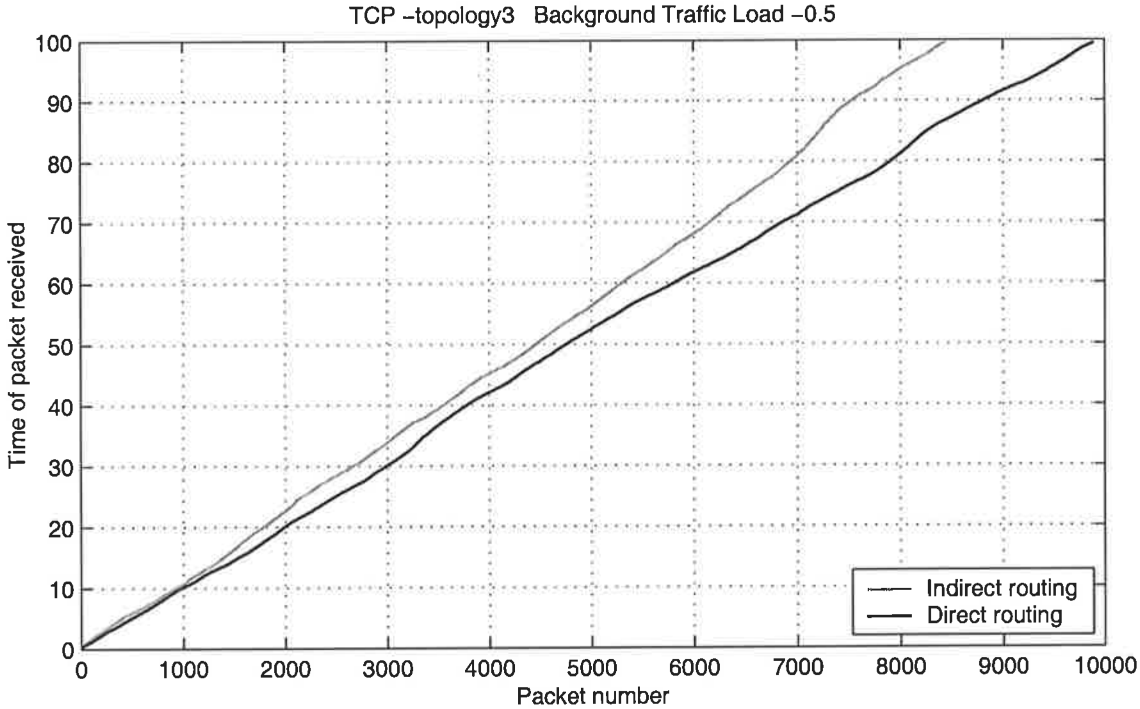


Figure 4.2.3-2: TCP performance under 0.3 background traffic load



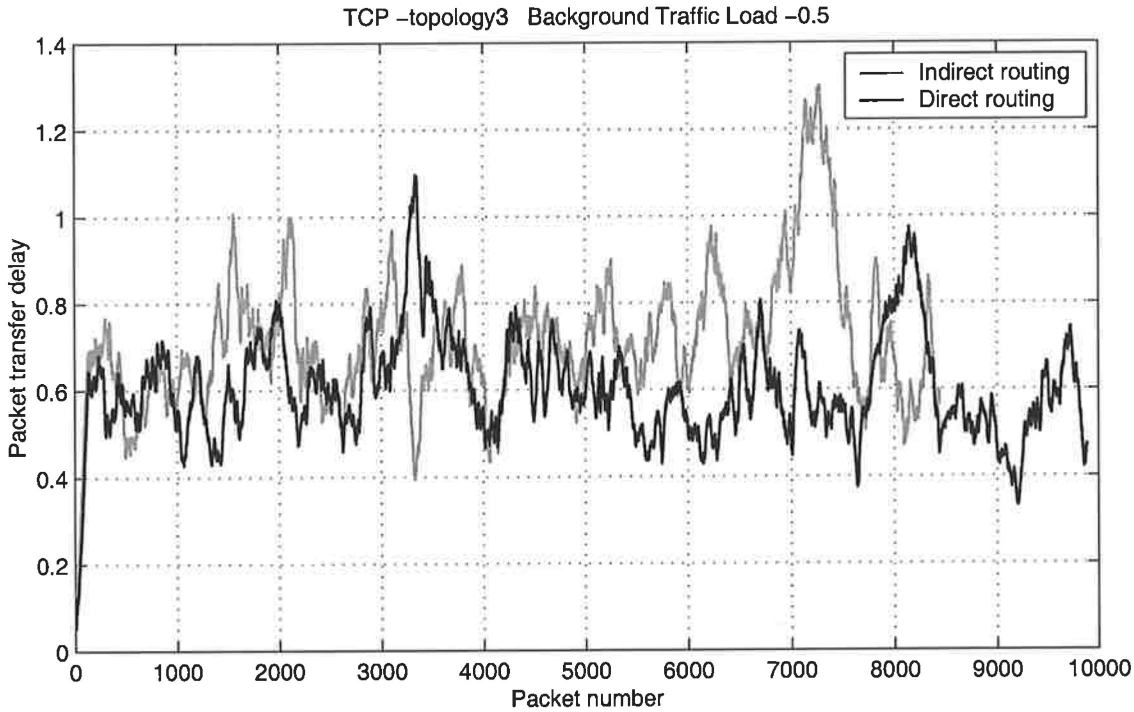
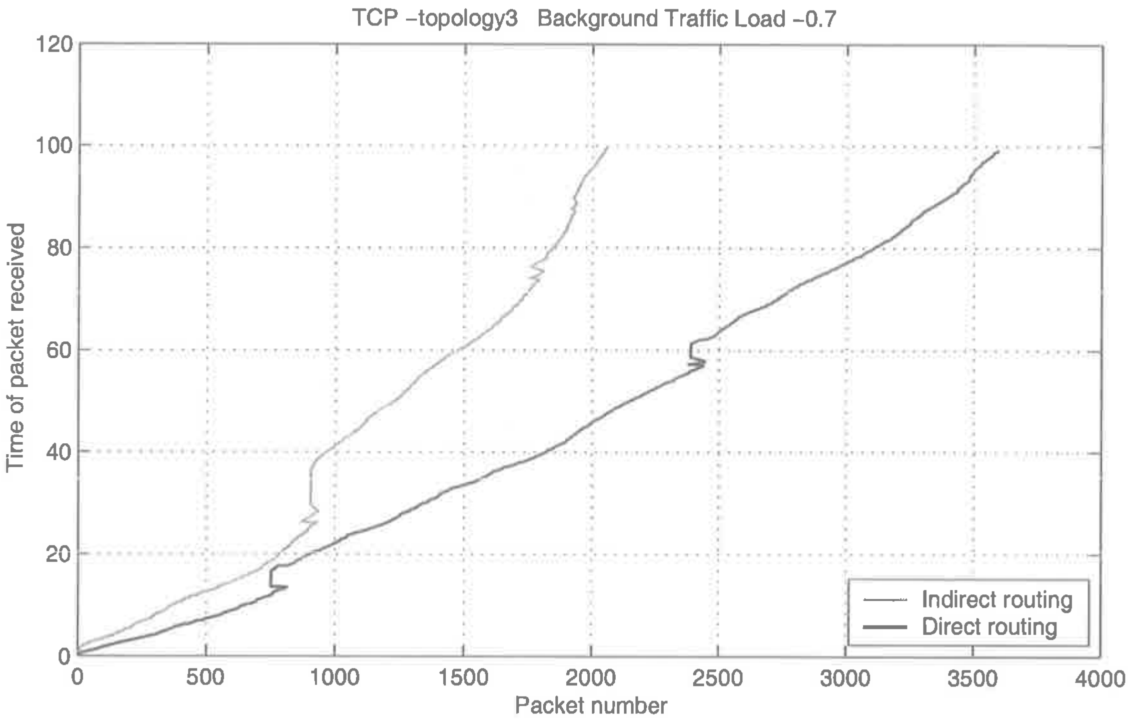


Figure 4.2.3-3: TCP performance under 0.5 background traffic load



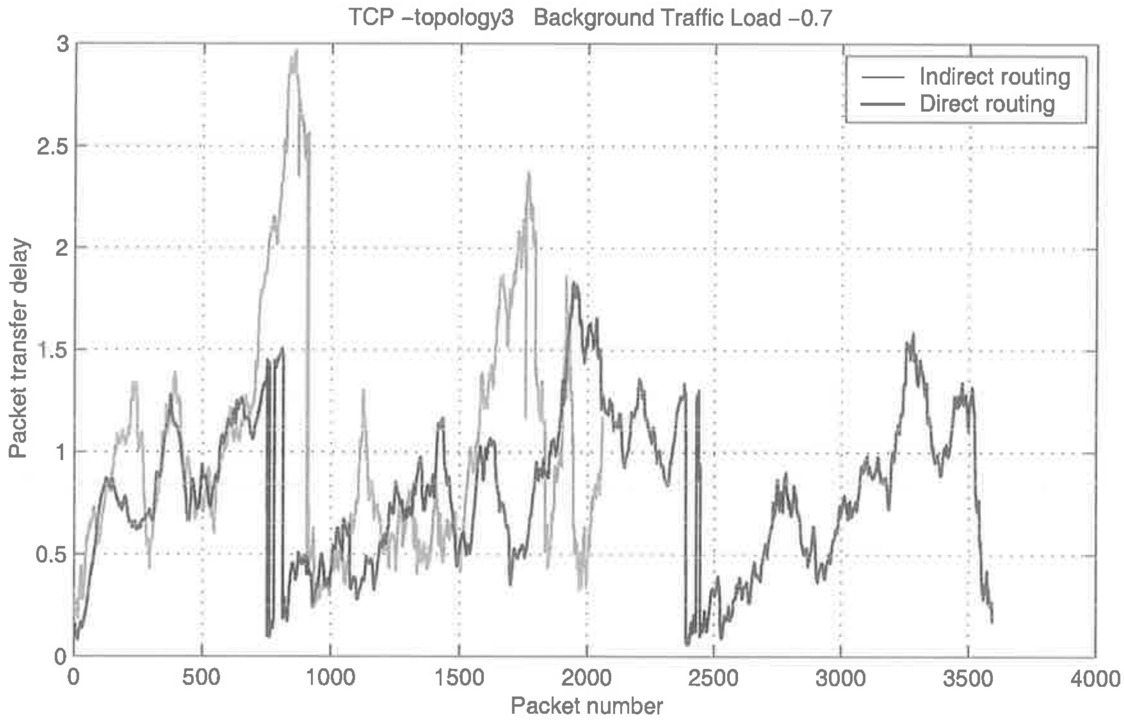
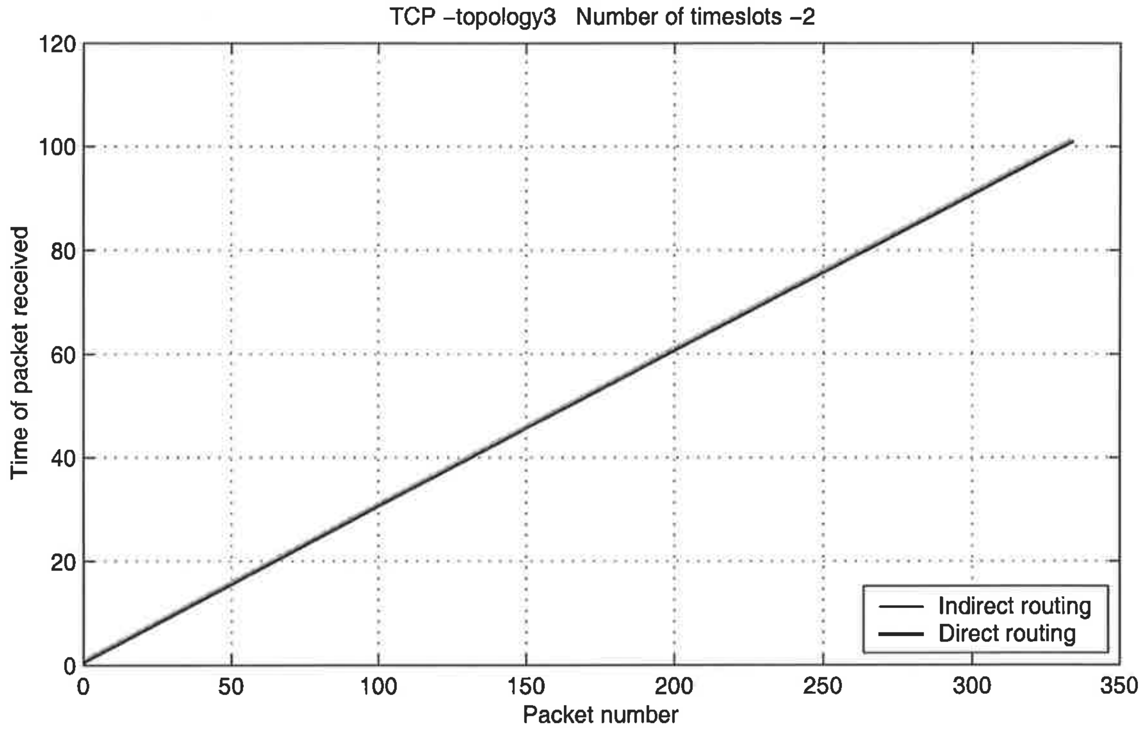


Figure 4.2.3-4: TCP performance under 0.7 background traffic load



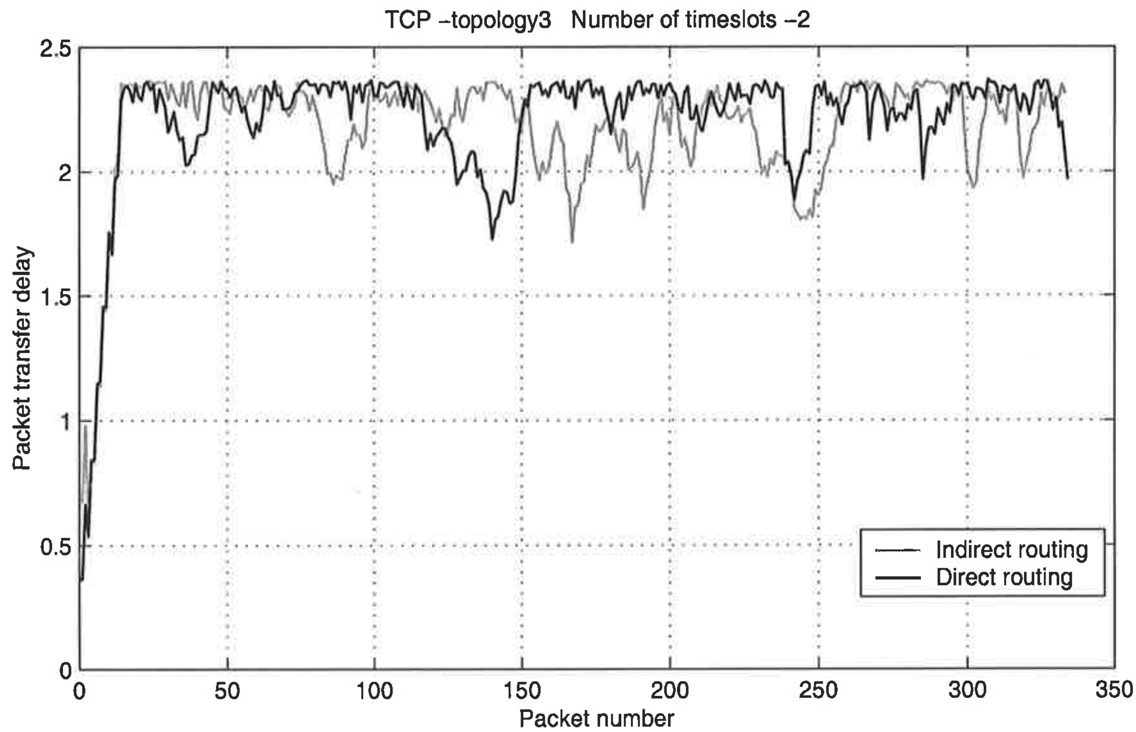
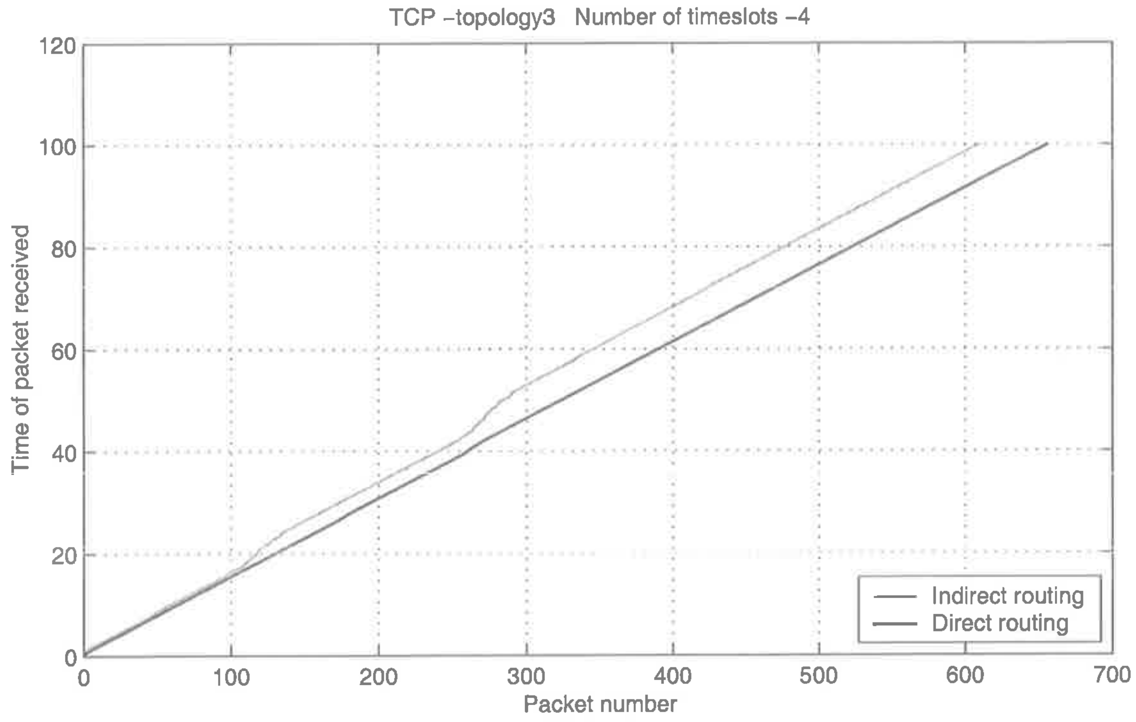


Figure 4.2.3-5: TCP performance under 0.7 background traffic load with 2 timeslots operation



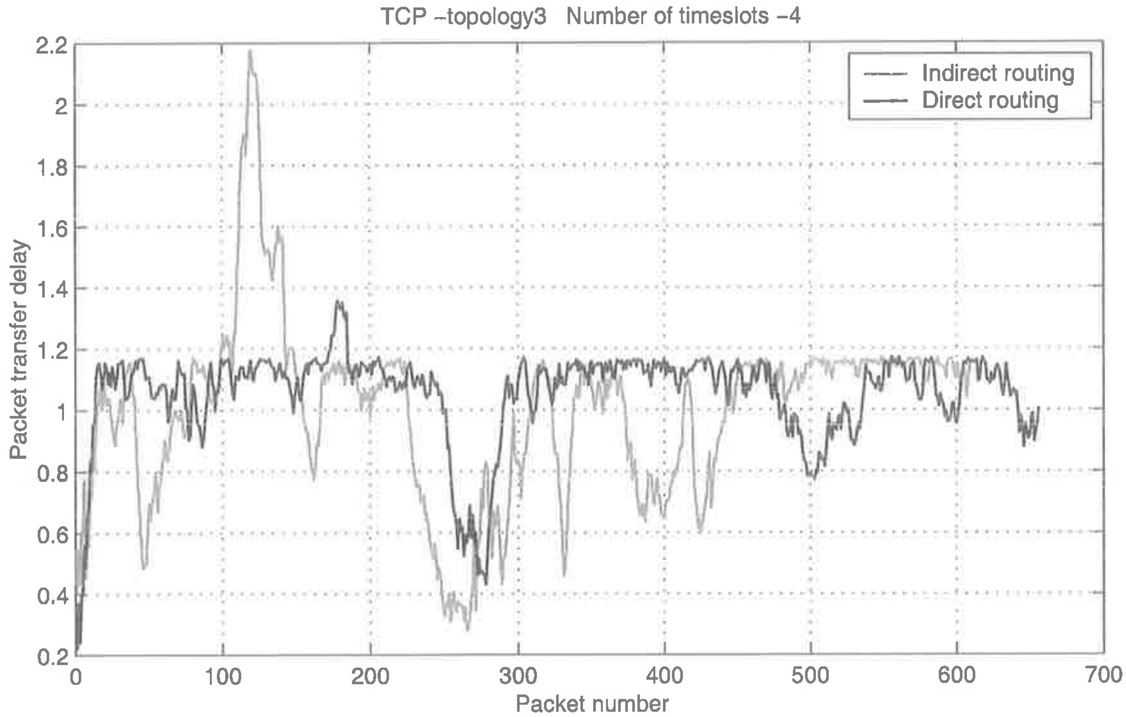
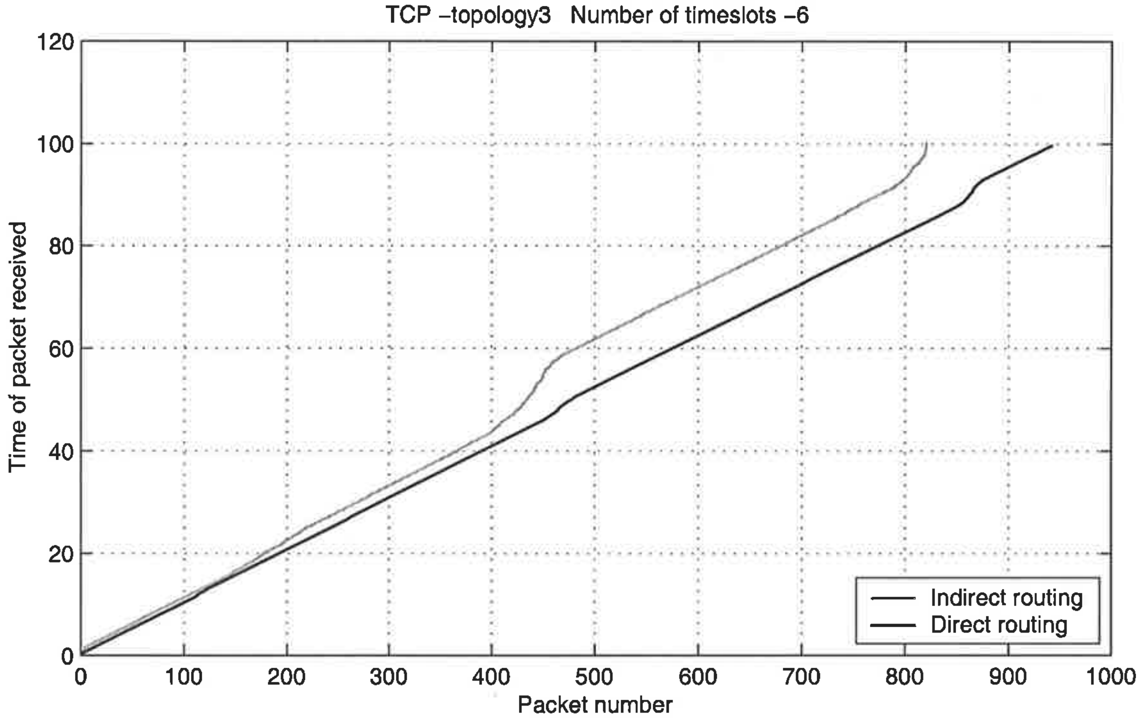


Figure 4.2.3-6: TCP performance under 0.7 background traffic load with 4 timeslots operation



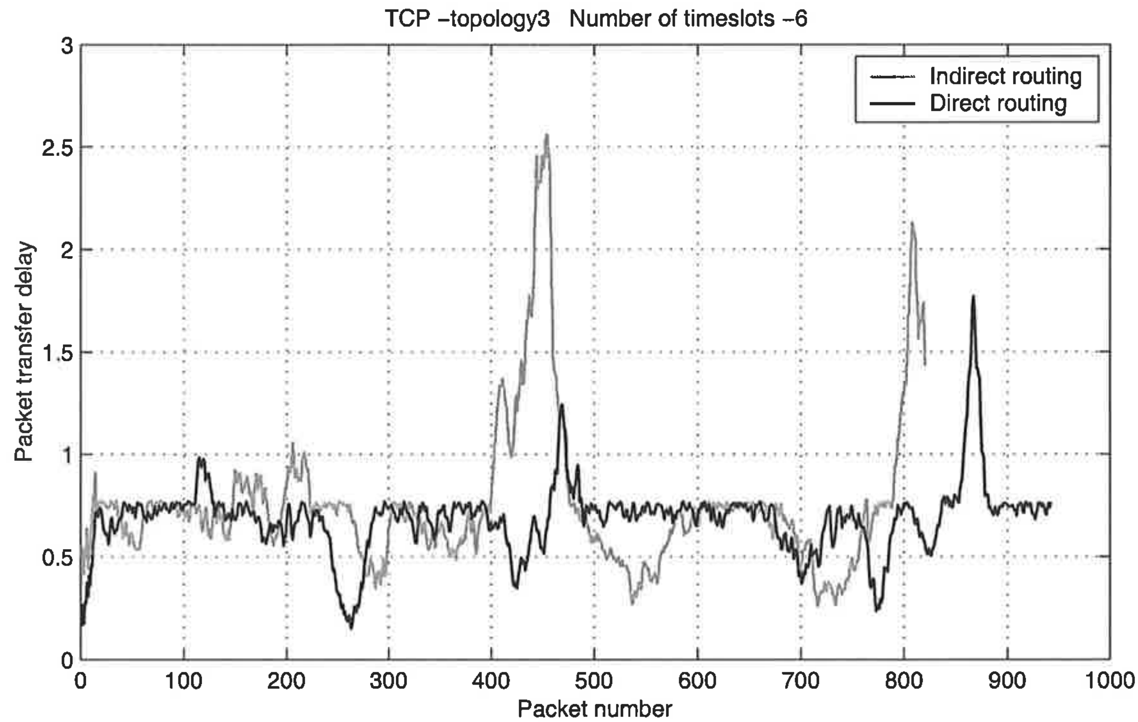
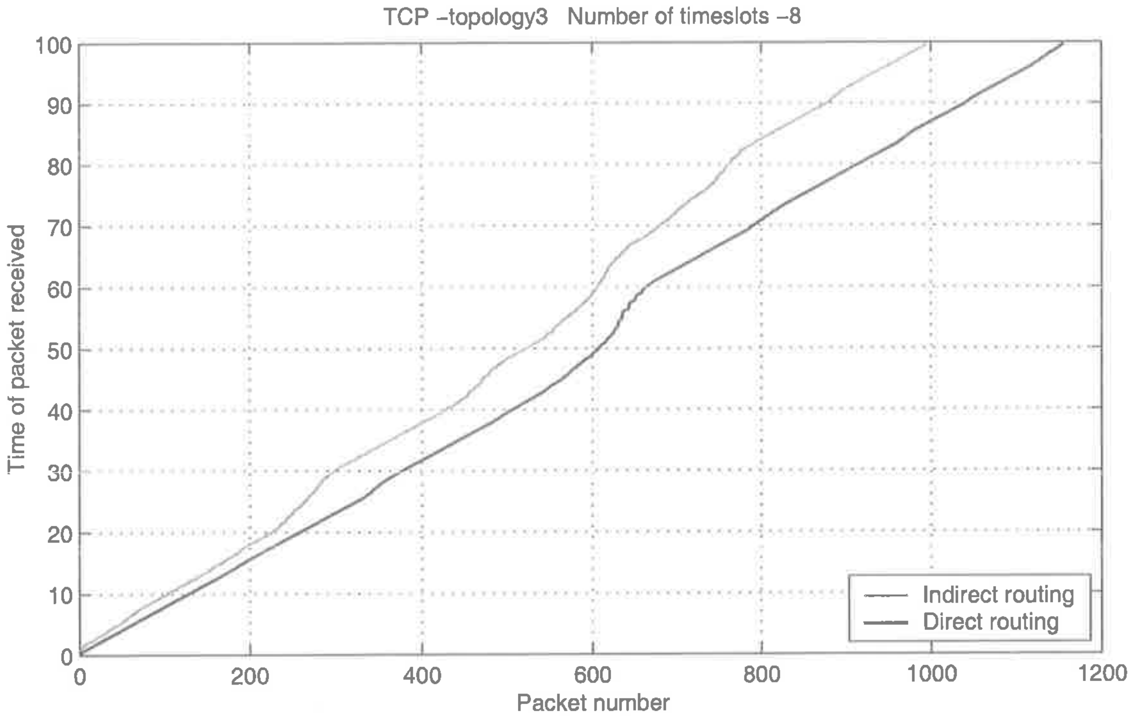


Figure 4.2.3-7: TCP performance under 0.7 background traffic load with 6 timeslots operation



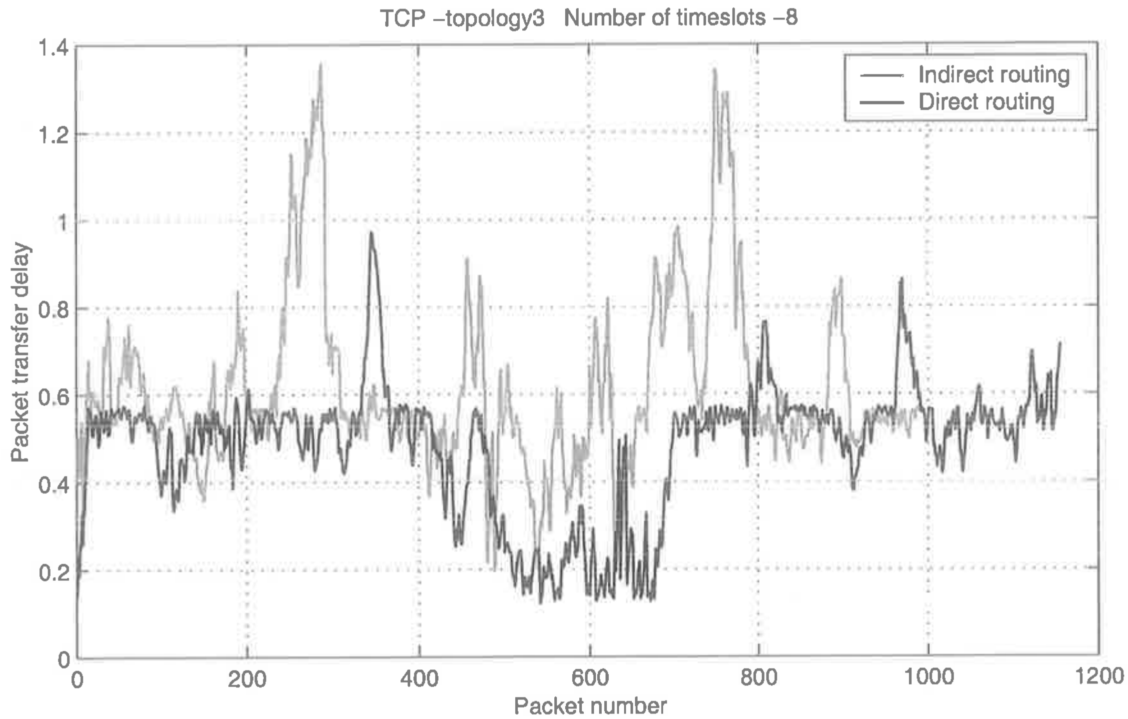


Figure 4.2.3-8: TCP performance under 0.7 background traffic load with 8 timeslots operation

Chapter 5

Conclusions

In this thesis, we summarize the basic concepts of mobile networking protocols and involved routing issues and make a comparison of indirect and direct routing in traffic performance. The traffic performance is examined based on TCP and UDP performance. Furthermore, the performance is examined for broadband and narrowband mobile user respectively. The former has an emphasis on bottleneck within the public Internet while the later has a bottleneck within the local area, the subnet where the mobile node is attached. We design four topology scenarios that describe the relative difference between the indirect and direct routing in the length of routing path (number of hops). From topology 1 to 4, the relative difference is 0%, 50%, 100% and 433% respectively. Topology 1 has an emphasis on encapsulation overhead since there is no difference between the two routings in length. We examine the performance under various background traffic loads for the studied topologies. And given a certain background traffic load, the performance is examined via changing the radio link capacity, which is described by GPRS timeslot operation.

5.1 Performance Summary

In overall, the encapsulation overhead has no significant effect on both UDP and TCP performance. In most of cases, we cannot achieve statistically significant performance difference between the two routings from topology 1. As to the rest three topologies, the improvement of TCP performance from direct routing over indirect routing is enhanced as the background traffic load increases. For broadband mobile user, the improvement of UDP performance has no strict relationship with background traffic load. For narrowband mobile user the improvement is enhanced with the increasing of background traffic load. Given a certain background traffic load, the improvement of both TCP and UDP performance becomes more apparent when more and more

5.2 Future Potential Studies

The simulation work presented in this thesis is based on a simplified modeling of the Internet. The real Internet structure and the practical effect of background traffic on the Internet are complicated, and it is exceedingly hard to characterize the traffic and network model [35]. The new studies on the Internet traffic modeling would provide more reasonable platform for performance analysis based on the Internet traffic.

The increasing importance of IP based mobile communications requires more efficient mobility and radio resource management.

Mobile IP adopts analogous technology used in GSM's call delivery to establish data routing. The home address and care-of-address in Mobile IP can be considered as the counterparts of MS's ISDN number (MSISDN) and Roaming number (MSRN) in GSM. Currently supposed Mobile IP is not suitable for micro mobility. The inefficiency of handover in Mobile IP impedes its moving towards applications in cellular level. New studies on Cellular IP [36], Mobile IP Regional Registration [37] and Local and Indirect Registration for Anchoring Handoffs [38] are trying to address this problem. [36] is designed to provide local mobility and handover support (mobility within a Cellular IP Network), and can interwork with Mobile IP to support wide area mobility (mobility between Cellular IP Networks). Cellular IP Gateway is introduced to connect Cellular IP Network to regular IP network and serves as the mobile node's foreign agent. Cellular IP Base Station, the Cellular IP Node that has a wireless interface, establishes the mobile node's interface to the network. Basically, the mobility management in Cellular IP shares the ideas used by GSM system. [37] is intended to reduce the signaling cost and improve the performance of handover via Regional (local) Registration when the distance between the visited network and the home network of the mobile node is large and the mobile node may change its location (care-of-address) frequently. In the suggestion, Gateway Foreign Agent (GFA) plays a core role in the mobility management. Under the GFA is a set of FAs. During the initial registration of mobile node with its home network, it uses the address of the GFA as its care-of-address. Therefore, the mobile node will not change its care-of-address when it changes

FA under the same GFA. In this case, the registration procedure due to changing FA can be performed in a local level. By introducing Anchor FA (AFA), [38] implements Local (regional) Registration. After successful registration with the home network, the FA involved during the registration acts as an AFA for future registrations within the visited domain (network), which are performed between the mobile node, the involved FA(s) and the AFA. The mechanisms, suggested above for micro-mobility management, are more or less analogous to the technology used in the traditional GSM and GPRS service but far from perfect and need to be further developed. On the other hand, the GPRS is being or will be deployed in current GSM system in the near future while Mobile IP has a long way to go towards its commercial Internet-wide deployment. The evolution of GPRS towards UMTS [39][40] provides a chance to harmonize the development of GTP and Mobile IP. It is possible to create a universal standard of IP mobility for whether potential UMTS or Mobile IP users. Furthermore, it is desired to establish an efficient mechanism to enable seamless roaming between heterogeneous systems such as UMTS/GPRS and Mobile IP. ETSI specification [41] presents a solution for this aim.

We have demonstrated that the effect of bottleneck of narrowband access for mobile user usually makes direct routing insignificant. With efficient radio resource management and the deployment of broadband access such as UMTS Terrestrial Radio Access (UTRA) which is based on W-CDMA, efficient implemention of optimal routing would be increasingly anticipated.

Appendix A UDP Performance

A.1 Performance for broadband mobile user

Packet transfer delay

Table A.1-1: Packet transfer delay vs. background traffic load under indirect routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	40.12	41.94	44.48	47.23	51.34	72.39	285.02
	Topo2	44.88	46.97	49.85	53.01	57.23	81.18	303.58
	Topo3	59.77	62.82	66.45	70.71	76.03	109.08	422.75
	Topo4	79.64	83.59	88.52	94.30	101.59	146.07	531.17
Std (ms)	Topo1	0.32	0.59	0.38	0.63	0.86	7.75	93.55
	Topo2	0.23	0.44	0.46	0.83	0.81	11.29	119.31
	Topo3	0.29	0.82	0.49	0.89	0.85	15.74	137.37
	Topo4	0.34	0.77	0.56	0.92	1.24	14.29	139.59

Table A.1-2: Packet transfer delay vs. background traffic load under direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	39.59	41.49	43.96	46.73	50.35	70.06	288.14
	Topo2	29.58	31.11	32.92	35.16	37.78	54.63	191.70
	Topo3	29.55	31.13	32.87	35.14	37.93	55.25	214.25
	Topo4	14.90	15.59	16.51	17.68	19.05	27.72	101.41
Std (ms)	Topo1	0.36	0.55	0.47	0.80	0.87	8.30	77.54
	Topo2	0.19	0.45	0.36	0.91	0.54	7.73	89.20
	Topo3	0.13	0.49	0.35	0.77	0.64	9.19	57.50
	Topo4	0.22	0.43	0.27	0.61	0.63	9.54	61.83

Table A.1-3: Relative difference between indirect and direct routing in packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rlt dif (%)	Topo1	1.34	0.84	1.20	1.09	1.97	3.33	-1.08
	Topo2	51.71	50.99	51.42	50.76	51.49	48.60	58.36
	Topo3	102.25	101.79	102.17	101.24	100.44	97.42	97.32
	Topo4	434.56	436.02	436.19	433.45	433.29	426.91	423.78

Standard deviation of packet transfer delay

Table A.1-4: Standard deviation of packet transfer delay vs. background traffic load under indirect routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean(ms)	Topo1	2.21	3.22	4.27	5.11	6.38	39.34	258.47
	Topo2	2.30	3.46	4.52	5.50	6.62	41.65	255.43
	Topo3	2.71	4.01	5.21	6.40	7.54	50.41	317.99
	Topo4	3.05	4.65	6.00	7.37	8.86	62.78	364.39

Table A.1-5: Standard deviation of packet transfer delay vs. background traffic load under direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	2.18	3.25	4.25	5.12	618	37.81	253.89
	Topo2	1.87	2.82	3.73	4.49	5.33	35.96	206.59
	Topo3	1.86	2.82	3.67	4.34	5.41	34.65	228.67
	Topo4	1.39	1.94	2.62	3.15	3.90	26.19	157.80

Table A.1-6: Relative difference between indirect and direct routing in standard deviation of packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rlt Dif (%)	Topo1	1.38	-0.82	0.47	-0.17	3.30	4.03	1.80
	Topo2	22.85	22.92	21.04	22.30	24.32	15.82	23.64
	Topo3	45.70	42.20	41.89	47.30	39.39	45.46	39.06
	Topo4	120.12	139.68	128.59	133.78	127.23	139.66	130.92

Table A.1-7: Relative difference between indirect and direct routing in variance of packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rlt Dif (%)	Topo1	2.78	-1.64	0.94	-0.34	6.71	8.22	3.64
	Topo2	50.93	51.09	46.50	49.57	54.55	34.15	52.87
	Topo3	112.29	102.20	101.32	116.98	94.30	111.59	93.38
	Topo4	384.53	474.46	422.54	446.51	416.35	474.36	433.23

A.2 Performance for narrowband mobile user

A. Fixed GPRS radio link capacity

Table A.2-1: Packet transfer delay vs. background traffic load under indirect routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean (ms)	Topo1	76.9	78.9	81.4	84.2	87.9	114.7	319.4	1168.1
	Topo2	81.4	83.9	86.6	89.6	93.9	123.0	344.6	1349.7
	Topo3	96.5	99.6	103.2	107.6	112.8	146.5	491.3	186.2
	Topo4	116.6	120.7	125.7	131.4	138.9	182.5	622.6	2503.8
Std (ms)	Topo1	0.26	0.49	0.44	0.55	0.90	20.07	73.37	262.92
	Topo2	0.25	0.66	0.50	0.79	0.97	23.41	96.53	356.50
	Topo3	0.31	0.52	0.48	0.99	0.87	10.63	147.02	371.66
	Topo4	0.32	0.87	0.63	1.11	1.05	8.28	105.55	462.91

Table A.2-2: Packet transfer delay vs. background traffic load under direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean (ms)	Topo1	75.7	77.6	80.1	83.1	86.4	112.0	314.3	1203.7
	Topo2	65.8	67.4	69.1	71.2	73.8	94.1	250.5	946.8
	Topo3	65.7	67.3	69.0	71.0	74.1	89.6	286.7	948.8
	Topo4	50.9	51.8	52.6	53.7	55.1	63.7	152.9	482.7
Std (ms)	Topo1	0.32	0.62	0.55	0.90	0.53	14.53	78.36	366.16
	Topo2	0.24	0.62	0.30	0.71	0.45	13.18	81.37	293.54
	Topo3	0.20	0.56	0.37	0.49	0.60	7.56	115.63	252.45
	Topo4	0.15	0.40	0.30	0.58	0.52	7.88	75.25	226.90

Table A.2-3: Relative difference between indirect and direct routing in packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Rlt dif (%)	Topo1	1.64	1.58	1.58	1.32	1.76	2.44	1.60	-2.95
	Topo2	23.86	24.49	25.29	25.81	27.26	30.63	37.56	42.55
	Topo3	46.92	47.95	49.57	51.47	52.24	63.62	71.40	96.25
	Topo4	129.06	133.16	138.92	144.67	152.25	186.55	307.11	418.76

B. Fixed background traffic load

Table A.2-4: Packet transfer delay vs. number of timeslots under indirect and direct routing (background traffic load = 0.2)

Timeslots			3	4	5	6	7	8
Indirect routing	Mean (ms)	Topo1	78.9	70.0	64.6	61.1	58.3	56.4
		Topo2	83.8	74.7	69.4	65.8	62.9	61.3
		Topo3	99.8	90.6	85.0	81.4	78.9	76.8
		Topo4	120.7	111.8	106.3	102.6	100.1	98.2
	Std (ms)	Topo1	0.65	0.69	0.47	0.45	0.46	0.49
		Topo2	0.56	0.45	0.59	0.63	0.38	0.50
		Topo3	0.65	0.63	0.55	0.55	0.58	0.48
		Topo4	0.61	0.63	0.58	0.61	0.54	0.60
Direct routing	Mean (ms)	Topo1	77.9	68.7	63.1	59.6	57.0	55.1
		Topo2	67.3	58.2	52.7	49.4	46.7	44.6
		Topo3	67.3	58.2	52.8	49.2	46.6	44.7
		Topo4	51.7	42.7	37.3	33.6	31.1	29.1
	Std (ms)	Topo1	0.77	0.39	0.51	0.49	0.48	0.64
		Topo2	0.61	0.26	0.37	0.51	0.50	0.35
		Topo3	0.45	0.50	0.35	0.41	0.42	0.44
		Topo4	0.41	0.41	0.35	0.43	0.44	0.30

Table A.2-5: Relative difference between indirect and direct routing in packet transfer delay (background traffic load = 0.2)

Timeslots		3	4	5	6	7	8
Rlt dif (%)	Topo1	1.32	1.89	2.47	2.58	2.27	2.24
	Topo2	24.5	28.4	31.7	33.0	34.9	37.7
	Topo3	48.3	55.6	61.1	65.4	69.2	71.8
	Topo4	133.6	161.7	184.9	205.6	221.9	237.8

Table A.2-6: Packet transfer delay vs. number of timeslots under indirect and direct routing (background traffic load = 0.6)

Timeslots			3	4	5	6	7	8
Indirect routing	Mean (ms)	Topo1	112.8	103.6	97.1	89.6	89.6	87.8
		Topo2	119.6	113.9	105.3	102.4	97.5	96.7
		Topo3	147.2	138.7	132.2	128.6	123.2	124.9
		Topo4	185.2	173.2	167.3	164.8	159.1	159.3
	Std (ms)	Topo1	7.73	11.6	10.0	6.25	9.16	8.17
		Topo2	10.7	21.8	13.8	10.8	15.2	12.6
		Topo3	11.0	11.4	12.4	14.2	13.5	10.6
		Topo4	21.1	15.0	16.6	11.2	15.2	18.2
Direct routing	Mean (ms)	Topo1	109.9	101.3	95.0	89.9	88.1	86.9
		Topo2	91.5	81.5	77.2	72.4	69.4	68.6
		Topo3	90.7	81.6	76.0	72.4	69.6	68.1
		Topo4	64.7	54.7	48.1	46.1	43.4	41.5
	Std (ms)	Topo1	9.03	10.7	11.0	7.73	10.9	9.94
		Topo2	8.12	8.81	9.72	8.04	9.40	11.4
		Topo3	7.45	8.60	6.89	10.5	7.27	7.61
		Topo4	10.7	8.27	4.93	8.36	7.12	6.25

Table A.2-7: Relative difference between indirect and direct routing in packet transfer delay (background traffic load = 0.6)

Load		3	4	5	6	7	8
Rlt dif (%)	Topo1	2.63	2.32	2.18	-0.35	1.66	1.12
	Topo2	30.7	39.7	36.4	41.5	40.6	41.1
	Topo3	62.2	70.0	73.8	77.7	77.1	83.5
	Topo4	186.1	216.9	247.6	257.7	266.9	283.8

Appendix B TCP performance

B.1 Performance for broadband mobile user

Packet transfer delay

Table B.1-1: Packet transfer delay vs. background traffic load under indirect routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	263.2	313.6	384.0	485.4	661.1	849.3	1007.3
	Topo2	271.1	321.1	394.4	510.0	685.0	867.3	1057.9
	Topo3	272.8	326.2	402.2	525.4	712.8	957.6	1162.4
	Topo4	287.9	345.1	425.9	556.9	760.9	1046.6	1369.7
Std (ms)	Topo1	1.53	4.60	9.77	15.82	39.28	88.41	127.07
	Topo2	1.70	3.86	7.39	22.08	43.88	72.55	139.84
	Topo3	1.44	2.74	10.11	17.60	37.69	80.28	129.52
	Topo4	1.06	3.91	8.18	14.98	31.73	73.61	160.59

Table B.1-2: Packet transfer delay vs. background traffic load under direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (ms)	Topo1	257.6	307.0	376.5	482.0	643.5	834.7	964.6
	Topo2	264.9	312.1	381.3	476.8	625.0	790.0	878.1
	Topo3	264.9	314.1	380.4	474.1	614.0	786.9	860.2
	Topo4	275.2	317.5	376.2	458.6	568.3	696.1	724.8
Std (ms)	Topo1	1.77	3.53	9.35	14.52	43.35	94.10	142.99
	Topo2	1.81	4.49	8.84	18.26	31.60	60.16	151.14
	Topo3	1.88	4.44	13.60	24.30	41.87	66.93	106.28
	Topo4	1.92	5.95	11.60	22.33	42.12	52.84	64.64

Table B.1-3: Relative difference between indirect and direct routing in packet transfer delay

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rlt dif (%)	Topo1	2.17	2.16	2.00	0.71	2.74	1.75	4.43
	Topo2	2.34	2.90	3.44	6.95	9.60	9.80	20.47
	Topo3	2.97	3.85	5.73	10.83	16.09	21.69	35.13
	Topo4	4.61	8.71	13.22	21.43	33.90	50.34	88.99

Number of packets (Goodput)

Table B.1-4: Goodput vs. background traffic load under indirect routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (NoP)	Topo1	21156	17992	14903	11967	8951	6060	2957
	Topo2	21220	18084	14881	11663	8800	5984	2796
	Topo3	21073	17802	14600	11319	8318	5505	2415
	Topo4	20909	17533	14287	10998	8079	4863	2154
Std (NoP)	Topo1	108.51	235.72	339.33	351.59	493.14	654.21	611.91
	Topo2	120.96	2.2.63	255.56	461.34	528.89	502.20	610.15
	Topo3	101.22	138.18	333.20	353.63	578.77	575.74	685.36
	Topo4	73.27	191.11	261.51	284.07	368.75	540.42	468.36

Table B.1-5: Goodput vs. background traffic load under direct routing

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Mean (NoP)	Topo1	21560	18337	15173	12042	9107	6004	3134
	Topo2	21698	18589	15381	12434	9511	6649	3686
	Topo3	21700	18482	15421	12515	9750	6843	3481
	Topo4	21947	19102	16200	13370	10843	7887	5097

(Continued)

Table B.1-5 (continued)

Std (NoP)	Topo1	130.0	185.3	325.5	331.6	667.6	716.8	686.9
	Topo2	134.4	245.7	322.4	425.9	515.1	647.7	588.6
	Topo3	141.4	239.6	472.0	584.1	656.6	815.6	736.5
	Topo4	145.0	341.6	477.8	619.3	799.6	1086.6	990.3

Table B.1-6: Relative difference between indirect and direct routing in goodput

Load		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Rlt Dif (%)	Topo1	1.91	1.914	1.81	0.63	1.74	-0.92	5.98
	Topo2	2.25	2.79	3.36	6.61	8.09	11.12	31.82
	Topo3	2.98	3.82	5.63	10.57	17.21	24.30	44.14
	Topo4	4.96	8.95	13.38	21.56	34.21	62.19	136.65

B.2 Performance for narrowband mobile user

Packet transfer delay

Table B.2-1: Packet transfer delay vs. number of timeslots under indirect routing

Timeslots		1	2	3	4	5	6	7	8
Mean (s)	Topo1	4.30	2.07	1.33	0.99	0.79	0.66	0.59	0.52
	Topo2	4.33	2.14	1.40	1.05	0.85	0.71	0.63	0.58
	Topo3	4.35	2.12	1.42	1.05	0.87	0.75	0.68	0.64
	Topo4	4.41	2.24	1.52	1.16	0.98	0.88	0.82	0.78
Std (ms)	Topo1	165.05	115.21	98.56	67.34	63.42	52.83	92.61	52.48
	Topo2	231.49	103.78	65.43	52.79	57.38	59.52	47.69	91.64
	Topo3	202.79	110.29	57.88	52.30	71.81	62.94	70.77	68.74
	Topo4	182.30	61.44	52.28	43.31	59.46	81.77	94.43	82.03

Table B.2-2: Packet transfer delay vs. number of timeslots under direct routing

Timeslots		1	2	3	4	5	6	7	8
Mean (s)	Topo1	4.35	2.11	1.35	0.98	0.78	0.68	0.59	0.52
	Topo2	4.35	2.15	1.40	1.03	0.82	0.67	0.58	0.51
	Topo3	4.39	2.15	1.39	1.02	0.82	0.68	0.58	0.50
	Topo4	4.48	2.26	1.48	1.10	0.86	0.71	0.61	0.54
Std (ms)	Topo1	130.46	85.11	91.58	70.36	68.62	61.70	58.73	62.47
	Topo2	193.08	91.97	67.65	67.60	49.46	59.19	39.92	48.50
	Topo3	123.21	81.86	77.06	72.28	48.59	50.98	50.35	53.53
	Topo4	97.48	48.23	60.04	28.57	67.32	71.06	42.46	33.68

Table B.2-3: Relative difference between indirect and direct routing in packet transfer delay

Timeslots		1	2	3	4	5	6	7	8
Rlt Dif (%)	Topo1	-1.17	-1.70	-1.00	0.76	1.14	-1.67	1.11	-1.10
	Topo2	-0.54	-0.45	0.11	2.59	2.95	5.39	9.29	14.96
	Topo3	-0.88	-1.38	2.29	3.07	6.73	11.23	16.99	27.74
	Topo4	-1.70	-0.90	2.14	5.52	14.29	23.80	34.04	44.94

Number of packets (goodput)

Table B.2-4: Goodput vs. number of timeslots under indirect routing

Timeslots		1	2	3	4	5	6	7	8
Mean (NoP)	Topo1	169.5	327.8	481.3	628.4	766.5	870.8	957.1	1055.3
	Topo2	167.2	329.5	479.7	631.1	765.0	877.6	970.7	1023.4
	Topo3	169.0	324.2	475.7	610.8	724.3	831.8	903.3	944.4
	Topo4	165.9	325.9	466.3	603.9	713.2	784.4	866.0	883.1
Std (NoP)	Topo1	4.53	14.39	27.23	33.32	49.07	68.35	141.07	135.75
	Topo2	9.79	12.62	30.65	38.58	52.98	87.20	104.54	172.37
	Topo3	5.51	23.07	31.63	47.90	63.61	80.83	95.20	120.63
	Topo4	9.80	13.28	41.28	52.73	63.71	85.57	119.45	124.98

Table B.2-5: Goodput vs. number of timeslots under direct routing

Timeslots		1	2	3	4	5	6	7	8
Mean (NoP)	Topo1	170.0	332.4	479.4	626.7	757.0	878.2	958.3	1042.1
	Topo2	170.2	332.6	487.6	644.1	789.5	917.6	1034.4	1144.0
	Topo3	170.5	333.4	489.7	647.8	790.4	923.3	1048.0	1148.3
	Topo4	171.6	335.3	496.5	660.4	816.6	953.9	1119.3	1250.9
Std (NoP)	Topo1	4.56	5.18	26.99	31.70	58.08	77.98	97.08	121.40
	Topo2	4.30	8.35	22.32	28.10	39.99	63.03	101.02	91.95
	Topo3	3.28	7.33	18.98	20.35	38.72	55.85	69.21	95.55
	Topo4	0.84	2.52	12.24	9.50	23.06	100.37	44.29	104.17

Table B.2-6: Relative difference between indirect and direct routing in goodput

Timeslots		1	2	3	4	5	6	7	8
Rlt Dif (%)	Topo1	0.29	1.42	-0.40	-0.27	-1.24	0.85	0.12	-1.25
	Topo2	1.76	0.93	1.66	2.06	3.19	4.56	6.56	11.79
	Topo3	0.85	2.82	2.95	6.06	9.13	11.01	16.02	21.59
	Topo4	3.42	2.88	6.47	9.36	14.50	21.62	29.25	41.64

References

[1] ETSI, General Packet Radio Service (GPRS); Service description; Stage 2 (GSM 03.60 version 6.3.2 Release 1997).

[2] C. Perkins, ed., IP Mobility Support, RFC 2002, Oct. 1996.

[3] ETSI, General Packet Radio Service (GPRS); GPRS Tunneling Protocol (GTP) across the Gn and Gp Interface; (GSM 09.60 version 6.4.1 Release 1997).

[4] ETSI, General Packet Radio Service (GPRS); Interworking between the Public Land Mobile Network (PLMN) supporting GPRS and Packet Data Networks (PDN) (GSM 09.61 version 6.1.0 Release 1997).

[5] ETSI, Support of Optimal Routing (SOR); Service definition (Stage 1) (GSM 02.79 version 7.0.0 Release 1998).

[6] ETSI, General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2 (GSM 03.64 version 7.0.0 Release 1998).

[7] ETSI, Numbering, addressing and identification (GSM 03.03 version 6.3.0 Release 1997).

[8] ETSI, General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol (GSM 04.60 version 6.3.1 Release 1997).

[9] ETSI, General Packet Radio Service (GPRS); Mobile Station - Serving GPRS Support Node (MS-SGSN) Logical Link Control (LLC) layer specification (GSM 04.64 version 6.5.1 Release 1997).

[10] ETSI, General Packet Radio Service (GPRS); Mobile Station (MS) - Serving GPRS Support Node (SGSN); Subnetwork Dependent Convergence Protocol (SNDCP) (GSM 04.65 version 6.5.1 Release 1997).

[11] ETSI, General Packet Radio Service (GPRS); Mobile Station (MS) supporting GPRS (GSM 07.60 version 7.1.0 Release 1998).

[12] M. Mouly and M. B. Pautet, The GSM System for Mobile Communications, 1992.

[13] G. Brasche, B. Walke, Concepts, Services, and Protocols of New GSM phase 2+ GPRS, IEEE Commun. Mag., Aug. 1997.

[14] J. Cai, D. J. Goodman, General Packet Radio Service in GSM, IEEE Commun. Mag., Oct. 1997.

[15] J. Postel, Transmission Control Protocol, RFC 793, Sept. 1981.

[16] M. Allman, V. Paxson, W. Stevens, TCP Congestion Control, RFC 2581, April 1999.

[17] J. Postel, User Datagram Protocol, RFC 768, Aug. 1980.

[18] J. Postel, Internet Protocol, RFC 791, Sept. 1981.

[19] S. Deering, R. Hinden, Internet Protocol, Version 6 (IPv6), RFC 2460, December 1998.

[20] R. Droms, Dynamic Host Configuration Protocol, RFC2131, March 1997.

[21] C. Perkins, IP Encapsulation within IP, RFC 2003, Oct. 1996.

[22] C. Perkins, Minimal Encapsulation within IP, RFC 2004, Oct. 1996.

[23] C. Perkins, D. B. Johnson, Route Optimization in Mobile IP, draft-ietf-mobileip-optim-09.txt, Feb. 2000 (work in progress).

[24] D. B. Johnson, C. Perkins, Mobility Support in IPv6, draft-ietf-mobileip-ipv6-12.txt, April 2000 (work in progress).

[25] Y. Cho, Y. Lin, H. C. Rao, Reducing the Network Cost of Call Delivery to GSM Roamers, IEEE Network, Sept./Oct. 1997.

[26] K. Thompson, G. J. Miller, R. Wilder, Wide-Area Internet Traffic Patterns and Characteristics, IEEE Network, Nov./Dec. 1997.

[27] M. E. Crovella, A. Bestavros, Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Trans. Networking, vol. 5, pp. 835-846, Dec. 1997.

[28] V. Paxson, S. Floyd, Wide-Area Traffic: The Failure of Poisson Modeling, IEEE/ACM Trans. Networking, vol. 3, pp. 226-244, June 1995.

[29] W. Willinger, M. S. Taqqu, R. Sheman, D. V. Wilson, Self-Similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, IEEE/ACM Trans. Networking, vol. 5, pp. 71-86, Feb. 1997.

[30] K. Fall, S. Floyd, Simulation-based Comparisons of Tahoe, Reno, and SACK TCP, Computer Communications Review, vol. 26 no. 3, pp. 5-21, July 1996.

[31] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP Throughput: A Simple Model and its Empirical Validation, ACM SIGCOMM 98.

[32] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, R. H. Katz, A Comparison of Mechanisms for Improving TCP Performance over Wireless Links.

[33] M. Cheng, L. Chang, Wireless Dynamic Channel Assignment Performance under Packet Data Traffic, IEEE J. Select. Area Commun., vol. 17, pp. 1257-1269, July 1999.

[34] S. Deering, Ed., ICMP Router Discovery Messages, RFC 1256, Sept. 1991.

[35] S. Floyd, V. Paxson, Why We Don't Know How To Simulate The Internet, Proceedings of the 1997 Winter Simulation Conference, 1997.

[36] A. Campbell, J. Gomez, C-Y. Wan, S. Kim, Z. Turanyi, A. Valko, Cellular IP, draft-ietf-mobileip-cellularip-00.txt, January 2000, (work in progress).

[37] E. Gustafsson, A. Jonsson, C. Perkins, Mobile IP Regional Registration, draft-ietf-mobileip-reg-tunnel-03.txt, March 2000, (work in progress).

[38] G. Dommety, Local and Indirect Registration for Anchoring Handoffs, draft-dommety-mobileip-anchor-handoff-00.txt, March 2000, (work in progress).

[39] E. Dahiman, B. Gudmundson, M. Nilsson, J. Sköld, UMTS/IMP-2000 Based on Wideband CDMA, IEEE Commun. Mag., Sept. 1998.

[40] ETSI, Universal Mobile Telecommunications System (UMTS); General Packet Radio Service (GPRS); Service description; Stage 2 (3G TS 23.060 version 3.2.1 Release 1999).

[41] ETSI, Universal Mobile Telecommunications System (UMTS); Architectural Requirements (3G TS 23.121 version 3.2.0 Release 1999).

[42] C. Huitema, Routing in the Internet, 1999.

[43] B. L. Welch, The Significance Of The Difference Between Two Means When The Population Variances Are Unequal, Biometrika, 29, 350-362, 1938.

[44] L. J. Tashman, K. R. Lamborn, The Ways and Means of Statistics, 1979.