*Technical Report*

UCAM-CL-TR-575
ISSN 1476-2986

Number 575

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

# Reconfigurable wavelength-switched optical networks for the Internet core

Tim Granger

November 2003

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

*http://www.cl.cam.ac.uk/*

This technical report is based on a dissertation submitted August 2003 by the author for the degree of Doctor of Philosophy to the University of Cambridge, King's College.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

*http://www.cl.cam.ac.uk/TechReports/*

Series editor: Markus Kuhn

ISSN 1476-2986

# Abstract

With the quantity of data traffic carried on the Internet doubling each year, there is no let up in the demand for ever increasing network capacity. Optical fibres have a theoretical capacity of many tens of terabits per second. Currently six terabits per second has been achieved using Dense Wavelength Division Multiplexing: multiple signals at different wavelengths carried on the same fibre.

This large available bandwidth moves the performance bottlenecks to the processing required at each network node to receive, buffer, route, and transmit each individual packet. For the last 10 years the speed of the electronic routers has been, in relative terms, increasing slower than optical capacity. The space required and power consumed by these routers is also becoming a significant limitation.

One solution examined in this dissertation is to create a virtual topology in the optical layer by using all-optical switches to create lightpaths across the network. In this way nodes that are not directly connected can appear to be a single virtual hop away, and no per-packet processing is required at the intermediate nodes. With advances in optical switches it is now possible for the network to reconfigure lightpaths dynamically. This allows the network to share the resources available between the different traffic streams flowing across the network, and track changes in traffic volumes by allocating bandwidth on demand.

This solution is inherently a circuit-switched approach, but taken into account are characteristics of optical switching, in particular waveband switching (where we switch a contiguous range of wavelengths as a single unit) and latency required to achieve non disruptive switching.

This dissertation quantifies the potential gain from such a system and how that gain is related to the frequency of reconfiguration. It outlines possible network architectures which allow reconfiguration and, through simulation, measures the performance of these architectures. It then discusses the possible interactions between a reconfiguring optical layer and higher-level network layers.

This dissertation argues that the optical layer should be distinct from higher network layers, maintaining stable full-mesh connectivity, and dynamically reconfiguring the sizes and physical routes of the virtual paths to take advantage of changing traffic levels.

# Acknowledgements

I would like to thank my supervisor, Ian Leslie, for his patience and guidance. Nortel Networks, and especially Peter Roorda, provided a useful starting point for discussion and information.

Many thanks to all those who took the time to proof-read this dissertation: Ian Leslie, Steven Hand, Andrew Moore, Jon Crowcroft, Penny Granger, Caroline Randall, Derek McAuley, and David Greaves. Thanks go to all members of the Computer Laboratory, and especially members of the Systems Research Group, who made my time in the lab enjoyable. Special commendation must go to Dave Stewart, David Spence, and Julian Chesterfield, who have all shared an office with me and fielded many inane questions.

My apologies to everyone who suffered along with me during the writing of this dissertation, especially Caroline.

This is for Isabelle, making it all worth it.

# Contents

# List of Figures

# List of Tables

# Glossary

## Terms, in order of appearance

**wavelength route:** a lightpath through the network which links two nodes who may not be physically connected together. Any intermediate nodes are configured to optically pass through this wavelength to create the route. This wavelength route can then be used to provide bandwidth to higher layers. Also referred to as a route. (pp 29, 94)

**waveband:** a set of wavelengths. The wavelengths carried on a single fibre can be demultiplexed into disjoint wavebands, typically all containing the same number of wavelengths. Each waveband can then be demultiplexed further into its component wavelengths. (pp 28, 94)

**waveband path:** where a given waveband joins two nodes in the network without that waveband being demultiplexed into wavelengths for the duration of the path. Intermediate nodes will demultiplex the incoming fibre into wavebands, then arrange for this waveband to be multiplexed into the correct outgoing fibre. Wavelength routes can be formed by reserving a wavelength inside multiple waveband paths arranged in series across the network. Waveband paths are also referred to as paths. (pp 29, 94)

**fixed demand problem:** where we have a fixed traffic matrix consisting of the number of wavelength routes required between each pair of nodes, and we have to provision a network to place all these routes. We typically have a fixed topology, but can decide how many fibres, switches, etc. to place at each point in the network. The aim may be to have the least equipment cost, but may also be to create a network suitable for use in the fixed topology problem. (p 91)

**fixed topology problem:** the network topology and resources are fixed, and we have a series of different traffic demands to place over that network. For each pair of nodes in the network we would like to place a number of wavelength routes between those nodes, where the number of required routes varies over time. The aim is typically to achieve a low blocking probability; a route is blocked when it is required but unable to be added to the network. (p 92)

**dedicated protection:** if a wavelength route has dedicated protection then there is another wavelength route acting as the protection route. This second route will be node-disjoint to the primary route, so that in the event of a single failure at least one of the pair will still be operational. If the primary route fails, traffic can be immediately switched onto the protection route. (p 95)

**shared protection:** if a wavelength route has shared protection then there exists another route which is node-disjoint to the primary route, to be used if the primary route fails. However along this protection route, bandwidth may be shared amongst many protection routes. The corresponding primary routes for all the protection routes sharing any bandwidth must also be node-disjoint, so that a single failure cannot fail two primary routes which both require the same shared resource. If the primary route fails the protection route must be correctly configured before traffic can be switched over. (p 95)

**FULL architecture:** at each node incoming fibres are demultiplexed into wavelengths, which are then switched, before multiplexing back into outgoing fibres. All incoming wavelengths may be received into electrical form to leave the optical network here, all outgoing bandwidth may originate from transmitters at this node. (p 97)

**WAVEBAND architecture:** at each node incoming fibres are demultiplexed into wavebands. These may be switched as entire wavebands, before being multiplexed into outgoing fibres, or they may be demultiplexed further into wavelengths. All these wavelengths may either be received into electrical form, or switched before being multiplexed into the start of a new waveband path. Since waveband paths are switched, new waveband paths may be created dynamically. (p 97)

**PATH architecture:** similar to the WAVEBAND architecture, but the set of waveband paths in the network is fixed at provisioning time. Wavelength routes are created by joining-up waveband paths in the network that have spare capacity. (p 100)

**FIXED architecture:** all wavelength routes are fixed in the network at provisioning time; each node will demultiplex incoming bandwidth into wavelengths and have a patch panel to either receive the bandwidth into electrical form or to multiplex into an outgoing fibre. (p 100)

# Acronyms

**AS:** Autonomous System, a collection of IP networks that have a unified routeing policy and are under single administrative control.

**ATM:** Asynchronous Transfer Mode, a network layer that switches fixed sized packets or cells. ATM is connection based, with circuits either permanently configured or created dynamically using some signalling protocol.

**DWDM:** Dense Wavelength Division Multiplexing, an extension of WDM. The wavelength spacing is smaller, allowing more wavelengths to be carried on a single fibre.

**FEC:** Forwarding Equivalence Class, a class of packets that will be routed the same by a MPLS-capable network. Packets are assigning to an FEC, this is indicated by some layer specific mechanism — an extra header in a packet switched network, or a (time slot, wavelength) tuple in a GMPLS network.

**fGn:** Fractional Gaussian Noise, a method of generating a stream of numbers that is approximately self similar.

**GMD:** Gaussian Marginal Distribution. The marginal distribution refers to the probability of any single item being a given value, and ignores the order in which these items occur. A GMD is where the marginal distribution matches the Gaussian or Normal distribution.

**GMPLS:** Generalised Multi-Protocol Label Switching, the concept of MPLS extended to different multiplexing schemes. MPLS defines a way to create virtual connections across packet-routed networks, where packets are routed based on a simple look-up at each node, done by assigning each packet to a FEC. GMPLS extends this for Time Division Multiplexing (TDM), Wavelength Multiplexing (WDM) and Fibre Multiplexing.

**IP:** Internet Protocol, the packet routeing layer of the TCP/IP suite of protocols. An IP network refers to a network where all data is encapsulated inside IP packets.

**IPON:** IP-Optical Node, in an optical subnet architecture this node is an IP network node which connects to an OXC to transmit packets over the optical network.

**MEMS:** Micro-Electro-Mechanical System, are where small components are physically moved by electrical signals. MEMS optical switches can use tiny mirrors which are held in a given position to reflect optical signals through the switch.

**OXC:** Optical Cross Connect, a node of the optical network. It can demultiplex, switch and multiplex incoming optical bandwidth. They may be connected to an IPON, in which case they will export virtual interfaces to allow the IPON to inject traffic into the optical network, and receive traffic from the optical network.

**SONET:** Synchronous Optical Network, a standard which defines bit rates for electrical and optical signals, how to multiplex these rates, and how to administer the network.

**TDM:** Time Division Multiplexing, where multiple signals are made to form a single data stream. The signals are split into sections, sections from all streams are then interleaved together to form the combined data stream.

**TCP:** Transmission Control Protocol, TCP packets are encapsulated inside IP packets, and are used to create a reliable data stream between two clients, which uses congestion control.

**WDM:** Wavelength Division Multiplexing, where multiple wavelengths are transmitted through a single fibre. The frequencies of those wavelengths are spread apart to give independent signals.

# Chapter 1

# Introduction

This dissertation presents the architecture of a high-capacity optical network that offers a virtual topology abstraction to higher network layers such as IP. This abstraction is designed to allow the transmission layer to exploit optical techniques such as wavelength division multiplexing and optical switching. The optical network acts as a subnet, providing stable connectivity between peer networks whilst dynamically reconfiguring the optical core to optimise utility.

It is the thesis of this dissertation that it is possible and desirable for an optical physical layer based on wavelength switching using DWDM to create and maintain a virtual topology distinct from the physical interconnection and to present this virtual topology to higher network layers such as IP. This dissertation includes work which assesses the possible gain from reconfiguring this virtual topology and how this may be achieved.

This chapter explores the motivations for the work, and outlines the structure of the dissertation.

## 1.1 IP Networking

The Internet phenomenon continues to grow, with the quantity of IP data traffic doubling every year [Coffman01]. With the introduction of new technologies such as cable modems and ADSL providing home users with more bandwidth, and the growth of corporate networks to support higher bandwidth services such as video conferencing, the traffic offered to the core of the network is constantly increasing.

The IP suite of protocols is based around the Internet Protocol (IP), which defines a connectionless, stateless, and unreliable datagram protocol. A variety of services is built on top of this. For example the TCP protocol adds a reliable, flow controlled, and duplex stream-based service; this accounts for a large portion of current traffic [Fraleigh03]. An alternative to TCP, accounting for a growing portion of current traffic, are transport layers used for media streaming that add the ability to transport real time streams of different media.

Whilst this stateless approach generally ensures that IP networks are initially scalable, flexible, and can recover from individual node or link failures, there is a large amount of associated overhead with this approach. Each individual IP datagram has to be forwarded at each hop, by using an operation to find the longest prefix match on the destination IP address. With the public access networks of today reaching end-to-end distances of over 40 hops for some routes and typically over 15 hops,[1] a large amount of work is replicated for each packet. Also, with an increasing scale of network, the connectionless approach leads to a large routeing convergence time: during these times parts of the network will be unreachable [Paxson97a]. Other problems include the relatively slow increase in the buffer memory access speed and the physical size and power consumed by the large routers needed for this per-packet processing.

## 1.2 Virtual connections

One method to reduce the processing required for a single packet to traverse a given network would be to increase the connectivity inside that network. With a fully physically connected mesh network, every possible destination is a single hop away. However, this network design is impractical, especially for geographically disparate networks, due to the inter-router cabling required.

An alternative to this approach is to create virtual connections between network nodes that are not physically connected. A given virtual connection will span multiple physical hops, but at each intermediate node this connection will be identified and switched to the next intermediate destination. The benefit lies in having a virtual connection identification scheme that requires fewer resources than the per-packet routeing algorithm. A virtual connection based scheme pushes the complexity of the network away from the data-band path and into the control-band, in the state and signalling overhead required to create, manage, and remove connections. This also has implications for failure recovery, where this state is lost.

With virtual connections, at any given node we have a series of nodes that are physically connected and also a set of destination nodes that are virtually connected; each of these can conceptually be considered reachable in a single hop.

This technique has been used previously, in packet switched networks such as X.25 and ATM. These categorise virtual connections as Switched Virtual Circuits (SVC), which are temporary and dynamically signalled through the network, and Permanent Virtual Circuits (PVC), which are long lasting and configured by the network operator. More recently Multi-Protocol Label Switching (MPLS) supports virtual connections through

[1] These figures are based on traffic measurement undertaken as part of this work, which is reported in more detail in Chapter 3

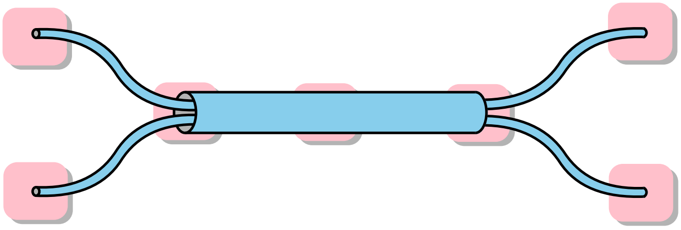

FIGURE 1.1: EXAMPLE OF TUNNELLING VIRTUAL CONNECTIONS

the concept of a Label Switched Path (LSP).

One feature of virtual connections is that we may be able to tunnel[2] one virtual connection through another. This could be used as a form of aggregation: if the physical route followed by several virtual connections coincides, then we might create a new virtual connection for this sub-path and tunnel the longer virtual connections through it. An example of this is shown in Figure 1.1. Here physical nodes are shown as the squares, with two longer virtual connections sharing their route tunnelled through the larger virtual connection. This dissertation distinguishes between a virtual connection that carries data from network ingress to network egress, referred to as a *route*, and a virtual connection that is solely used to tunnel other virtual connection through, referred to as a *path*. A route may pass through a number of paths whilst crossing the network.

One distinction between different virtual connection schemes is the method of identifying a virtual connection. A *virtual* connection refers to where there is an explicit identifier added to the data to determine the connection to which this data belongs, whereas a *physical* connection uses an implicit notion of a connection. MPLS and ATM would be classed as virtual; schemes based on Time Division Multiplexing or Frequency Division Multiplexing are physical. In contrast in this dissertation we instead use the term physical connection to refer to nodes that are directly connected, and the term virtual connection to refer to nodes that are connected via other intermediate nodes.

## 1.3 Optical Switching

One promising method of identifying virtual connections comes through the growing use of Wavelength Division Multiplexing (WDM) over single mode optical fibres as a transmission method. Independent streams of data are modulated using different frequencies and sent through the same piece of fibre. At the receiver, several parallel frequency sensitive filters can be used to separate the signals from each other. The term Dense Wavelength Division Multiplexing (DWDM) is used for the same technique, but when the gaps between adjacent wavelengths are smaller. The growth of these methods is largely due to the emergence of all-photonic amplifiers that amplify a wide

[2] In this context tunnelling refers to where one virtual connection is enclosed in another virtual connection

range of frequencies simultaneously and due to the bandwidth achievable through a single fibre – 50 THz corresponding to a possible maximum of 150 Tbit/s [Mitra01]. Current research systems are capable of using 160 wavelengths simultaneously, each transmitting at 42.7 Gbit/s to achieve 6.4 Tbit/s [Zhu03]. A further option is available to allow Time Division Multiplexing (TDM) on some or all of these wavelengths. Here the data on each wavelength is made up from multiple lower bandwidth signals; equal quantities of data from each signal are interleaved to form the aggregated signal. This creates more independent signals per fibre, each at a lower bit-rate.

A way to create virtual connections from an optical network is to use a wavelength, or possibly a time-slot on a wavelength, as the basic unit for switching. Along the virtual path, node $i$ will know that wavelength $WI_i$ on port $\alpha_i$ will be transmitted on wavelength $WO_i$ on port $\beta_i$. This may be achieved by converting the optical signals to electrical form, passing across a switch fabric then converting to optical form for transmission. However a faster method relies on the new technology of photonic switching, where there is no electrical conversion stage; the switch fabric is all-optical.

A virtual path can then be made by co-ordinating adjacent nodes to switch a section of the bandwidth from a source node to a given destination through a number of intermediate nodes. Once established, data sent by the source will undergo no further processing until reaching the destination. This leads to a decrease in network latency, moving congestion to the edges of the virtual paths and decreases the network infrastructure cost due to the removal of the processing element. The idea that the core of the Internet will evolve in a richly connected circuit switched optical network is also supported by considering the increasing difference between network traffic and router capacity, specifically memory access time and total power consumption [McKeown03], and considering the advantages gained from having a flatter more-meshed core optical network [Doshi01].

## 1.4 MPLS

Multi-Protocol Label Switching (MPLS) is an emerging Internet Standard designed to implement a virtual connection scheme. Here the IP header look-up function is replaced by a look-up operation on a packet label. This label might be a small pre-pended header, or stored in the lower layer protocol; the VPI/VCI fields on an ATM cell for example. At each step this label is checked and, together with the incoming port, gives sufficient information about the operations to be performed on this packet, such as forwarding the packet, or perhaps exchanging the label for another. This proposal has been extended for non-packet orientated networks, such as optical networks, through the Generalised MPLS (GMPLS) proposals. These have a hierarchy of label switched networks; from packet switched MPLS streams to Time Division Multiplexed (TDM) streams, Lambda Switch Capable (LSC) streams to Fibre Switch Capable (FSC) networks. Each of these surrounds the next stage of network. This dissertation concentrates on the inner two areas, with the outer boundary of the network proposed being the equivalent of LSC nodes.

The two approaches to integrating optical networks with GMPLS are the peer and overlay models. In the overlay model the two layers operate independently, each with their own control plane. The optical network controls setting up lightpaths whilst hiding the physical topology from the higher network layer, which controls traffic outside this optical core. The peer model has a single control plane that handles all areas of the network. Both have merits and disadvantages, which will be discussed in Section 2.4.3.

If the overlay model is to be employed, then the optical layer might operate using the principles outlined in this dissertation. Edge routers would run GMPLS over the provided optical core, and would provide the impetus to configure the optical network bandwidth. With the peer model, given the differences in techniques allowed by the different network types in GMPLS, from packet switched to fibre switched, it is likely that the GMPLS control plane will not treat network nodes homogeneously. For example since LSC and FSC network use wavelength frequency and fibre number as implicit connection identifiers, nesting labels is not a technology that is transferable from packet switched networks. For this reason the LSC and FSC networks could operate on the principles outlined in this dissertation.

## 1.5 Major Issues

When designing a connection based network layer using an optical DWDM infrastructure there are many issues to contend with, some of which are discussed here.

In assessing schemes that use virtual connections we have two main associated costs. The first is the cost of maintaining the state of each virtual connection. We expect to have to respond to changes in physical network connectivity and also change the virtual topology based on network management decisions. The second cost is the cost incurred by each packet at each hop along a virtual connection. We restrict the scope of this dissertation to schemes where there is no cost per hop. Processing packets per hop can either be done through converting part or all of each packet to electrical form at each hop, or by using some form of optical processing. We contend that one of the primary benefits from an optical core is the ability to remove the electrical processing bottleneck from the centre of the network, which would be lost if we needed per-packet processing to create virtual connections. The second option, optical packet processing, is beyond the scope of this dissertation and will not be considered further.

One consequence of not allowing per-packet processing is that, since we cannot use any data-band information to distinguish connections, we have to use control-band information. This means that between these control-band messages, connections will have a fixed available bandwidth.

### 1.5.1 What Do We Switch?

The optimum size of the unit of multiplexing (which is also the minimum path bandwidth) depends on several factors. On a fundamental level if the size of a single path is significantly greater than the amount of traffic on a route, then there is the potential for wasted bandwidth. This can be avoided by some form of route aggregation to fill excess capacity by joining routes with a common section into a single path. However, a fewer number of larger capacity paths mean fewer ports required on each switch and less information to share between nodes since the number of paths per node is smaller.

Conversely, smaller paths lead to more complex switching but possibly a more optimal utilisation of bandwidth. Given the limitations imposed by how finely we can divide a fibre into wavelengths, achieving smaller allocatable bandwidth might need the addition of Time Division Multiplexing to further split up a single wavelength. This increases the complexity at each node above that of merely requiring more ports. A smaller granularity of allocation might also require more frequent changes to the physical to virtual topology mapping, to keep this mapping in step with changes in traffic demand.

At the other end of the bandwidth scale, it is possible to amalgamate adjacent wavelengths to form wavebands that can be then used as an alternative switching unit. Whilst at some nodes wavebands will be switched intact, at other nodes these wavebands can be split up into their component wavelengths to either be terminated or re-amalgamated into other wavebands. Being able to switch a waveband intact allows the number of required switch ports per quantity of bandwidth to be reduced further, and can reduce the optical signal losses caused by cascading multiple demultiplexing filters with narrow pass-bands.

### 1.5.2 Why Do We Switch?

A network operator may have many reasons for wanting to reconfigure the network, primarily motivated by who is paying for what. A Virtual Private Network (VPN) has a very different set of requirements from a public access network. The former may require guaranteed bandwidth that is fault tolerant, while the latter needs a fair allocation of bandwidth between competing routes dealing with bursts of activity over many timescales.

### 1.5.3 Who Decides?

Current IP networks run a variety of protocols to discover topologies and route traffic that operates over a generally passive network. With the advent of Shared Protection Rings (SPRs) and automatic restoration under SONET, the underlying network is beginning to make a distinction between the physical network and the network that is seen by the IP layer protocols. The reason for this shift is the increase in speed gained by moving some of the functions of the network layer closer to the physical hardware. For example IP level restoration schemes rely on monitoring packet exchanges and triggering route recalculation after a number of missed messages; optical level schemes can directly monitor the received power of the signals.

The critical questions when designing an optical network using virtual paths to carry IP traffic are how much the IP layer knows about the capabilities of the underlying network, and how much the optical layer knows about the requirements of higher network layers. At one extreme, the peer network approach, the optical network could be entirely passive and export the raw interface of each switch at all nodes. At the other, the overlay approach, the optical layer could show a virtual network to the IP layer at all times and itself control how that virtual network maps to the physical one.

The major advantage of the peer architecture approach is the lack of possible duplication of protocols that occurs with the overlay approach. Topology discovery, routeing, and fault recovery are undertaken solely at the IP level. However, by careful consideration of the interaction between the two layers, it may be possible for the overlay solution to avoid this duplication, leading to a fast and efficient solution. This might include exporting a full virtual mesh topology so that the default IP route is always the single hop route. The exact width and physical path, or paths, taken by this virtual hop might change over time, but this causes no extra work for the IP layer. This has impact on the design and operation of the IP layer, which currently has limitations on the degree of connectedness at any single point: this is discussed further in Section 6.3.

### 1.5.4 How Much Do We Switch?

At each node in the optical network that interfaces with external IP networks, we will have a set of incoming fibres and will want to receive some of this bandwidth to forward onto the connected IP networks. Likewise we will have incoming bandwidth from the IP networks that we wish to add to the optical network.

One option is to have a full set of transmitters and receivers, such that all of the optical bandwidth can be either received or transmitted at each network node. However, this leads to a high equipment cost, especially if the majority of optical bandwidth passes through this node. A second option would be to have a smaller set of transmitters and receivers, so a proportion of the optical bandwidth can be added or dropped at each node. However, choosing the size of this subset is non-trivial, since we wish to reduce equipment cost without impacting on the ability to create new connections on a given wavelength.

## 1.6 Contribution

The thesis of this work is that it is possible and desirable for an optical physical layer based on wavelength switching using DWDM to create and maintain a virtual topology distinct from the physical interconnection and to present this virtual topology to higher network layers such as IP.

This dissertation presents the design of such a scheme. This is comprised of:

- the analysis of traffic patterns to assess the benefits from reconfiguring the network over a range of timescales;
- the design of the network elements used; optical switching capabilities, wavelength translation abilities, height of protocol stack required, etc.;
- the routeing algorithms used to determine a near optimal set of routes for a given topology and set of traffic demands, and the allocation of reserved capacity for failure recovery;
- the actions required by the optical network to set up required routes, to agree on route changes before and after network failure events, and to redistribute bandwidth allocations following observed changes in network behaviour; and
- the effects on the higher layer traffic of having a richly connected network layer that dynamically reconfigures itself based on traffic and failure stimuli.

The first key problem that this dissertation addresses is that of the timescale of change. In a dynamically adapting network we need to adapt to human management interaction, current traffic levels, and equipment failure events. It is known that traffic levels vary over all timescales, failure events and management interaction will happen over longer timescales. We concentrate on the ability to reconfigure based on traffic level changes, since this will also handle longer timescales. We would expect to find that the quicker we can reconfigure the more efficient our network will become. However other important factors include the minimum optical path size, the delay caused by network setup time for a reconfiguration, the packet loss we are willing to sustain, and the exact traffic patterns themselves.

Secondly we specify possible architectures for a reconfigurable optical network. One major benefit of using wavelength switching to create optical virtual connections is the possible reduction in equipment costs. Since we have no need to convert this wavelength into electrical form at any intermediate network node, we can reduce the number of receivers and transmitters at these nodes. Removing unnecessary components would allow us either to build a cheaper network with the same throughput, or to increase the capacity for the same total cost. We also contend that making our network reconfigurable will lead to benefits, since we are able to allocate bandwidth to traffic flows based on their current requirements rather than a predetermined fixed

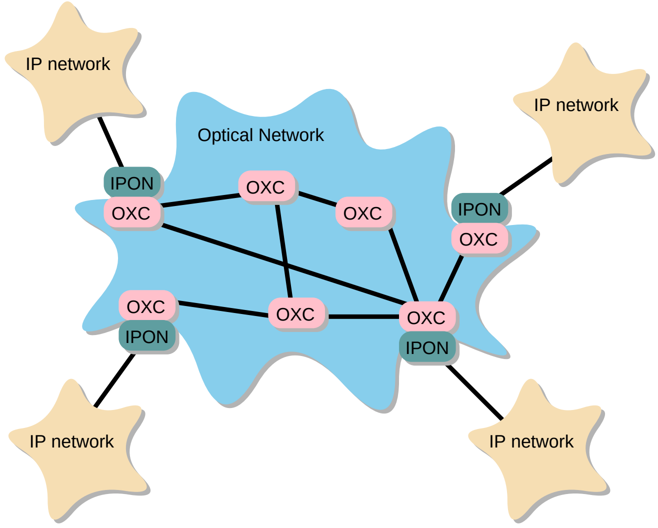

FIGURE 1.2: OPTICAL SUBNET ARCHITECTURE

allocation policy. However, the scope for reconfiguration may be limited by the cost reduction policy of removing unnecessary components: components unnecessary in one configuration may be needed for a subsequent configuration. Whether the ability to reconfigure will gain more than the advantages through minimising equipment cost is currently an open question, and may result in a compromise solution to gain optimum performance.

Another fundamental problem with designing a reconfigurable optical network is the routeing algorithm used. It needs to be able to cope with the dual demands of requiring an efficient solution to routeing the current set of virtual connections and being able to respond to changes in that set of connections in a timely manner. The algorithm must also take best advantage of the optical technology used, perhaps in exploiting a fast restoration mechanism. Another area we study is the interaction between this reconfiguring optical layer and higher network layers; the extent to which they should be separated or integrated. Although separation makes each layer conceptually easier, there is a risk of duplication of effort and reserved bandwidth at each layer. Depending on the virtual topology their may be other problems with a single fault in the optical layer causing multiple failures at the higher layers [Crochat00].

## 1.7 Proposed Architecture

The network solution proposed in this dissertation is summarised as follows. Figure 1.2 shows an overview of the network described. IP-Optical Nodes (IPONs) connect to external IP networks and participate in normal IP routeing protocols. They connect into the Optical Network by directly connecting with an Optical Cross Connect (OXC), which is an all-photonic switch with associated control processor. These switches are assumed to be DWDM capable,[3] providing a large number of possible connections carried by a single fibre that can be demultiplexed, switched and multiplexed. These operations are assumed to be data independent, although places in the optical network may exist where signals are restored/reshaped/retimed, which requires knowledge of the data signal format. OXCs are connected together by one or more fibres on a single link, these links forming a partially connected mesh network.

The optical network will provide a communication mechanism between every pair of IPONs, and will seek to maintain direct pair-wise virtual connections in the presence of changing network load and changes in topology due to failures and other situations. It will seek to optimise the size of these connections and the routes they take across the optical network, both of which are hidden from the higher level IP network.

The OXCs are not directly addressable by IP nodes outside the optical network. When receiving incoming traffic, the IPON will classify packets based on next IP-level hop destination and queue them subject to any priority mechanism that is appropriate. The OXC that is associated with that IPON provides a virtual interface for each possible destination IPON. These interfaces are used to inject packets into the optical network. At this point statistics such as average used bandwidth and packet loss due to buffer overrun will be collected and made available for analysis. This analysis will be used by the optical network to trigger adding a new virtual connection between this pair of IPONs or removing an underused connection. These virtual interfaces will also contain some distance metric to be used at the IP routeing level. This may take into account the bandwidth available along this virtual route or other factors such as route length, or may be fixed to stabilise external routeing protocols.

Once an OXC that is connected to an IPON is notified of another OXC-IPON pair it attempts to add a route between these two OXCs. Once the route is up it will then advertise this new virtual connection to its IPON. It is assumed that there are out of band control connections between neighbouring OXCs, possibly using a reserved wavelength, where data sent on this wavelength is converted to electrical form and processed by the OXC to run a routeing protocol.

In any OXC-IPON pair we do not assume that that IPON is capable of filling the outgoing optical bandwidth from that OXC or receiving the full incoming optical bandwidth to that OXC; we will provision to add and drop a fraction of the optical throughput of the OXC to the IPON. This will reduce the cost of the network for the same total throughput; however the precise fraction chosen will influence how reconfigurable our network will be. Our proposed solution uses the concept of a *waveband* to define this fraction.

A waveband is a set of wavelengths; each fibre is subject to a two-stage demultiplexing process where it is first split into wavebands, then each waveband may be further demultiplexed into wavelengths. We build an optical *waveband path* from a waveband by switching it through a series of nodes without ever demultiplexing into wavelengths

[3] Section 6.3.2 discusses whether DWDM should be combined with TDM capability

during the path. At each end of the path the waveband is split into wavelengths, and we can provide the facility here either to drop these wavelengths to an IPON, or to combine a wavelength with others to form the start of a new waveband path. In this way we form a *wavelength route*, an all-optical route made from a series of waveband paths which connect two IPONs.

Using a two-stage multiplexing process has some intrinsic advantages when considering optical propagation through multiple network nodes. However, it is also a convenient tool to help describe the degree to which we affect the reconfigurability of the network by removing optical components. For example, one promising strategy identified in this dissertation is to have a fixed set of waveband paths and to place new wavelength routes using spare capacity in these paths. This requires optical switches, transmitters, and receivers only at the end of each waveband path.

It may be advantageous to have the facility to extend these optical paths beyond a single Autonomous System (AS). For example if a large amount of traffic was following the same route through multiple ASs, a single lightpath might be created to travel this shared route. At the edges of one AS the lightpath will be switched to the next AS without transferring it to electrical form. To be practical this would have to rely on several factors. Neighbouring systems would need compatible optical hardware, and some mechanism for exposing enough information about this hardware to set up this path co-operatively. The advantages from co-operating would need to be demonstrated to the administrators of each AS, rather than being an independent decision by a single AS. This is an important area for the future of optical networks, but beyond the scope of this work.

The advantage of maintaining the difference between optical and IP layers is that we are minimising the job of the IP layer. Since each IP-aware node in this network is always connected to every other IP-aware node, we have static connectivity. Over time the data carried between two IPONs will vary, and the exact number and physical path of optical routes will attempt to match this variation. By reducing the number of IP-aware hops that a packet has to travel through across this network to one, the packet is required to be examined, processed, and queued only once.

One disadvantage of this approach is that in the event of a failure where the optical layer cannot recover full-mesh connectivity, some higher-level recovery mechanism would still be required. Also, since we are only processing packets at the edge of the network, we lose the ability to statistically multiplex or selectively drop packets in the centre of the network. There are also scaling issues with current routeing protocols and router implementations with a very large degree of connectivity that would need to be addressed.

## 1.8 Dissertation Outline

Chapter 2 outlines other relevant work done in this field.

Chapter 3 looks at real network traffic in order to determine the appropriate response timescales for an optical network core, and to ascertain how this is related to the granularity at which we allocate bandwidth.

Chapter 4 investigates the architectural possibilities for optical network designs, setting out possible ideas and problems. It also details the assumptions behind the simulator used to test these problems.

Chapter 5 gives the results of simulations designed to guide our design of the optical network layer.

Chapter 6 discusses the implementation of the optical layer, and how this will affect the IP layer protocols.

Finally, Chapter 7 summarises the main arguments of this dissertation and gives possibilities for further work.

# Chapter 2

# Background

In this Chapter we provide context, focusing on the topics covered by this dissertation. In Section 2.1 we look at established networking protocols which form the bulk of the Internet traffic that any new core will have to be capable of supporting.

In order to assess the implications of a new network design we choose to simulate a high level network. This requires a network topology and traffic model; these are investigated in Section 2.2.

Since we intend to use an optical DWDM network core, knowledge of the capabilities of current and future optical networking elements is essential. Section 2.3 lists the primary elements inside an optical network, and assesses their limitations or possible scope for exploitation.

Finally in Section 2.4 we look at other schemes that perform some form of optical circuit switching to carry IP traffic.

## 2.1 Networking

This section covers the basic knowledge of existing network structures and protocols. In designing an optical network it is essential to understand what capabilities it is responsible for, and what capabilities are provided by other network layers. The basic layering model of the Internet is explained in Section 2.1.1. Since our optical network will have to interact with the top half of the network stack, Section 2.1.2 describes in more detail the common protocols currently in use there. The concept of subnets and their interaction with routeing protocols is covered in Section 2.1.3.

### 2.1.1 Network Stack

It is common for communication between two end systems to be broken down into a series of layers. One traditional model is the Open System Interconnection (OSI) model, which defines seven layers. The physical layer is the hardware onto which a bit stream can be sent. The data link layer may provide error control, frame synchronisation and flow control on shared access media. The network layer adds routeing across multiple hosts and networks; the transport layer adds end-to-end error control and congestion control. The session layer handles conversations between end systems. The presentation layer adds the ability to reformat the data between network and application format, for example changing the byte ordering or adding encryption. Finally the application layer refers to the application running on the end system, which may have to deal with user authentication and the syntax of data carried.

Having a well-defined stack is good for designing conceptual systems and allowing independence of implementation similar to object oriented code design. In some cases layers are split, for example the data link layer is often split into the Media Access Control (MAC) layer and the Logical Link Control (LLC). In other cases layers are merged, typically for performance gains. For example the network layer could direct traffic bound for a single destination server to one of a cluster of machines based on the application level characteristics of the data. For IP networks the application, presentation, and session layers are usually merged to form a single application layer.

### 2.1.2 Internet Protocols

The Internet Protocol suite is based around the idea of encapsulation, with higher-level services built on top of more fundamental ones. The base level is IP, which roughly correlates with the network level in the OSI model. This defines the IP datagram, the object that is sent from one host machine to another, which contains a header and payload. The header contains the necessary information for intermediate hosts to be able to forward the packet onto its next destination. The IP datagram service is unreliable, since it may be discarded by intermediate hosts, and connectionless, since each packet is dealt with independently. The Maximum Transmission Unit (MTU) is the largest packet size that can be carried by any of the transmission media used on the path. It is typically 1500 bytes from the predominant use of Ethernet in end-system networks, but may be up to 64 Kbytes.

The payload of an IP datagram contains a packet from a higher layered network protocol, predominantly User Datagram Protocol (UDP), Transmission Control Protocol (TCP), or Internet Control Message Protocol (ICMP). ICMP is used to transmit out-of-band knowledge of network functionality; for example notification of network errors, a warning that an IP packet failed to reach its destination due to too many links having been traversed, or a redirect message to a host informing it of a better next hop destination.

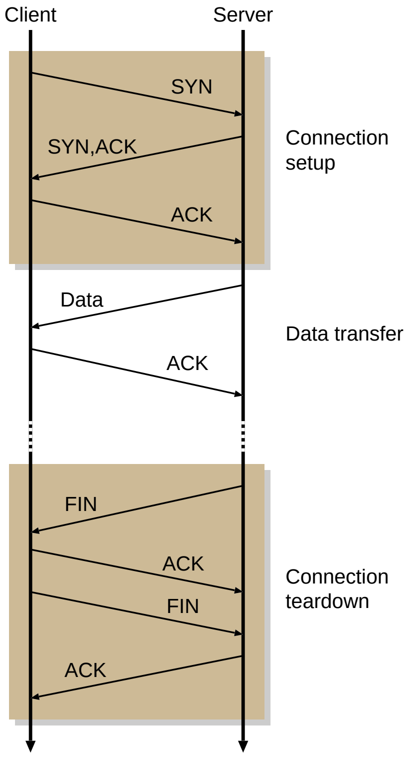

FIGURE 2.1: TCP PACKET EXCHANGE EXAMPLE

UDP adds the ability to identify individual entities within a single host machine by adding port numbers to the addresses of source and destination. It also adds protection against corruption of data through an optional checksum and correct framing of data if lower layers fragment data. However the datagram flow at the UDP level is still connectionless and unreliable. It is typically used for flows of data where retransmission is not needed, for example in a real time stream where the value of data is zero after a given delay between transmission and arrival.

TCP is a connection-orientated protocol supporting duplex communication between two processes on different machines. It is reliable in the sense that it deals with receiving out of order packets and lost packets. It changes its behaviour in response to network conditions with a flow control mechanism based on a sliding window acknowledgement system. Standard implementations have been extended a number of times and much research work has been done to alter subtly its behaviour, especially in its start up behaviour and response to lost or timed out packets. Data carried by TCP accounts for a large proportion of Internet traffic, and under ideal conditions this protocol shares the available bandwidth at a bottleneck point in the network between all competing flows. The simple version of connection setup and teardown is shown in Figure 2.1. Time progresses downwards and packets are shown leaving each host and arriving at the other at a later time. Connection setup consists of the client initiating the connection with a SYN packet.[4] After a positive response with a SYN-ACK

[4] a TCP packet with the SYN flag marked. SYN stands for synchronised, and SYN packets contain the initial sequence number essential for the reliability mechanism

packet, the client replies with an ACK packet,[5] Our simple example shows a single data packet transfer, followed by an acknowledgement of this data; however the data transfer section may transfer data in both directions and connections may be very long lived. Termination is initiated by one host sending a FIN packet. This is acknowledged, and a further FIN packet is sent to confirm that both directions of data flow have terminated, which again is acknowledged.

TCP is used as a transport for many higher-level protocols. For example the Hyper Text Transfer Protocol (HTTP) is carried in the payload of a TCP connection and is commonly used for transferring web pages. Peer-to-peer file-sharing systems which are growing in popularity commonly use TCP for large file transfers.

Network-level routeing protocols can be divided into two groups; inter-Autonomous System (AS) and intra-AS. An AS is typically defined as a network or groups of networks under single administrative control, where full topology is known. For example an AS might be a university, a business, or an Internet Service Provider (ISP). Routeing information is exchanged within an AS using an Interior Gateway Protocol (IGP), of which Open Shortest Path First (OSPF) is an example. It is a link state algorithm, in that each router within the AS advertises its local knowledge of current topology to other routers when that knowledge changes. Each router can then construct independent routeing tables based on given link metrics. It includes facilities for class of service routeing, load balancing, and dividing a single network up into subnets.

An Exterior Gateway Protocol (EGP) is used to route between adjacent Autonomous Systems: most commonly the Border Gateway Protocol version 4 (BGP4). A router belonging to two ASs is called a gateway and these exchange routeing tables containing AS reachability information, sending the affected parts of their routeing table when a change occurs. To allow the best routes to be chosen, routes have metric assigned to them, typically on policy-based management decisions.

### 2.1.3 Routeing Area Hierarchy

The current Internet has independence between different routeing hierarchies. At the top level an EGP, such as BGP, creates routes by sharing knowledge of AS adjacencies. Each packet is matched to a network prefix, which has an AS path saying which is the next hop AS to use for this packet. Details are also forwarded about entry points into each AS for each network prefix.

Within each AS, an IGP is used in a co-ordinated manner to forward packets from the entry point into the AS to the exit point as given by the EGP. The internal network structure of the AS is hidden from other ASs, and only the gateway points which are connected to both ASs are visible.

[5] a TCP packet with the ACK flag marked, used to acknowledge the receipt of data up to a given sequence number

To make best use of the addresses available in a network prefix a similar notion of routeing area independence is achieved using *subnets*. This is where the set of valid addresses in a network prefix is split into subnets.

The routeing within each subnet is logically separate, with specific hosts acting as gateways between each subnet routeing area. This is similar to the independence between neighbouring ASs. It allows an optical subnet to operate inside a larger AS, and have an independent routeing algorithm and transport system based on optical circuit switching.

### 2.1.4 Summary

We have explored the idea of a layered network model, and observed that frequently the layers become merged for convenience or performance reasons. We have briefly described some of the key protocols in use in IP networks that provide packet delivery, routeing, and connection based flow control and congestion avoidance. We have looked at how subnetting allows logical separate routeing and transport systems to be used inside an IP network.

In IP networks the IP layer assumes that the lower level is a passive point-to-point packet transmission mechanism, and builds a multi-hop packet delivery system on top of this. Inside an AS, using a protocol such as OSPF, the IP layer adds fault recovery by sharing reachability information.

Having an optical layer which can reconfigure the topology visible by the IP layer raises the issue of how these two should be combined or separated.

## 2.2 Simulating Networks

Simulating network operation is an essential tool for exploring the design of new network architectures and routeing protocols. It allows us quickly and easily to test various combinations of network design against a repeatable set of events to assess their relative benefits. To ensure that results gained from simulation are accurate, it is important to design the simulator to match closely the environment where the network will be employed. The first consideration for the simulator used in this dissertation is the network topology generation, discussed in Section 2.2.1. The second consideration is the traffic used for the simulation. This is discussed in Section 2.2.2, which sees the progression of model used for traffic generation from the simplest Poisson model to the more realistic self-similar models, based on the observation of real traffic. Section 2.2.3 covers how this type of traffic may be generated, and Section 2.2.4 outlines the idea of non-stationarity, where the distribution of traffic changes over time.

The results of network simulations are often criticised, since it is often hard to select sufficient detail from the situation being simulated whilst keeping the simulation simple enough to be practical. Details that seem insignificant often have critical impact once the implementation is tested for real. The simulation of a large aggregation circuit switched network undertaken in this work is easier than simulating a packet switched network, since we only have to simulate the setting up and tearing down of whole circuits rather than individual packets. To ensure that our results are relevant we need to motivate correctly our choice of topology and the traffic model which drives circuit-level events.

### 2.2.1 Topology Generation

Many graph types are used to simulate and model networks. They come in three main types: randomly created graphs, regular topology graphs, and real life examples.

Using real life network topologies [Zaumen91, Krishmaswamy01, Davis01] has the advantage that results have applicability to the current situation. However this may produce algorithms optimised for that particular network and not transferable to networks of the future, which may be different in structure from current networks.

Regular topologies, such as lines, stars, rings, etc., can be useful for simulation [Noel00] or mathematical [Ramaswami98, Subramaniam99] studies of networks, but real networks tend not to be regular so results may not be wholly relevant.

There is a wide range of random topology generators. A single network can be created using a simple probabilistic model [Waxman88], or a model that generates a power law in the degree of nodes [Barabási99].

The Georgia Tech Internetwork Topology Models (GT-ITM) [Zegura97, Calvert97] creates hierarchical networks using a top-down recursive procedure, replacing some nodes in one level of network by a new network, which is scaled down in size to fit in the space taken up by the original node. Each network is created by adding links between randomly placed nodes using a probability distribution based on factors such as distance between the nodes. One drawback with this approach is there is nothing to guarantee that the networks created are entirely linked together. Achieving a network with a given number of nodes would either mean repeatedly generating networks until one was connected, or adding a new stage to the network generation algorithm.

Tiers [Doar96] is an alternative approach to the idea of multi-hierarchical networks seen above. It generates individual networks based on a minimum spanning tree across randomly placed nodes, adding links to obtain a specified level of redundancy. It then can interconnect these networks using a three level hierarchy of types of networks: Local Area Network (LAN), Metropolitan Area Network (MAN), and Wide Area Network (WAN).

BRITE [Medina01] integrates representative models of Internet topology, measurements of real networks, and a flexible approach to new models into a two level hierarchical network design tool.

The exact degree to which AS-level network topology displays a power-law has been questioned, when more sources for detecting AS links are used [Chen02]. However this work does agree that for the connectedness of the largest ASs the power-law does seem to fit. Generators using power-law connectivity algorithms tend to match both AS and router-level measured topologies more accurately than ones using a rigid hierarchical design [Tanmunarunkit02]. These models are primarily for generating large multiple-AS topologies of more than 1000 nodes: how to model a single AS with fewer nodes is less clear.

### 2.2.2 Traffic Modelling

To assess the performance of a system by simulating traffic flows, an accurate traffic model is essential. Given some timescale between measurements we wish to be able to say that the number of bytes or packets arriving in each time interval is $X_i$.

Pioneering work on traffic characterisation was done on early telephone networks; it introduced the Poisson model where call inter-arrival times are exponentially distributed, and calls last for an exponential length of time [Erlang09]. These models have useful analytical properties: for example, the aggregation of multiple Poisson process is also Poisson. It is also known that the multiplex of a large set of processes converges to a Poisson distribution, where the individual processes are not necessarily Poisson [Palm38].

To model data traffic, where we either do not have the necessary aggregation or the individual packet-arrival processes do not aggregate to a Poisson process, other models have been developed [Frost94]. These include Markov and Markov-modulated processes [Heffes86], where we modulate a Poisson process with each Markovian state having a different Poisson rate parameter. An alternative method for capturing bursty traffic is the set of autoregressive functions. The simplest example is the linear autoregressive model $AR(p)$, where each new sample $X_i$ is a linear combination of the previous $p$ samples with random noise added. A more complex example is the Autoregressive Integrating Moving Average (ARIMA), which adds error terms and moving averages [Box76].

Observations of several real systems have found that the sequence $X$ has self-similar and long range dependent characteristics, both for Ethernet traffic [Leland94] and WAN traffic [Paxson95]. The traffic models above do not capture this feature. This is significant since adding long range dependence changes the tail of the distribution, possibly making previously rare events such as buffer overruns in a queueing system much more likely. A comprehensive bibliography of the area of self-similar traffic can be found in [Willinger96].

Long range dependence is defined to be where the autocorrelation function of a sequence decreases slower than exponential. If $X_k^{(m)} = (X_{km-m+1} + \cdots + X_{km})/m$ then an exactly self-similar traffic is defined to be where the process $X$ for all $m = 1, 2, \cdots$, $VAR(X^{(m)}) = \sigma^2 m^{2H-2}$ and

$$r^{(m)}(k) = r(k) = \frac{1}{2}\left((k+1)^{2H} - 2k^{2H} + |k-1|^{2H}\right), k \geq 0 \qquad (2.1)$$

where $r^{(m)}$ is the autocorrelation function of $X^{(m)}$. Exact self-similar traces are very rare, so the normal definition is of asymptotically, or second order, self-similarity. Second order self-similar traffic can be defined to be where for large $k$

$$r^{(m)}(k) \rightarrow r(k), \quad \text{as} \quad m \rightarrow \infty \qquad (2.2)$$

Intuitively this means that there are bursts on all timescales; 'traffic "spikes" ride on long term "ripples," that in turn ride on still longer term "swells,"' [Leland94]. A consequence of self-similarity is that multiple independent self-similar streams will aggregate together and preserve the self-similarity, preserving bursts at all timescales. This differs from short range dependent traffic, such as that generated from a Poisson model, which when aggregated becomes less bursty. For the core of the network, with a very high level of aggregation, the effect of this is that since self-similarity implies long range dependence, traffic levels will still have the same long range dependence characteristics.

A feature of self-similar processes is that the degree of self-similarity can be defined using a single parameter; the Hurst parameter $H$ [Hurst51, Taqqu85]. From Equation 2.1, $H = 0.5$ describes a short range dependent process. As $H$ increases to $1$ the degree of long range dependence and self-similarity increases. There are several well known methods for estimating the Hurst parameter [Beran94, Rose96, Veitch99].

It has been suggested that network traffic contains multiple components, the majority of traffic being from a Gaussian self-similar source — the mice — and the remainder from a very bursty set of high bandwidth connections — the elephants [Sarvotham01a]. Each high bandwidth connection is the coincidence of a large file size transfer and a fast network link, causing it to dominate all other traffic at that time. The degree by which total traffic has a marginal Gaussian distribution increases as the level of aggregation increases [Paxson97b], and also depends on the horizontal aggregation present in the measurements [Kilpi02]. For the purpose of traffic engineering an algorithm has been developed to classify traffic originating from given network prefixes as either mice or elephants [Papagiannaki02]. This algorithm attempts to lengthen the time that any given network prefix retains the same classification.

### 2.2.3 Self-similar Traffic Generation

Schemes exist to generate traffic that is comparable to that observed in networks, in that they exhibit self-similarity. For instance, the multiplex of many traffic sources from a Pareto generator [Willinger95] or the Random Midpoint Displacement method, which progressively subdivides intervals and sets the sample at the middle of the interval from an appropriate distribution [Lau95], exhibit self-similarity. The ARIMA process can be extended to Fractional ARIMA [Hosking81], and methods using transforms of Fractional Gaussian Noise [Paxson97b], or Fractional Brownian Motion and extensions [Véhel97] can also be used. Fraction Gaussian Noise, although it requires a Gaussian marginal distribution, is possibly the most practical method since it is entirely described by a single parameter, $H$, and fast methods exist to general approximate sample traces [Paxson97b].

### 2.2.4 Non-stationarity in Network Traffic

The other property commonly measured in networks is non-stationarity [Thompson97]. Non-stationarity occurs when the distribution from which samples are drawn for the time process $X_i$ changes with $i$, for example the mean or standard deviation changes. The major example of this is the daily cycle when traffic is measured from one company or institution. The difference in the mean traffic level between the peak hour and the slowest hour is likely to be greater than can be accounted for by a single stationary self-similar process. Other Internet variables, such as packet inter-arrival time, have also been observed to show non-stationarity, with the distribution changing with network load [Cao01b].

User-driven non-stationary events can be divided into two classes; either repeatable cycles of activity such as daily or weekly cycles, or one-off *flash crowd* events [Barford01, Barford02, Jung02]. These include large software releases, news events, and websites supporting live events. For the end-systems involved these can be catastrophic events leading to unavailability, and may lead to measurable effects in the core network depending on how spatially and temporally concentrated the traffic is. For core networks, events such as equipment failures leading to re-routeing are likely to have a significant effect.

Determining the difference between long range dependence and non-stationarity is difficult. With long range dependence we may get long periods of time where the mean over that period is different from the mean of the whole distribution. However this may also arise from a short range dependent distribution whose mean has a piece-wise linear function — it exhibits non-stationarity. An extension of the variance time estimator for the Hurst parameter can help tell the difference between these two cases [Teverovsky97]. If the distribution is known to be long range dependent, an alternative is to use an estimator that yields unbiased estimates under some non-stationary events [Roughan99].

For a network designer the critical aspect is not the mathematical basis of the underlying traffic, but the consequences of that traffic on the design of the network. It is clear that traffic shifts over relatively large timescales do occur and the network has to cope with these traffic shifts, whether or not they are generated by a stationary long-range dependent process or non-stationarity. Although having precise mathematical models may add credence to network simulations, it is not obvious how accurate these models have to be to motivate correct network design. It has been shown that for systems such as a finite buffer queue that periodically lose history information when the buffer overflows, the long-range dependent structure is ignored beyond a correlation horizon [Grossglauser99]. This work also shows the importance of the marginal distribution over the exact degree of self-similarity as measured by the Hurst parameter. A traffic model must therefore have the correct marginal distribution and the correct degree of long-range dependence over the critical timescales that exist in the particular system being simulated.

### 2.2.5 Summary

In this section we have looked at some of the considerable body of work in designing simulations. We have seen that network topologies can be generated to be representative of current multi-hierarchical networks. Since we are concentrating on the operation of an optical network within a single administrative boundary, we are interested in single hierarchy networks, typically generated using spanning trees with redundancy.

We surveyed the history of traffic generation, through to the identification of self-similar characteristics in observed network traffic. This is embodied by the notion of bursts on many timescales, and will have consequences for the timescale over which we reconfigure our network. We can generate self-similar traffic using published algorithms. We have noted that observed network traffic is non-stationary, where the parameters of the traffic distribution change over time. This will affect our analysis of observed traffic and also require collecting simulation results under a range of traffic conditions.

## 2.3 Optical Technology

Any optical network design has to be not only flexible enough to cope with the evolution of optical network technology, but also able to exploit and be constrained by the current and likely development in the field. A comprehensive description of optical network technology can be found in [Ramaswami02], and details of optical network components can also be found in [Borella97].

One goal of understanding the abilities and limitations of optical hardware is to be able to make a realistic set of conditions for the Routeing and Wavelength Assignment

(RWA) problem. In a generic optical network we have optical nodes, containing traffic sources and sinks, and some switching capability. Optical fibres connect some of these nodes, and have sets of wavelengths carried on them. Fibres are typically homogeneous; they all carry the same set of wavelengths. We have some traffic demand, where we wish to form optical paths from a source at one node, through fibres and switches at intermediate nodes to a sink at the destination node. The solution to the RWA problem shows the route each lightpath takes through the network and which wavelength is used on each fibre along that path. The specific restrictions placed on this problem may vary greatly, depending on the capabilities of the optical hardware to be used. For example, if we assume full wavelength conversion, the RWA problem degenerates into a simpler routeing capacity problem.

The RWA problem is known to be NP-complete [Even76, Chlamtac92]. Heuristic solutions have been proposed to solve this either as separate routeing and wavelength assignments [Bala95a, Ramamurthy03] or as a single procedure to perform both routeing and assignment [Zhang95, Mokhtar98, Cinkler00].

Formulations of the RWA problem into an Integer Linear Program (ILP) have been constructed [Ramaswami95, Tornatore02] which give optimal solutions. These typically are too computationally expensive for large networks, but can give optimal bounds useful for benchmarking heuristic solutions. An approximate polynomial-time algorithm has also been developed to provision networks [Hauser02].

Historically routeing protocols tend to be simpler in design than ILP or approximate polynomial algorithms, using heuristics and only locally optimal solutions. This makes them easier to design and formally verify correctness of implementations, and easier to distribute between large numbers of processing elements.

### 2.3.1 Link Capability

The propagation of light through a fibre optical cable is complex, but in outline light waves travel along the core of the cable, reflecting off the boundary between the core and the cladding. The cladding is very similar to the core, but designed to have a higher refractive index, which allows the light waves to be guided through the core. Early multimode fibre had a relatively thick core, 50–100 $\mu$m, which allows several modes of propagation, each mode referring to how frequently the light wave reflects off each edge, which is ultimately determined by the angle of entry into the core.

The major limitation with this approach is that different modes propagate at different speeds leading to intermodal dispersion. While using a graded index fibre, where the refractive index of the fibre changes continuously from the centre to the cladding, can decrease this dispersion, when transmitting over a distance of a few kilometers practical systems are limited to a bit rate of under 20 Mbit/s.[6]

[6] Optical systems are typically limited by the product of the bit rate and transmission distance: multi-

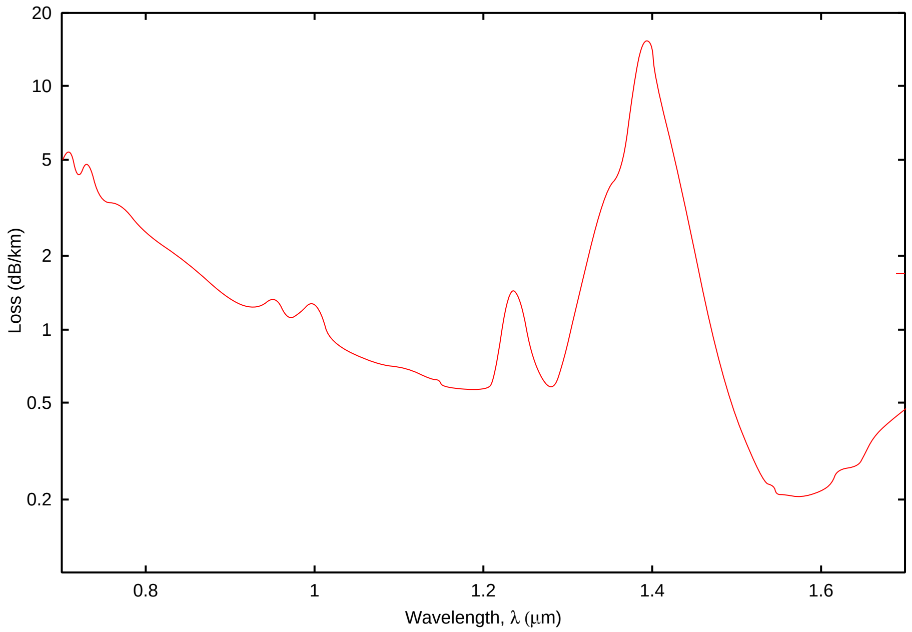

FIGURE 2.2: LOSS PER KM OF FIBRE AS A FUNCTION OF WAVELENGTH OF LIGHT USED

Single mode fibre removes intermodal dispersion by allowing only one propagation mode since the core width is around 8–10 $\mu$m, a small multiple of the wavelength of the signal.

The two major causes of signal loss through fibre are material absorption and Rayleigh scattering. The effects of these can be seen in Figure 2.2. The local minima correspond to different frequency bands used in optical networks. The peaks between bands are mostly caused by absorption by water vapour in the fibre; this has been reduced over the last few years by the use of newer types of commercial fibre. The minimum point is around 0.25 dB/km of loss, which enables a distance travelled of around 100 km before the signal to noise ratio drops too low.

Other factors affect the transmission of data through single mode fibre. Since fibres are typically not entirely cylindrical, they are slightly birefringent; the propagation speed depends on the polarisation of the wave. Since a lightwave consists of two orthogonally polarised modes, these will propagate at different speeds leading to Polarisation-Mode Dispersion (PMD). Chromatic dispersion occurs since the propagation time is frequency dependent, and some optical pulses are chirped; the exact frequency changes slightly with time. This means that pulses can broaden, shorten, and even be reversed over time. These effects can be controlled and exploited by dispersion compensating fibre, which changes the refractive index of the fibre to create enough waveguide dispersion to compensate for chromatic dispersion at a given wavelength.

Non-linear effects occur at higher bit-rates. Scattering occurs as energy is transferred from one wavelength to another longer wavelength; Stimulated Raman Scattering is

mode fibre can support higher bit rates, but only over shorter distances.

an instance of this. Whilst for a DWDM system power between different wavelengths might be affected, with shorter wavelengths losing power, the effect can also be used for amplification purposes. Other non-linear effects include self-phase modulation, where high intensity pulses can become chirped since the refractive index of fibre has an intensity dependent component. For close channel spacing or a system using dispersion compensated fibres, four wave mixing can create new signals at related frequencies to existing signals.

With correct choice of the properties of the fibre used, the current available long haul DWDM systems typically use channel spacing of 50 GHz and up to 128 wavelengths at 10 Gbit/s spanning 4,000 km before full regeneration is required. A full discussion of light propagation in optical fibres is covered in [Ramaswami02, Chapter 2].

### 2.3.2 Amplifier Capability

To achieve the quoted spans between full regeneration when the signal loss is around 0.25 dB/km, optical amplification is necessary. This enables amplifying the optical signal without having to convert it to electrical form, preferably over a wide range of frequencies used by a DWDM system, and with a high-output flat gain spectrum; we wish to avoid amplifying some signals more than others. Two main technologies are used for optical amplification: Erbium-Doped Fibre Amplifiers (EDFA); and Raman amplifiers.

EDFAs exploit the fact that the gap between two energy levels for erbium ions corresponds to a wavelength range of 1525 nm to 1570 nm, coinciding with the minimum point of attenuation in a fibre. The optical source is combined with a pump laser. This is set at the correct wavelength for the reverse transition between levels, either 980 nm or 1480 nm, and the stimulated and spontaneous emissions of photons will occur at the signal wavelength range. The stimulated emission accounts for the signal amplification; reflections need to be prevented in order to stop this becoming a laser source, while spontaneous emissions account for some of the noise introduced by the amplifier. The natural gain spectrum for EDFAs is not flat: there is a peak in amplification at around 1532 nm. This can be countered by adding filters [Giles90, Toba93].

Raman amplifiers exploit the non-linear effect of Stimulated Raman Scattering. Here a pump laser provides amplification to signals around 100 nm above the laser wavelength. This allows other regions of wavelengths to be used for optical systems. Raman amplifiers are typically used for long haul systems, and they benefit from the pump laser propagating backwards along the signal. The gain is strongly linked to the pump laser power, but backwards propagation decreases the effect of a varying pump power on the total signal amplification. This also decreases the effect of crosstalk between different DWDM signal wavelengths caused by the depletion of the pump power from amplifying one wavelength affecting other wavelengths.

Placement of optical amplifiers in a network has been carried out for passive star networks, where each node in the tree-like architecture is connected to an optical star that forwards all incoming wavelengths onto all outputs apart from where the signal originated. Each node has a fixed transmitter and tuneable receiver; these architectures could be used as LANs or small MANs. One simplifying constraint is that all wavelengths should be equally powered. This is a result of the near-far effect: two wavelengths originating different distances away will have attenuated different amounts resulting in a decrease in the gain an amplifier can deliver, limited to reduce the saturation of the higher powered signal. Constraining all wavelengths to be equally powered at each node leads to a simplified problem [Li94, Ramamurthy98a]. However for small and medium sized networks allowing wavelengths to be unequally powered reduces the number of amplifiers needed [Ramamurthy98b].

One limitation with optical networks is the speed at which new lightpaths can be added to an existing optical network. Current practice is gradually to increase the power of a new wavelength whilst checking for crosstalk and other effects on other pre-existing lightpaths traversing the same optical components. This is done because each amplifier that the connection uses has a separate dynamic control loop to equalise the gain for that amplifier: these could interact to give a longer period of instability. For small numbers of amplifiers the period of signal degradation following a sudden change in input power is of the order of milliseconds [Madamopoulos02]. However is it insightful that current operating practice for adding a new wavelength route to an existing network requires several minutes to bring the power up to usable levels [McAuley03].

### 2.3.3 Multiplexing

Multiplexing and demultiplexing optical signals is an essential component in WDM networks; optical signals need to be split and joined depending on the frequency of the signal. There are two main types of demultiplexer, active and passive. Passive types, such as stimax gratings [Laude84] or arrayed waveguide gratings [Vellekoop91], can typically be also used as multiplexers. Acoustically tuneable filters [Smith90], an example of an active demultiplexer, have the added ability to dynamically select multiple wavelengths but perform in an inferior manner to passive components on the two main goals: to minimise loss of the pass-band frequencies and to minimise signal from rejected bands. Other goals include thermal stability and flat pass-band stability to cope with slight changes in actual wavelength frequency.

To cope with demultiplexing a larger number of channels than a single component can manage (currently around 40 on commercially available products), a multistage approach is needed. Here demultiplexers can be cascaded using either a banded or an interleaved approach.

Banded multistage multiplexing splits the set of incoming frequencies into disjoint adjacent groups — wavebands — as shown in Figure 2.3. These wavebands are separated in the first stage of demultiplexing; each band is then further demultiplexed into wave-

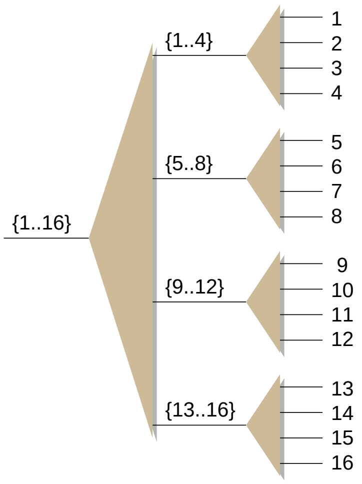

FIGURE 2.3: EXAMPLE OF A MULTI-STAGE DEMULTIPLEXER USING WAVEBANDS, WITH 4 BANDS OF 4 WAVELENGTHS EACH

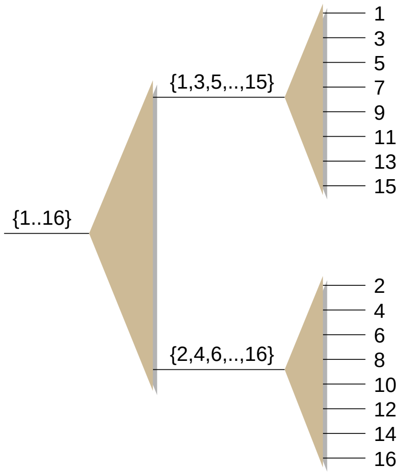

FIGURE 2.4: EXAMPLE OF AN INTERLEAVED MULTI-STAGE DEMULTIPLEXER

lengths during the second stage. Often the wavebands are separated by an unused wavelength frequency to reduce the crosstalk between bands. The use of wavebands outside multiplexing has been investigated; for example switching wavebands at each optical switch rather than switching wavelengths [Bala95b], or limiting the tuneability of receivers [Labourdette97] such that each receiver can tune to any wavelength in a given band. Routeing wavebands rather than wavelengths requires fewer routeing elements given the same total bandwidth, and hence a reduction in signalling used in the network [Austin01]. A summary of potential advantages to both the hardware level and network design can be found in [Gerstel00c]. Using a mix of waveband and wavelength routeing, an integer linear program has been formulated [Lee02]; although too complex for real networks it leads to an heuristic algorithm. This uses shortest path routeing, and groups routes into wavebands based on their destination node before assigning them to wavelengths. This method is applicable only for provisioning fixed traffic demands, and is evaluated only in terms of the reduction in switch size.

Interleaving [Chiba01] splits an incoming fibre into two, with odd numbered wavelengths on one output and even numbered wavelengths on the other. An example of this is shown in Figure 2.4. This makes the second stage easier since the channel spacing has doubled. In the remainder of this dissertation we will assume that we are using some form of banded approach, although the same effect might be achieved with some form of interleaved multiplexer.

### 2.3.4 Switch Capabilities

Currently deployed optical switches are essentially large electrical switches where incoming light signals are first decoded by an array of receivers, then switched across the electrical backplane. Outgoing data is then converted back to an optical signal before being transferred through the network. Whilst this is a flexible approach it is limited by the speed of the backplane, the large number of receivers and transmitters required, and most critically by the power consumed and the problems in dissipating the resulting heat. To differentiate this form of optical switching from one using an optical backplane, where signals remain in optical form throughout the switch, these newer switches are referred to as *all-optical* or *photonic* switches. In the remainder of this dissertation unless otherwise stated, an optical switch refers to a photonic switch with an optical backplane.

Photonic switches switch optical signals without converting them to the electrical domain. They have several characteristics, depending on the use to be made of them in the network. Switches used for packet switched networks obviously need a faster switching time than those used for providing circuit switched networks; from the order of nanoseconds up to tens of milliseconds.[7] Other characteristics are *crosstalk*; signal interference between logically independent paths through the switch, *insertion loss*; the power lost from signals travelling through the switch, and *extinction ratio*; the reduction in signal strength when on a single path the switch moves from a pass state to a blocking state.

Another important characteristic of switches is their blocking characteristic. A blocking switch can be in a state where an unused input port cannot be connected to an unused output port. Switches can be nonblocking and are divided into categories of necessary assumptions as to how they achieve this. Strict-sense nonblocking switches can link any two unused ports regardless of previous operations. Wide-sense nonblocking switches algorithmically find ways to connect ports such that any future connection can be realised. Rearrangeable nonblocking switches may require that existing connections be rearranged in order to connect a new pair of ports.

For medium scale switches, a number of 2x2 switches can be connected together in one

[7] Here we give examples of splicing speed, the time between adjacent packets or circuits. 1 ns at 10 Gbit/s corresponds to 10 bits, while recent optical circuit switches advertise switching speeds to tens of milliseconds.

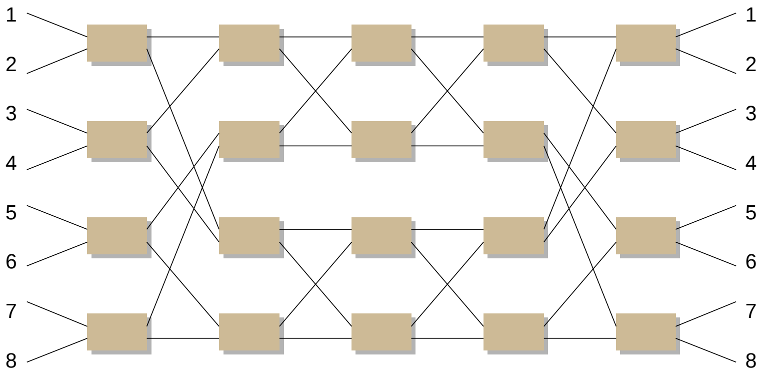

FIGURE 2.5: EXAMPLE OF AN 8X8 SWITCH CONSTRUCTED FROM 20 2X2 SWITCHES, USING THE BENEŠ ARCHITECTURE

of several configurations. These configurations determine the number of 2x2 switches required for an NxN switch, the nonblocking sense of the switch, and the difference between the minimum and maximum path length through the switch. This path length determines the signal loss of the switch. An example of the Beneš [Beneš65] architecture is shown in Figure 2.5. This is a rearrangeably nonblocking architecture that uses $\frac{n}{2}(2log_2(n) - 1)$ 2x2 switches for a $n$x$n$ switch, one of the most efficient architectures.

In the regime of large numbers of fibres and channels, the logical size of the switch is critical. Large switches, containing thousands of input and output ports, are created either by Micro-Electro-Mechanical System (MEMS) switches or through combining smaller switches in some configuration.

The MEMS devices used in optical switches are arrays of tiny mirrors fabricated in silicon that are controlled to steer lightpaths around the switch [Neukermans01]. They can be built either in a flat configuration with one fixed axis, or with two variable axes to steer light through 3d space. These devices are currently capable of switching hundreds of ports, are nonblocking, and some have switching speeds of around 10 ms [Calient02, Aksyuk02]. However, large MEMS switches are very complex devices, requiring a large amount of electronic control circuitry, and have yet to be demonstrated in a large scale deployment environment.

### 2.3.5 Wavelength Conversion

One constraint on creating optical paths through a network is the wavelength continuity constraint. This states that the same wavelength has to be used on each link in the path, and is an essential consequence of using photonic switches. This constraint might result in a routeing failure when there is spare capacity on all links along the path, but not for the same wavelength. To avoid this constraint, wavelength conversion may be used. There are two fundamental ways of achieving this: opto-electrical and photonic.

Opto-electrical conversion changes the incoming signal to electrical form and then retransmits it. This may involve several stages, which are called 1R, 2R, and 3R. 1R regeneration detects the incoming analogue signal and retransmits; it is entirely modulation format transparent. 2R regeneration adds reshaping; here we detect the digital pulses and retransmit those. 3R adds retiming onto 2R; we have to know the bit rate to be able to retime the digital signal and use either a global clock or clock recovery scheme.

Photonic conversion [Elmirghani00] relies on physical properties: cross-gain modulation using optical gratings, cross-phase modulation using interferometers such as Mach-Zehnder Interferometers, or four wave mixing. By combining some effects, 3R regeneration can be gained [Chiaroni97]. However, photonic conversion is an immature technology, and it is not clear how long it will take for it to become more cost effective than opto-electrical conversion.

When designing an optical network node, the amount of conversion can be varied. No conversion implies that traffic entering a node on a wavelength will leave that node on the same wavelength; this is known as the wavelength continuity constraint. Full conversion implies that traffic entering a node can leave on any wavelength regardless of input wavelength. Between these two extremes we have limited conversion of various kinds. Some predetermined wavelengths might have full conversion, whilst the remainder have none, or some wavelengths might have the ability to be converted to a small set of target wavelengths, or any combination of the two.

The benefits to network design from allowing conversion have been established, with figures of 40% improvement on wavelength re-use [Ramaswami95] and a reduction of 50% in the required fibre length capacity required [Strand01]. However, allowing unlimited conversion at every network node would result in this performance improvement being achieved at a large cost.

Research has led to the use of sparse conversion, firstly with a scheme where nodes capable of complete wavelength conversion are spread equally [Subramaniam96], which finds that the benefits from equipping more nodes with conversion suffer from diminishing returns. Analytical models [Subramaniam99] and simulation studies [Xiao99] have both shown that a more optimal approach to placing converters gains more benefits than random placement, to the stage where the difference between full conversion and sparse conversion is minimal.

A further area of research has been into the use of limited conversion, typically motivated by the possible use of four wave mixing techniques to realise an all-optical converter. Using limited conversion can be shown to achieve performance close to full conversion [Sabella98, Sharma00]. For offline analysis of limited topologies, performance can be identical [Ramaswami98], and limited conversion decreases the worst-case scenario of the number of required wavelengths to entirely satisfy a set of traffic demands [Gerstel99].

With networks increasing in scale, multifibre networks are becoming more common, where between two nodes there are multiple fibres, and therefore multiple sets of identical wavelengths. These are equivalent to a single fibre network with limited conversion capabilities [Ramaswami02, Section 8.2.3], perhaps explaining the observed phenomena that the gain from allowing conversion decreases exponentially as the number of fibres increases for multifibre networks [Jeong96, Karasan98].

### 2.3.6 Protection of Lightpaths

Given that an optical network is complex and relies on many components, failures can happen in a number of ways. Some failures would render a lightpath or set of lightpaths completely unusable: a fibre cut for example. Some will degrade a lightpath so the bit error rate increases to an unacceptable level; bad thermal stability of lasers or multiplexing equipment might cause this. Finally, some failures might leave lightpaths operable, but stop reconfiguration: for example the failure of switch control hardware.

Analysis of the duration of link failures can shed light on the relative occurrence of different types of failure [Iannaccone03]. This analysis from a Tier-1 IP backbone network found that 10% of failures last longer than 45 minutes; fibre cuts and other major failures that required human intervention are certain to be in this category. 4% took between 15 and 45 minutes and probably required substitution or maintenance of hardware equipment. The remainder, 86%, took less than 15 minutes, with 46% of failures lasting for less than a minute; these were caused by resetting line cards and routers, congestion causing false alerts, and temporary failures of the transmission layer. The conclusion from this work is that network failures were at least a daily event, and a large proportion are caused by resetting optical and IP hardware devices.

There exists a range of mechanisms to deal with recovery from a failure. They differ primarily in the speed of recovery from a failure, the range of failures covered, and the excess capacity needed to recover. At one extreme is the Synchronous Optical Network (SONET), an OSI layer 1 standard for multiplexing a wide range of transmission rates of digital data onto an optical transmission system. It has a large Operation and Maintenance (OAM) overhead, allowing fine control over provisioning traffic and comprehensive error detection and reporting. For a wide area network it can be used as a point-to-point link, or more normally as a ring network utilising automatic protection. Here, fibres are allocated in primary and protection pairs. When traffic on the primary fibre fails, it is automatically sent on the protection fibre which circles the ring network in the opposite direction.

For this ring-based protection, data capacity is decreased by a factor of two, but the recovery speed is designed to be less than 50 ms. At the other extreme, we may have optical fibres being used as a transport network for an IP network. Here, no reserve capacity is needed, but there is no restoration of connectivity, and higher level mechanisms are required to restore paths across the network; these typically take seconds or minutes to operate.

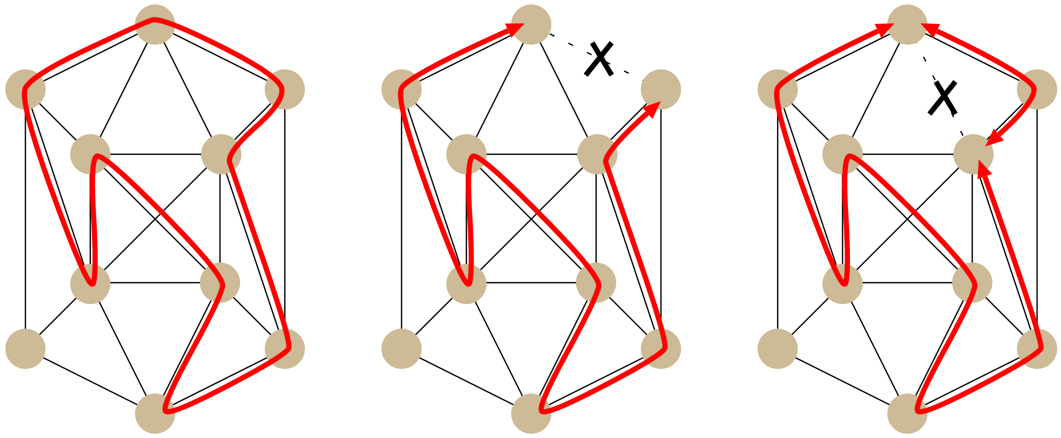

FIGURE 2.6: AN EXAMPLE OF A P-CYCLE IS SHOWN IN THE LEFT FIGURE. THE CENTRE FIGURE SHOWS THE FAILURE OF A LINK ON THE P-CYCLE, USING THE REMAINDER OF THE CYCLE FOR PROTECTION BANDWIDTH. THE RIGHT FIGURE SHOWS THE TWO PROTECTION PATHS USED IN THE EVENT OF A FAILURE OF A SPANNING LINK

One area of study has been the interoperability of any protection system in the optical layer with other network layers. Clearly different network layers have detailed and accurate knowledge about any failures in their layer, so are best placed to respond in a timely fashion with appropriate protection mechanisms. However, this can lead to duplication of effort, with multiple layers responding on the same failure stimulus. Different optical layer protection schemes may have various advantages over higher network layers [Gerstel00a], and the correct balance between optical layer protection and higher layer protection requires co-ordination and depends on the exact network parameters [Colle02]. Problems occur when protection issues are not co-ordinated in any way at the time networks are provisioned, since a single failure at the optical layer might lead to multiple unrecoverable failures at higher layers [Crochat00] depending on the mapping from virtual optical topology to IP topology.

Assuming some form of protection is necessary in the optical layer, there is a variety of schemes available, each requiring different types of lower level hardware to implement [Gerstel00b]. There are two main forms of optical protection system for optical mesh networks: ring or cycle based, and path based [Baroni00]. Ring based systems are an adaptation of SONET ring network protection, typically requiring dedicated protection bandwidth but yielding fast reconnection times. Path based systems are more connection orientated and can be more bandwidth efficient, but may disrupt traffic for longer. Typically systems in which a joint optimisation of primary and protection bandwidth is performed, with protection carried out on an end-to-end basis through the network, are the most efficient in terms of spare capacity requirement for circuit switched networks [Xiong99].

An extension of ring protected mesh networks uses precomputed cycles ('p-cycles') through the network [Grover98]. These p-cycles can protect against link failure not only on the cycle, but also links not in the cycle but spanning two nodes in the cycle,

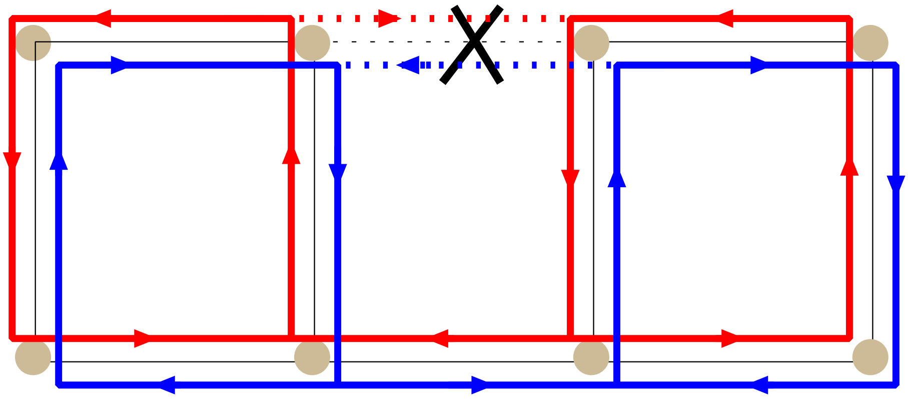

FIGURE 2.7: A LOOP-BACK PROTECTION EXAMPLE, WITH PRIMARY DIGRAPH MARKED IN BLUE AND THE CONJUGATE PROTECTION DIGRAPH SHOWN IN RED

as shown in Figure 2.6. This leads to a reduction in the protection capacity needed, although it relies on full conversion being present. This reduction in capacity is theoretically as good as can be achieved using a path based system [Stamatelakis00].

An alternative extension to this form of protection that operates at ring-like speeds over a general mesh network is generalised loop-back [Médard02]. Here digraphs are formed with the primary and protection digraphs being conjugates of each other, shown in Figure 2.7 with the primary digraph in blue and the protection digraph in red. In the event of a failure, the node upstream of the failure can flood traffic in the protection digraph: a negative acknowledgement system accepts the first path found to the intended destination and quenches other paths. Generalised loop-back is intended to work with two sets of primary and protection digraphs, where the two primary digraphs are conjugates of each other. This allows routeing to behave independently from the protection algorithm, and each link is assigned to a primary digraph on a link by link basis.

Path based protection systems typically fall into two categories. The first requires that dedicated protection bandwidth is allocated for each primary path. Traffic can be dual-fed through each path, and the destination can select traffic streams based on quality. The majority of systems use the second approach, which exploits capacity savings by multiplexing protection paths. This is allowed where two primary paths are disjoint: since a single failure cannot disrupt both, their respective protection paths may share bandwidth. This reduction in capacity required is at the cost of lower performance in the event of multiple failures, and extra latency in activating the backup path since intermediate nodes need to configure the path.

An integer linear program has been formulated to optimise the capacity saving from shared path protection, with a corresponding linear program relaxation and heuristic method to form a valid solution [Sridharan02]. For medium sized networks this may be fast enough to perform in a centralised fashion. Considering a restricted sharing scheme, where disjoint end-to-end routes routes can be shared, an approximate polynomial-time algorithm can be used provision networks [Hauser02]. Other constraints may exist for protection paths; fibres that are link disjoint may be carried in the same physical duct and so fall into the same Shared Risk Protection Group (SRPG). A heuristic has been developed [Zang03] which, although can fail depending on the exact duct and link topology, typically gets close results to the proposed Integer Linear Program formulation. However the progression from this heuristic algorithm for provisioning networks to dealing with dynamic traffic is not explored in their work and is unclear.

Distributed algorithms for path-based shared protection schemes have also been constructed. One option is for nodes not to share any state but to use a distributed search algorithm to find protection paths [Doshi99]. Using distributed signalling this achieves a sub-second time from failure to activating protection. An alternative approach is to share network state amongst nodes, allowing precomputed disjoint primary and protection paths to be allocated on call admission [Austin01]. This approach also incorporates bundling together demands that share path end points, leading to a reduction in the complexity of the network.

An extension to the reduction of required protection bandwidth can be achieved by compromising the requirement that 100% of primary paths must survive a single node or link failure [Mohan01]. Here we allow a primary path to share bandwidth with other protection paths. Assuming that the mean lifetime of a path is less than the mean time between failures, the majority of paths will be unaffected by this.

One area where mesh path protection decreases in efficiency is that of sparse networks. Here there are fewer alternative paths and less chance for sharing protection bandwidth. Advances can be made by reducing the network to those nodes with a degree of three or more which operate full mesh protection, whilst special cases are made for chains of nodes of degree two [Grover02].

### 2.3.7 Relative Cost of Components

When designing a new network we are primarily interested in the performance in relation to the real cost of the network: to be able to compare fairly the performance of two networks they must be equal in price. This is especially difficult when designing networks using components that are currently in research and development, since the effects of mass production or other technological breakthroughs will have a large effect on prices. When considering current components, manufacturers tend to be reticent about releasing information.

It is clear that a price model seeking to cost every component accurately is impossible, so the approach used in this dissertation is to simplify the model of equipment used to its most significant items. This reduces the scale of the problem to a manageable one, since we can run a range of simulations altering the relative prices for each component to get an idea of how stable our results are in the face of changing or inaccurate price information.

The cost model used in this dissertation is covered in Section 4.3.3.

### 2.3.8 Summary

In this section we examined the enabling optical technology that combines to make an optical network: using the low attenuation region of optical fibres, amplified by all-optical amplifiers, together with the potential to demultiplex and switch entirely within the optical domain.

We have discussed the possibility of a multi-stage multiplexing solution to problems caused by the increasing number of wavelengths present in a single fibre. One approach is to use wavebands: contiguous sets of wavelengths that can be treated as a single unit. This allows the possibility of a multi-stage switch, where some signals are optically switched as a whole fibre, some switched as a waveband, and the remainder switched as individual wavelengths. This has the capability of reducing the cost of the optical components necessary to support an equal quantity of bandwidth, over the normal approach of splitting all fibres into wavelengths.

Wavelength conversion has widely been regarded as essential for increasing the utilisation of optical networks; however this can easily be done only by converting optical signals to electrical form and back, since all-optical methods are limited. As optical networks grow in size, supporting multiple fibres on a single physical link, the benefits from explicit conversion decrease due to the effective conversion gained from the multiple copies of each single wavelength.

One advantage of an optical network could be the possibility of fast protection switching. Since we can detect loss of signal quickly, by reserving capacity, high priority routes can be moved to the backup capacity with minimal data loss. Protection schemes vary by how they determine the spare capacity; primarily through whether it is route or link based, and whether it is dedicated or shared. Studies using optimal static routeing algorithms show that route based shared protection schemes use the least backup capacity, which may offset the extra time taken to activate a spare backup route.

## 2.4 Virtual Connection Schemes

There has been considerable work into extending the IP routeing mechanism, which is currently done by finding the longest prefix match at each node on the destination address and the IP routeing database to determine the next hop.

Several commercial implementations have been developed to remove the necessity for this costly operation. *IP Switching* and the *Cell Switched Router* are both based on detecting promising flows to switch using online traffic measurements, discussed in Section 2.4.1. *Tag Switching* and *Aggregated Routeing and IP Switching* are both topology based, allowing computed routes in the network to assign classes of packets to flows, discussed in Section 2.4.2. These systems have been generalised by Multi-Protocol Label Switching (MPLS) proposed by the Internet Engineering Task Force (IETF), which is discussed in Section 2.4.3.

Research into ways to adapt the virtual topology of an optical network to match dynamic traffic is discussed in Section 2.5.

### 2.4.1 Traffic Based Switching

Ipsilon Networks' proposed solution was to integrate ATM low layer technology with IP traffic to form IP Switching [Newman98]. By using soft state in the ATM hardware to cache recently seen IP flows, subsequent packets from the same flows can be switched by the ATM hardware rather than routed. Initially default single hops paths were configured, and all packets were reassembled and routed at the IP level. Flows were identified by spotting packets with matching source and destination IP addresses, or possibly by matching port numbers as well. Two flow identifications schemes were used: either requiring that a threshold number of packets have been seen from the same flow within a given time period, or relying on spotting well-known port numbers that are likely to be good candidates for switching. Once a flow had been identified, requests to switch this flow were propagated upstream. Packets could then be switched rather than routed. Established flows were deleted after 60 seconds with no packets.

Toshiba's Cell Switched Router [Katsube97] was a similar proposal based on using ATM hardware in the core of the network to switch identified IP flows, using triggers such as the detection of a TCP SYN packet or similar network flow analysis, or possibly being integrated with schemes such as RSVP.

One disadvantage of these schemes is that they use a relatively low level of aggregation for flow definition, so scaling in the middle of a large network will be a problem. To be practical, these schemes still have to be able to route the majority of packets to cope with flows too short in duration to identify [Thompson97].

More recently TCP Switching [Molinero-Fernández02, Molinero-Fernández03] takes the position of switching all packets as application-based flows, with admission control taking place on connection setup to allow peak rate allocation for each flow. Where the maximum connection bandwidth is much smaller than the link bandwidth, which is likely in the core of the network, this results in similar average completed data transfer time to packet switching while using an entirely circuit switched core.

### 2.4.2 Route Based Switching

Cisco Systems' Tag Switching [Rekhter97] was the forerunner to MPLS. Each packet would be labelled with a tag, encoded either as a shim header in front of the network layer header, directly in the data-link layer header, or as part of the network layer header. Packet forwarding was done by switching on these tags. Tag switching deals with scaling principally by assigning tags based on their destination address. This assumed that network nodes are capable of merging multiple incoming connections to a single outgoing path. It also required explicit routes to be added to enable load balancing between alternative routes.

IBM's Aggregate Route-Based IP Switching (ARIS) was a similar proposal which relied on the property of connection merging to implement network egress based routeing paths: all packets destined for a given egress point traversed a tree rooted at that egress point. Explicit *establish* messages were sent from the egress points in a reverse path multicast style until every other network node had route information for that egress point.

### 2.4.3 Multi-Protocol Label Switching

MPLS (Multi-Protocol Label Switching) [Rosen01] sets up virtual connections through the network by attaching a tag to each packet that belongs to a given FEC (Forwarding Equivalence Class). FECs are defined to be the set of packets that are routed in the same way within the network that MPLS is operating over. At each hop routeing takes place by looking up the tag in the routeing database. Operations such as merging streams or tunnelling connections can be supported by having a stack of labels. The nature of the tag depends on the network hardware; packet orientated systems might pre-pend a packet header, while ATM systems might use the VPI/VCI slots to encode the FEC label. A version of the standard that extends MPLS to non-packet switched domains has been proposed as Generalised MPLS (GMPLS) [Banerjee01], which introduces Time Division Multiplexed (TDM), Lambda Switch Capable (LSC), and Fibre Switch Capable (FSC) networks. FECs can be logically associated with the time slot or lambda together with the incoming port.

As described in Chapter 1, the two models for integrating optical networks with GMPLS are overlay and peer; hybrid models allow a mixture of overlay and peer networks to operate side by side. Overlay networks conceal the optical core, hiding the physical topology, and allowing a separate control plane for configuration. At the edge of the optical core there is an interface to the next level: a TDM switched network under the GMPLS model. This separation allows the optical core to be specialised, dealing with bandwidth allocation, error detection and protection tightly connected to the underlying hardware. It also reduces the number of hops that the higher network layer sees, since crossing the optical core counts as a single virtual hop rather than multiple physical optical interconnections. Due to this, routers at the edge of the core have many more routeing adjacencies: possibly all the other edge core routers if a full virtual mesh exists.

The peer network deals with the whole network as a set of GMPLS capable nodes under a single control plane. This reduces the possible overlap in signalling and computation load; only one set of topology discovery, routeing and restoration algorithms is needed. End to end routes now pass through more routers, but each router has a simpler task since it has relatively fewer routeing adjacencies than with the overlay model.

The work presented in this dissertation is the operation of an optical subnet, which is capable of being fitted into the GMPLS overlay model as the optical network layer. It would carry traffic from the GMPLS layer, which would in turn be responsible for aggregating the traffic using packet and time division multiplexing. Whilst in the peer model there is no separate optical layer, the operation of the LSC and FSC networks could well integrate some of the ideas contained in this dissertation.

## 2.5 Virtual Topology Adaptation

Recent research on adapting the virtual topology of an optical network, the set of all wavelength routes, has shown that it is possible to adapt to daily traffic cycles using only past traffic statistics [Gençata03]. However this uses a thresholding scheme based on the mean traffic computed over the previous five minutes, ignoring the critical timescale of the time taken to overflow packet buffers at the network egress.

Other recent work has looked at a single bidirectional fibre ring network where one node leads to a backbone network and the remainder connect to distribution networks [Lee03]. Performance of a reconfigurable network was assessed, where each access node has the ability to add or drop a limited number of the wavelengths on the fibre, but uses tuneable optical components to choose which wavelengths. This was found to be within 10% of the capacity of a network where each node is able to access all wavelengths. This work did not consider the effects of a reconfiguration on the existing lightpaths in the network nor did it integrate the extra cost of tuneable components in the performance model.

### 2.5.1 Summary

In this section we have covered some of the research and commercial projects which use optical networking to implement virtual connections. We can divide them into two classes, depending on the method used to classify packets. Traffic based classification spots flows by keeping track of packet statistics, and will suffer from the high aggregation levels present in core networks.

Route based switching is more promising, since it uses knowledge of routes through the network to classify packets on egress. MPLS has emerged as the front-runner in these proposals, and the generalised version GMPLS integrates optical networks that are Lambda and Fibre Switch Capable.

## 2.6 Conclusion

In this chapter we have covered the background knowledge needed for the rest of this dissertation. We started by looking at common Internet protocols in use now, then the areas needed to perform network simulations. Finally, we turned to the specific optical components needed to construct an optical network, together with existing schemes that use virtual connections over an optical network.

We next need to assess the feasibility of an optical network that reconfigures to support a changing traffic load. To do this we need some idea of the timescale over which the network would have to respond, and how this timescale relates to the possible gain in network utilisation from reconfiguring. This is discussed further and investigated in the following chapter.

# Chapter 3

# Traffic Measurement

The previous chapter contained some of the advances in optical networking technology which allow virtual connections to be built from a continuous wavelength route, passing through multiple nodes without being translated to electrical form. It is clear that with these advances it is now possible to reconfigure optical networks much faster than previously. By looking at representative network traffic this chapter explores how the potential gain from reconfiguring our network is related to how often we reconfigure.

## 3.1 Introduction

The emergence of optical amplifiers has made long distance optical links possible, but usually all traffic arriving at a node would be converted to electrical form before being routed or switched and then converted back to optical for the next hop. Now, with the advent of optical switches, it is possible to create virtual paths through the network that leave the signal in optical form. Since the route is optically switched we are now able to reconfigure the network to change the virtual topology presented to the higher network layers. MEMS devices can switch in around 10 ms so, although not suitable for packet switching, they may be used profitably in reconfiguring a circuit-switched network.

Reconfiguring optical networks may lead to benefits if we can track the demands placed on the network by traffic or management led stimuli. At the simplest level we can use slack capacity in one area of the network to satisfy excess load in another area. The new virtual connections may be longer, routeing around the congested area; however since each route is optical we are only increasing the propagation delay rather than adding extra queueing delays.

The potential gain from such a system depends primarily on the traffic load placed on the network. If the traffic levels across the network do not vary over time then there will be no gain from reconfiguring the network. A key parameter is the magnitude of the change in traffic load compared with the minimum size of a virtual connection, and how fast this change occurs. If the traffic between two points in the network is always smaller than a single virtual connection, then again our gain from reconfiguring is limited. Similarly if the traffic load varies over the scale of many virtual connections, or varies faster than we can respond, we are faced with a high management overhead of setting-up and tearing-down many connections.

Since we are designing a network infrastructure for the future it is impossible to directly measure the exact traffic that it will have to deal with. Applications and services that use our network will change, as will the number and location of users. The proposed change in design to the network core will also affect traffic patterns.

One feasible approach is to measure what current conditions are, and either use some multiplexing scheme to scale traffic to the levels that will be experienced in the future, or look for features of the traffic that will remain in the future.

We can then analyse the potential performance gain for a reconfiguring circuit switched network, based on our approximate knowledge of network traffic. This requires the development of algorithms to perform the dynamic bandwidth allocation for a traffic trace; we can then quantitatively measure the utilisation benefits from being able to reconfigure our network, and how that relates to the timescale over which we can change bandwidth allocation and the relationship between the change in traffic levels and the size of our virtual connections.

This analysis is carried out for a single stream only: in a real network we would need shifts in demand between different traffic streams. If all streams are correlated, and increase and decrease in phase, then there is no scope of network-wide capacity reduction since we would simply require peak rate allocation.

## 3.2 Measurement Apparatus

Measurement data was taken from the link between the University of Cambridge and JANET [JAN], the UK Academic Network, using a measurement infrastructure in place from the Nprobe project [Hall01]. The data rate through the link, peaking at around 39,000 packets per second, is higher than most available traffic archives. Within the provisions of the agreement to measure at this point, we have access to anonymised IP addresses that preserve netblock and AS information to enable aggregation of traffic flows. This is in contrast with some existing data sets that encrypt addresses, destroying the locality information. Finally we are able to filter unnecessary detail from packet headers to minimise the storage and processing times when dealing with large quantities of data.

The placement of the probe machine was such that the traffic measured was almost entirely traffic travelling from within the University to outside, or vice-versa. The small

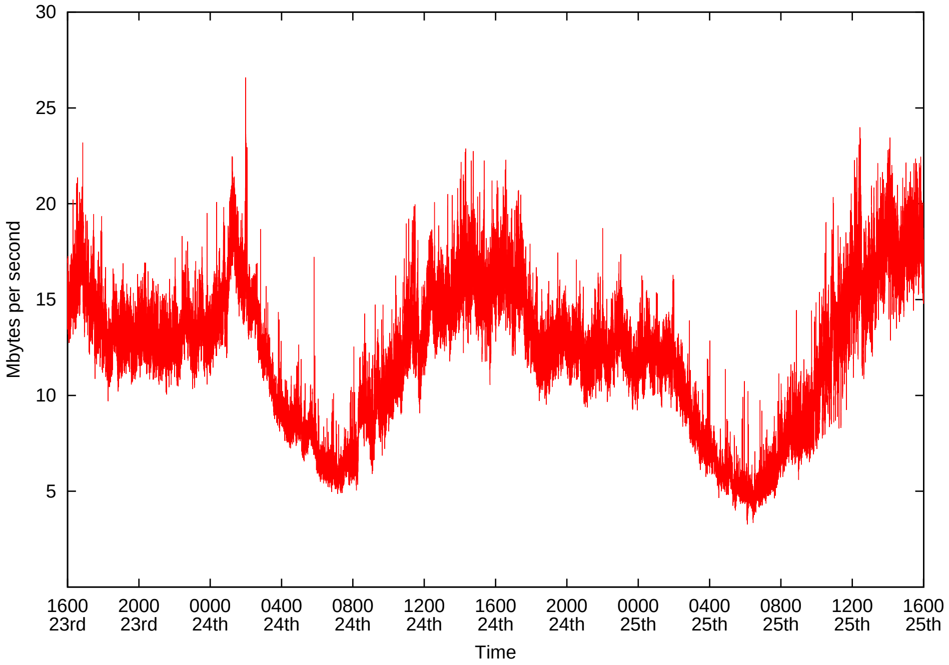

FIGURE 3.1: MBYTES PER SECOND

amount of internal traffic measured was router-to-router information and was negligible. The trace measured all IP datagrams, storing their source and destination addresses, packet size, timestamp to 500 $\mu$s resolution, and whether a packet was a TCP SYN or TCP FIN packet. Since the existing infrastructure dealt with online analysis of higher network protocols, code was written that interfaced with the packet capture engine and stored the per packet information. This was compressed to be around 9 bytes per packet, taking advantage of the compression gained from hashing local addresses. The amount of non-IP traffic at this measurement point had previously been found to be negligible. The trace lasted 48 hours, and was taken mid week at the start of Easter Term 2002, when network load was likely to be high with most members of the University in residence.

## 3.3 Basic Statistics

Firstly the data was analysed to compare it with similar reported traffic traces in the literature. Figure 3.1 shows total traffic as bytes per second. This link averages around 15 Mbyte/s, or around 120 Mbit/s. A clear daily cycle can be observed with a main busy period between noon and 2am in each 24 hour period. Figure 3.2 shows the equivalent graph for packets per second for the first 48 hours of the trace. A similar trend is observed.

Figure 3.3 shows the number of TCP SYN and TCP FIN packets per 10 s, smoothed using 10 s bins so the two lines do not obscure each other. This gives a sustained average

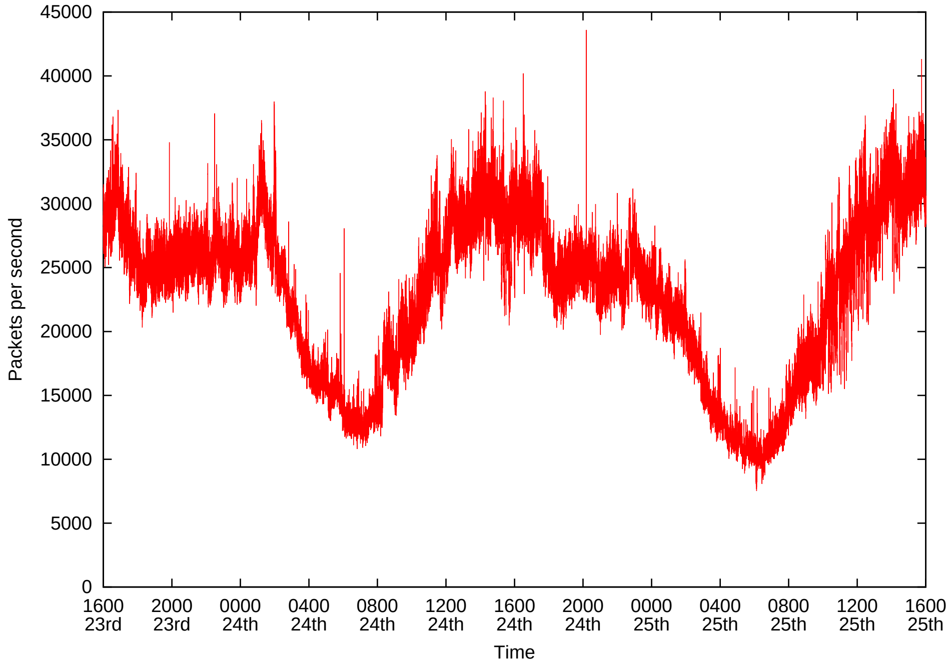

FIGURE 3.2: PACKETS PER SECOND

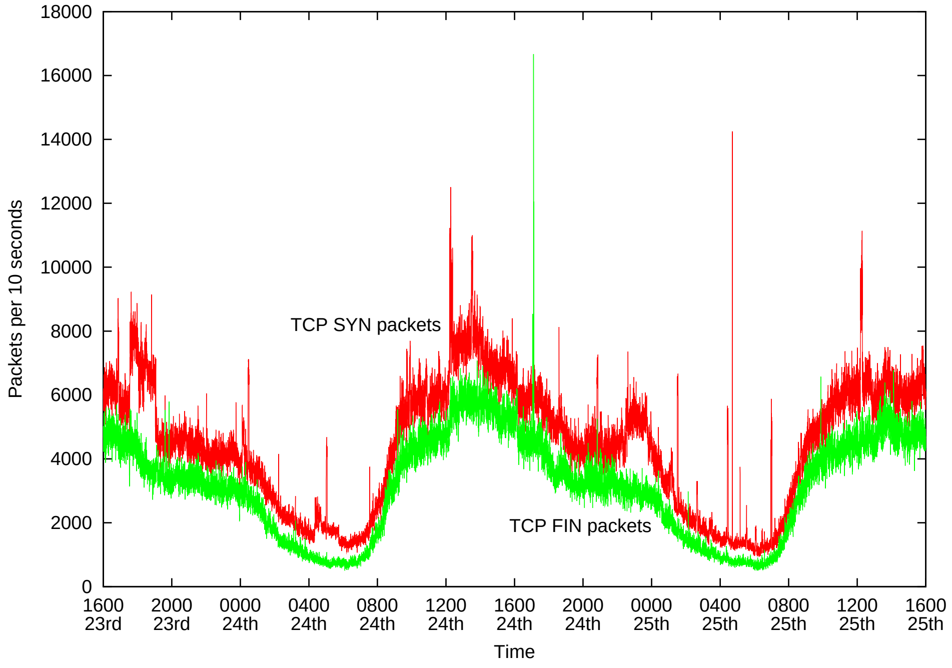

FIGURE 3.3: TCP SYN AND FIN PACKETS PER 10 S

of around 250 completed TCP connections per second during peak hours, since each connection closes using two FIN packets.

In this figure we see some clear effects; firstly that the number of SYN packets is consistently greater than the number of FIN packets. This is expected since SYNs may be dropped by a server as a form of admission control when the server load is high, so will be retransmitted more frequently than FINs. Also a well-known web client terminates the TCP connection responsible for an HTTP1.0 download by replying to the server-sent FIN with a reset, rather than a second FIN. However some other effects can be seen here, as at about 1800/23rd, 0500/24th, 1300/24th, and 2300/24th the difference between the number of SYNs and FINs increases for a period of a couple of hours on each occasion. Investigation of the data reveals these to be host scans — a remote host testing entire IP netblocks for local hosts by attempting to start up a TCP connection — which are almost certainly automated network attacks. This is a good illustration of where even for a large institution the activity of single individuals can be visible at large levels of aggregation. Another visible effect is the FIN spike seen just after 4pm/24th, involving over 24,000 FIN packets sent to a particular internal host over a 45 second period. This might either be a mis-configured program or a deliberate denial-of-service attack.

## 3.4 Path-based Aggregation

In order to extrapolate future traffic by multiplexing streams together it is useful to know how current streams multiplex together to create the measured traffic. Therefore we need some way of splitting the measured traffic into different streams. Using the model of virtual connections we estimate how the measured traffic would be carried through the JANET AS. Traffic bound for the same destination AS would take the same path through the network, as would traffic whose AS path follows the same set of intermediate ASs. Whilst in the future the concept or scale of an AS may vary, this approach is not only practical but the only reasonable method of splitting the measured traffic. The concept of splitting traffic is covered further in Section 6.2.1.

A simple example is shown in Figure 3.4. Here we consider the outgoing traffic case, with traffic flowing from left to right. The circles mark ASs; traffic can either terminate within this AS or be passed to a neighbouring AS. The quantity of traffic terminated is marked in the circle; a total of eight units are sent. If we assume that we can allocate optical paths for any traffic path carrying two units or more then the red lines indicate where optical paths would be created. Since no traffic terminates at Destination A, a path carrying four units travels direct to Destination B. Four units are also carried to Destination C, but since two units terminate here and the paths to D and E only take one each, the red optical path terminates here.

This model is realistic if some form of inter-AS lightpath assignment is operation; hence it is beneficial to organise traffic into separate lightpaths even though they follow the

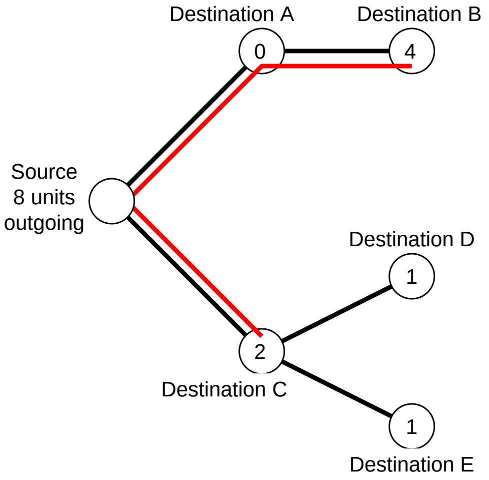

FIGURE 3.4: AN EXAMPLE AS MAP SHOWING HOW MUCH DATA IS CARRIED TO EACH DESTINATION AS.

same route through our AS. Without this assumption we would only be interested in the next hop AS, assuming that different neighbouring ASs peer with us at different points in our network. Even if no inter-AS lightpath assignment is used this model may still be relevant if the size of a single AS increases over time.

External addresses from the measured traffic data were extracted and routes to them were calculated using jroute [JRoute], a traceroute-like program designed to give good performance using parallel TCP probes and the ability to overcome most firewalls. This was performed the day after the trace was taken to capture the topology closest to that in use during the measurement period. These were combined with BGP table information from the Route Views project [RViews] to give the AS path used by the measured traffic. The BGP table used was that archived at the time of the measurement period. Optical paths were then determined as described above, with a threshold for minimum optical path size of 2.5 Mbit/s — around 2% of the total average traffic. This data rate was not chosen to be representative of a possible optical network design; in current optical networks a wavelength typically carries 10 Gbit/s of data. It was chosen to break the measured traffic into a small set of streams conducive for analysis.

The final AS map is shown in Figure 3.5. The root of the map, JANET, is labelled. Links are colour coded to indicate the average bandwidth traversing that link, using a log scale to map bandwidth to colour index. Links in red carried approximately 2 Mbit/s, yellow averaged 20 Kbit/s, and green 1 Kbit/s. Dark blue links carried a total of 100 Kbit over the 48 hour period. At each branching point the angle-range given to each child AS is proportional to the number of ASs linked to by that child. The end of each optical path is labelled with the number of that path. Traffic carried by that path goes to all the children of the terminating AS, apart from ones covered by a longer optical path. The average AS path length measured was 3.5, when weighted by the quantity of traffic over that AS path. This is biased by the large amount of traffic to the web caches situated in JANET: excluding single hop AS traffic increases

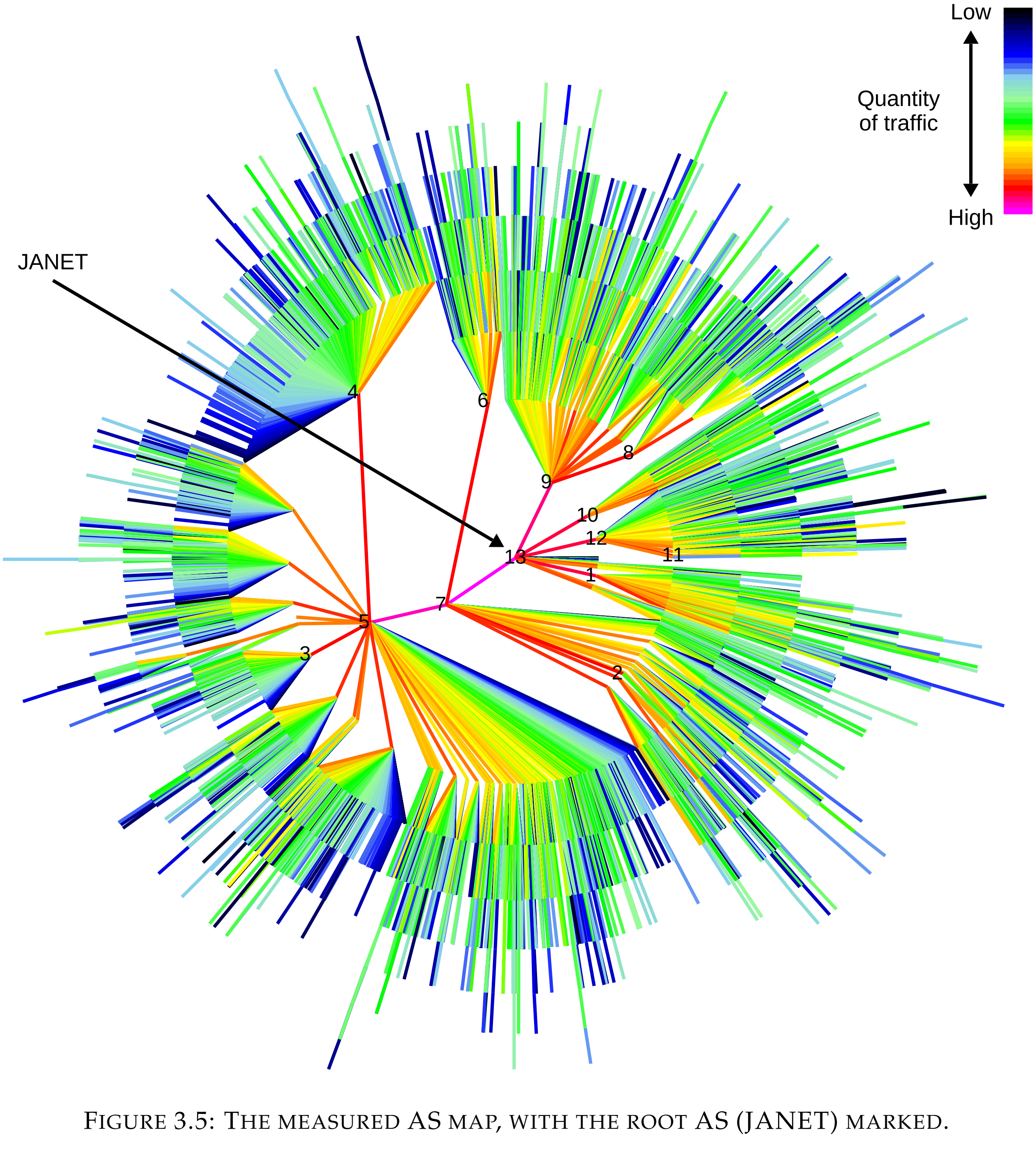

FIGURE 3.5: THE MEASURED AS MAP, WITH THE ROOT AS (JANET) MARKED.

the average AS hop distance to 4.

## 3.5 Time Series Analysis

In order to determine how this traffic will aggregate together it is useful to know the statistical properties of this traffic. Network traffic has been found to contain non-stationarity and self-similarity, and also to approach a Gaussian Marginal Distribution (GMD) at high levels of aggregation. The level at which traffic aggregations exhibit a GMD is not obvious, and if the aggregate has a GMD the properties of the component parts of that aggregate may not. Since some tests for self-similarity rely on a GMD, and knowing that traffic has a GMD allows easier statistical tests to be carried out, this is investigated first.

In the following section we use $X_i$ to refer to the quantity of traffic in time period $i$, where there are $n$ time periods. The notation $E[D]$ refers to the expected value of the distribution $D$, and in the special case of $X$ we use $\mu = E[X]$ and $\sigma = \sqrt{E[(X-\mu)^2]}$ for the mean and standard deviation respectively.

We use Quantile-to-Quantile (QQ) plots to give a visual analysis of marginal distributions. Here the sample data $X_i$ is ordered, and for each sample we calculate the number of standard deviations it is away from the mean. From our comparison distribution we calculate the predicted number of standard deviations from the mean as given by the placing in the ordered samples. As an example, assume we are testing against a GMD where our measured distribution $X$ has $n_X = 1000, \mu_X = 10, \sigma_X = 1$. If the 10th lowest sample value is $8$, this gives a value of $-2$ since $8 = \mu_X - 2 \cdot \sigma_X$. Since this is the 10th lowest sample, if $G$ has a GMD we calculate $d$ where $P(G_i < \mu_G + d \cdot \sigma_G) = 0.01$; $d = -2.054$. In the QQ plot we would then add a point at $(-2.054, -2)$.

### 3.5.1 Gaussian Marginal Distribution

Two periods of 10,000 s (around 2 hours 47 minutes) were identified in the trace, the first at high load and the second and low load. The high trace was from around 2200/23rd, the low from 0700/24th. Short time periods were used to minimise the effects of any non-stationarity present.

Figure 3.6 shows the QQ plot of the marginal distribution of the high traffic trace against the GMD. Although there is a marked turn up at both ends, the plots, taken with 100 ms, 1 s, and 10 s bins, are all close to the GMD. The departure from GMD implies that there are more bursts than expected and fewer very slack periods in traffic.

From the same high period Figure 3.7 shows the QQ plots for the traffic separated on the paths from Section 3.4, and placed in 1 s bins. This clearly shows that whilst some streams are as close to a GMD as the full traffic mix, some have a markedly non-GMD.

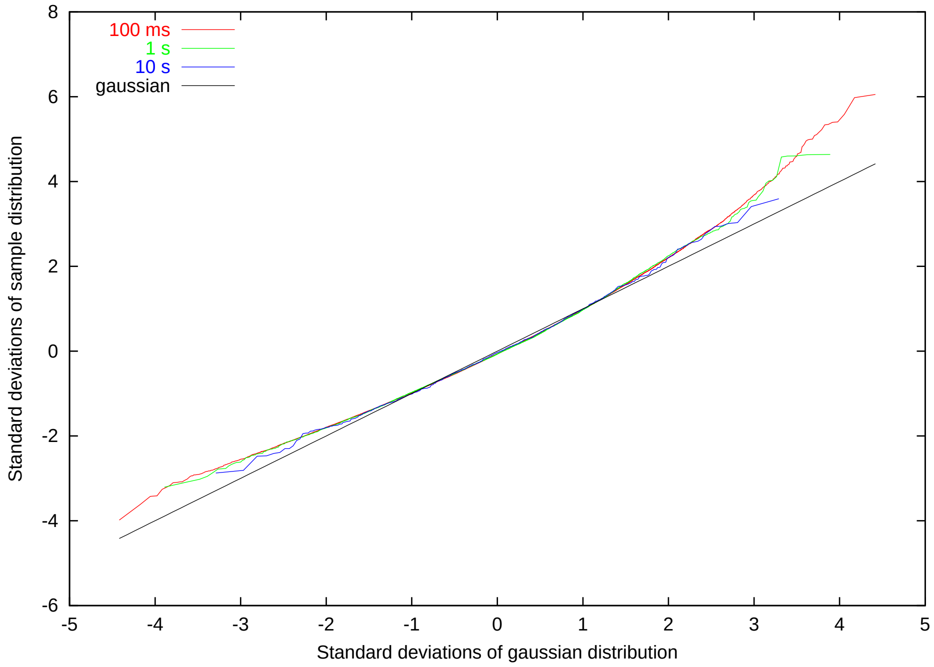

FIGURE 3.6: HIGH LOAD 10,000 S; QQ PLOT OF MARGINAL DISTRIBUTION AGAINST GMD, USING 100 MS, 1 S AND 10 S BINS

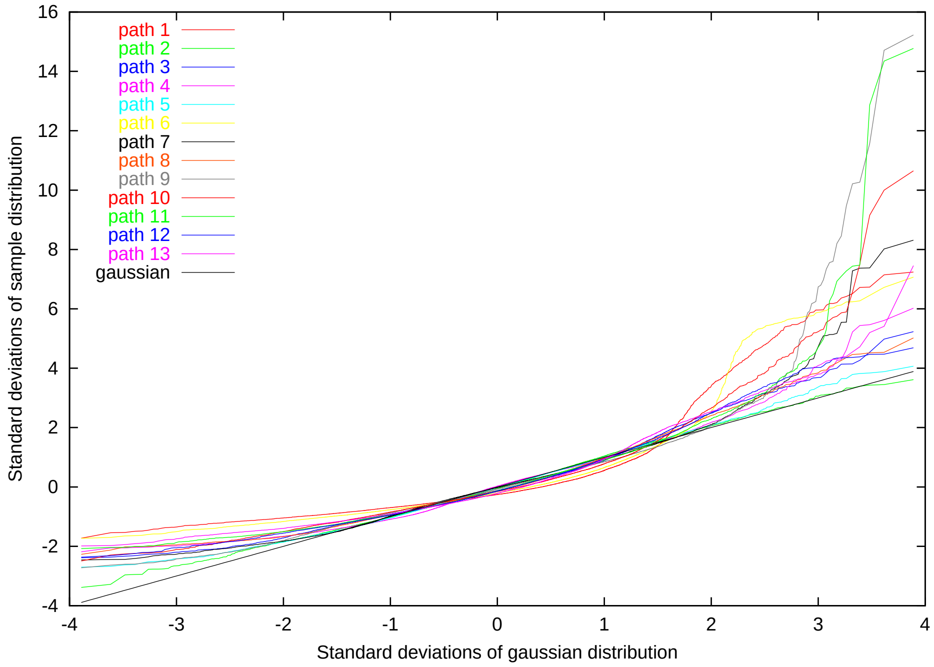

FIGURE 3.7: HIGH LOAD 10,000 S USING THE SPLIT TRAFFIC STREAMS; QQ PLOT OF MARGINAL DISTRIBUTION AGAINST GMD USING 1 S BINS

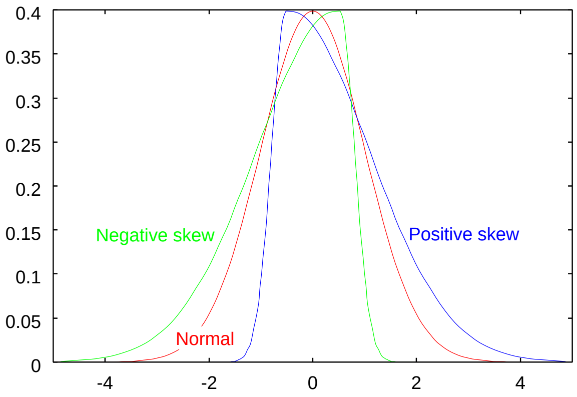

FIGURE 3.8: GMD SHOWN WITH THE EFFECTS OF SKEW

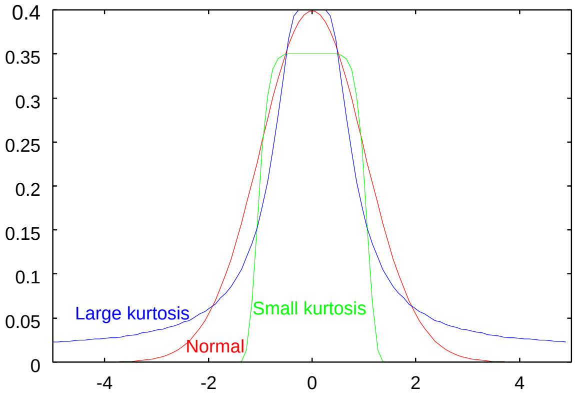

FIGURE 3.9: GMD SHOWN WITH THE EFFECTS OF KURTOSIS

This can be shown quantitatively by using formal statistical analysis.

$$\sqrt{\beta_1} = \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \tag{3.1}$$

$$\beta_2 = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} \tag{3.2}$$

Statistical tests for a GMD effectively test the third and fourth moments of the marginal distribution, which measure the skew and kurtosis. The skew, calculated using Equation 3.1 measures how symmetrical the distribution is; a positive value indicates a larger upper tail, a negative value indicates a large lower tail. The GMD has a skew of zero. This can be seen in Figure 3.8. The kurtosis is the fourth moment, calculated using Equation 3.2, and indicates how extended both tails of the distribution are. The GMD has a kurtosis of 3; smaller values indicate smaller tails and vice versa. These effects can be seen in Figure 3.9.

Rather than test for each statistic, omnibus tests have been introduced which combine the tests; Shapiro and Wilk's W test has been found to be the most reliable [Shapiro68], but it is not typically implemented for sample sizes greater than 2000 due to the large number of coefficients needed. A simpler test is the D'Agustino D test, with the result transformed into the Y statistic, and significance bounds calculated using the Cornish-Fisher expansion [D'Agostino86].

$$D = \frac{\sum_{i=1}^{n}(i - \frac{1}{2}(n+1))X_i'}{n^{\frac{3}{2}}\{\sum_{i=1}^{n}(X_i' - \overline{X})^2\}^{\frac{1}{2}}} \tag{3.3}$$

The D statistic is calculated using Equation 3.3, where $X_i'$ is the sorted list of $X_i$. The expected value of D is approximately $1/(2\sqrt{\pi})$, with the standard deviation is asymptotically

| bin size | 100 ms | | 1 s | | 10 s | |
|---|---|---|---|---|---|---|
| 95% bounds | -1.97 | 1.94 | -2.02 | 1.89 | -2.16 | 1.75 |
| all traffic, high | 17954 | | -5.19 | | -0.909 | |
| all traffic, low | 5069.27 | | -29.1 | | -2.57 | |
| path 4 | 17831 | | -3.07 | | 0.252 | |
| path 5 | 18018 | | 2.48 | | 1.4 | |
| path 2 | 17762 | | 4.28 | | 2.37 | |
| path 8 | 17537 | | -19.4 | | -4.79 | |
| path 7 | 17161 | | -23.7 | | -4.8 | |
| path 11 | 16831 | | -50.2 | | -7.99 | |
| path 12 | 17292 | | -26.8 | | -9.81 | |
| path 9 | 17007 | | -68.2 | | -10.6 | |
| path 3 | 17706 | | -26.5 | | -12.3 | |
| path 13 | 17369 | | -40.6 | | -13.9 | |
| path 10 | 17069 | | -76.3 | | -24.9 | |
| path 6 | 15787 | | -148 | | -50.8 | |
| path 1 | 13831 | | -177 | | -52.7 | |

TABLE 3.1: RESULTS OF TESTING FOR GMD. IF THE MEASURED Y STATISTIC IS OUTSIDE THE BOUNDS LISTED, WE REJECT THE HYPOTHESIS THAT THE MARGINAL DISTRIBUTION HAS A GMD WITH 95% CONFIDENCE

$$\left[\frac{12\sqrt{3}-27+2\pi}{24\pi n}\right]^{\frac{1}{2}} = \frac{0.02998598}{\sqrt{n}}$$

so the D statistics can be transformed into an approximate standardised variable with zero mean and unit standard deviation, Y, as follows:

$$Y = \frac{\sqrt{n}(D-0.28209479)}{0.02998598}$$

This Y statistic can be easily visually and graphically compared to significance bounds at varying confidence levels. These formulae use figures to the given level of significance following from their use in [D'Agostino86].

Table 3.1 shows the results of applying the D test. The traces tested are the aggregate traffic during the two 10,000 s periods identified previously, and each of the separated traces during the high activity period. All these 15 traces were tested with 100 ms, 1 s, and 10 s bins, and the separated traces ordered by the absolute value of the Y statistic with 10 s bins.

The 95% error bounds mean that if the measured Y statistic is outside those bounds we can reject the null hypothesis that our measured sample is from a GMD with 95% confidence. For all samples where we look at bytes arriving during each 100 ms period

| | Y | | $\sqrt{b_1}$ | | $b_2$ | |
|---|---|---|---|---|---|---|
| all traffic, high | -5.19 | 2.03 | 0.552 | 0.417 | 3.43 | 2.86 |
| all traffic, low | -29.1 | -2.31 | 0.954 | 0.401 | 6.65 | 3.08 |
| path 4 | -3.07 | 6.70 | 0.453 | 0.271 | 3.42 | 2.63 |
| path 5 | 2.48 | 5.38 | 0.217 | 0.162 | 2.90 | 2.71 |
| path 2 | 4.28 | 6.05 | 0.193 | 0.157 | 2.78 | 2.68 |
| path 8 | -19.4 | -7.9 | 0.755 | 0.566 | 3.74 | 3.05 |
| path 7 | -23.7 | -1.39 | 0.658 | 0.176 | 5.52 | 2.96 |
| path 11 | -50.2 | -6.73 | 1.69 | 0.609 | 15.5 | 3.11 |
| path 12 | -26.8 | -7.75 | 0.906 | 0.591 | 4.20 | 3.10 |
| path 9 | -68.2 | -7.02 | 1.94 | 0.23 | 21.7 | 3.26 |
| path 3 | -26.5 | -16.9 | 0.79 | 0.624 | 4.00 | 3.36 |
| path 13 | -40.6 | -24.3 | 1.13 | 0.892 | 4.52 | 3.46 |
| path 10 | -76.9 | -24.9 | 1.67 | 0.745 | 9.61 | 3.81 |
| path 6 | -148 | -39 | 2.63 | 1.07 | 13.5 | 4.38 |
| path 1 | -177 | -70.9 | 2.71 | 1.49 | 12.5 | 6.15 |

Table 3.2: Results of testing for heterogeneous connection bandwidths. For each statistic, the left column is the full trace; the right column is the beta trace with the most bursty connections removed. If the Y statistic is outside -2.02 and 1.89 we reject the hypothesis that this distribution has a GMD with 95% confidence. For a GMD, $\sqrt{b_1} = 0, b_2 = 3$

we can reject our null hypothesis; these distributions almost certainly do not have a GMD. At the 1 s timescale we are closer to a GMD, especially with the high aggregate trace and paths 4, 5, and 2, but we again reject the hypothesis in all cases. Smoothed over 10 s we accept that the high aggregated traffic has a GMD. The separated streams show a wide variation: some are also accepted or nearly accepted that they have a GMD whilst others are easily rejected. The trace taken during the low period exhibits a significantly higher Y statistic than the high aggregate trace at 1 s and 10 s timescales.

These results confirm the informal results from Figure 3.6 and Figure 3.7; the aggregate traffic is fairly close to a GMD, especially at longer timescales, whereas the separated traffic streams have a wide variety — some curves are fairly straight, whilst others diverge strongly.

An explanation of some of these findings is that is it possible for a very small number of individual connections to dominate measured traffic during a burst, typically a single connection. We can use a thresholding method to separate the traffic into dominant connections and the remainder [Sarvotham01b]; the dominant 'alpha' connections are conjectured to be the coincidence of a large file transfer with a high bandwidth network route. For each one second interval where the total traffic is larger than a threshold value, we remove the bytes that are due to the largest connection within that time interval. This extraction of the 'alpha' bursty connections leaves the remainder 'beta' traffic. The results are shown in Table 3.2, with the traces ordered as for Table 3.1. The threshold value used was the mean of the traffic plus five times the standard deviation.

For the beta traffic both $\sqrt{b_1}$ and $b_2$ are closer to the GMD values of 0 and 3 respectively, with the most improvement from the thresholding to the most non-GMD traces. The Y statistic shows that the beta traffic is closer to having a GMD, although only one beta trace falls within the 95% confidence interval. However, while for the separated traces this transformation helps to explain a large proportion of the non-GMD, the effects are critically determined by the level of the threshold value. For traces close to a GMD the kurtosis figure drops below 3, indicating that we are classifying too much traffic as alpha: for traces far from a GMD the transformation is not strong enough. This is understandable since the threshold used was set for each trace based on their mean plus standard deviation; perhaps a better classification scheme would be to fit the central part of the marginal distribution to a GMD, ignoring the upper tail, and use the standard deviation from this fit.

Closer investigation of the paths reveals a possible correlation between the burstiness of a path and its geographical destination. If the maximum bandwidth available for a set of connections is low, then traffic should be closer to having a GMD since you have no possibility of an alpha connection. Two of the closest GMD paths are 4 and 5; path 4 terminates at Alternet, and path 5 at ASN-QWest, both of which are US networks. They both use a connection which is shared amongst a large number of peering networks; this trans-Atlantic link is a well known source of congestion for access to some US networks. These traces were also taken in the evening, daytime in the US, where American networks are most likely to be loaded. Conversely, the least GMD path, path 1 is also an American network, Abilene. However this serves academic, government institutions, and high-performance research networks, and has a dedicated direct link from JANET. Of the other non-GMD paths path 6 is Ebone, a European network, path 10 is GEANT, a European research network similar to Abilene, and path 9 terminates at LINX, the London Internet Exchange.

Since the path destinations are only part of the total journey, typically only two or three AS hops out of maximum of around ten, the correlation between path destination and closeness to a GMD of observed traffic is not perfect, although clear differences are observable. With the commercial collapse of several of these networks since these measurements were taken, notably KPN-QWest marked number 7 on Figure 3.5, it would be instructive to compare these results with current measurements.

The major conclusion to draw from this data is that even though the majority of traces have a distinctly non-GMD, when aggregated together they form a traffic trace that is very close to a GMD. This can be explained from a connection basis, since when there is a high level of aggregation the high bandwidth 'alpha' connections will be numerous enough that they start to multiplex together and appear less bursty. There is also an upper limit on the bandwidth that is available in this measuring environment — the maximum line bandwidth connecting this measurement point with the rest of the Internet. A connection that has this full bandwidth will dominate a lower activity trace more than a high activity trace.

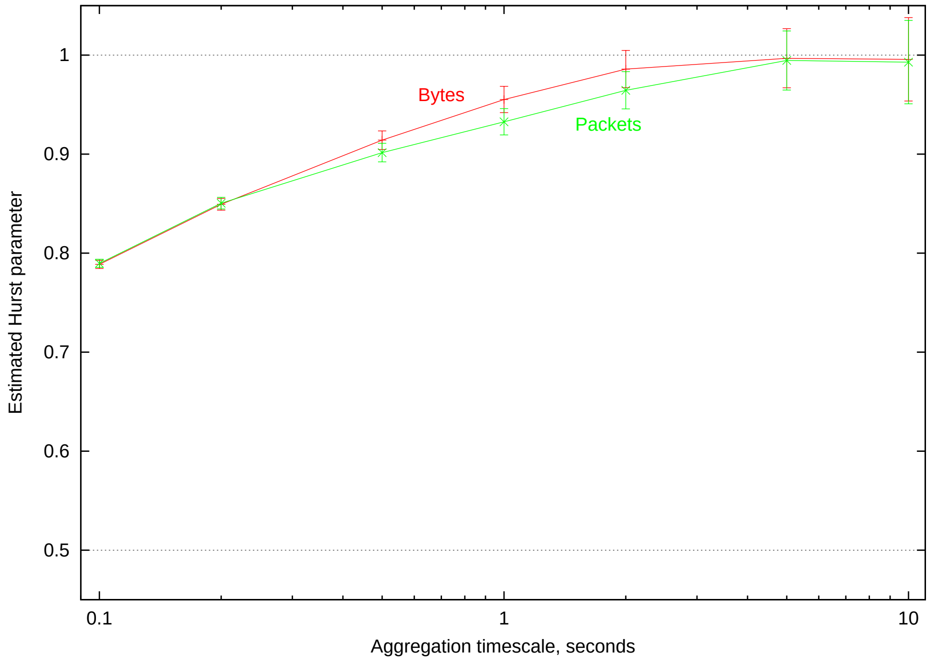

FIGURE 3.10: HURST PARAMETER ESTIMATION FROM TOTAL HIGH NETWORK LOAD TRACE. THE TRACE WAS AGGREGATED OVER DIFFERENT TIMESCALES AND THE HURST PARAMETER OF BYTES AND PACKETS WAS ESTIMATED, WITH 95% CONFIDENCE INTERVALS AS SHOWN.

### 3.5.2 Self-similarity

It has widely been reported that Internet traffic is self-similar. In this section we try to confirm this using the standard Hurst parameter estimation techniques. Whittle's approximate Maximum Likelihood Estimator (MLE) [Whittle51] has been shown to have statistically suitable properties allowing confidence intervals to be calculated, unlike graphical based methods such as the R/S statistic and variance-time plots. Whittle's MLE works by fitting the sample data to a known distribution. In this case we will use fractional Gaussian noise as the distribution since our data has an asymptotic GMD.

While the Hurst parameter measures the degree of long range dependence in our data, exact self-similarity would demand the same Hurst parameter for the data aggregated over any timescale, whilst second order self-similarity requires the Hurst parameter to converge as the aggregation timescale increases. With finite time series, especially when the sample size is deliberately reduced to remove any large non-stationary effects, both definitions can be hard to distinguish precisely.

Figure 3.10 shows the Hurst parameter and 95% confidence intervals from Whittle's MLE, using code from [Beran94] and coded into the R statistical language [Ihaka96]. The full trace from the busy period investigated in Section 3.5.1 was aggregated over different timescales. The longest timescale used was 10 s, which gave 1000 samples. The Hurst parameter and confidence intervals were estimated for the amount of bytes in each time period, and also the number of packets in each. Lines marking the limits of the Hurst parameter at 0.5 and 1 are also shown; these are the two limits of the Hurst parameter, where 0.5 indicates no long-range dependence, and the degree of long-range dependence increases as the Hurst parameters gets closer to 1.

Whilst we can clearly discount either process being exactly self-similar there is a clear trend towards a Hurst parameter approaching 1 as the aggregated timescale increases. At smaller timescales either short range dependence dominates, or the smaller Hurst values are due to the distribution moving away from a GMD as seen in Table 3.1.

### 3.5.3 Correlation Distributions

In the presence of both long range dependence and non-stationarity it is hard to accurately distinguish one from the other. Since there are many ways that a time series may be non-stationary, there are no clear statistical tests for non-stationarity. Weak non-stationarity occurs when either of the first two moments of the distribution change, the mean or variance. The dividing line between a non-stationary event and a long-term traffic 'swell' is not clear. To sidestep this problem we look at the correlation structure of our measured traffic. This is motivated by our ultimate goal; to know how frequently we need to reconfigure our optical network. For operational purposes it doesn't matter whether a shift is caused by long range dependence or a non-stationary event, since the effect is the same.

Figure 3.11, Figure 3.12 and Figure 3.13 all show the distributions of lags up to a 24 hour, 90 minute, and 155 s lag respectively. Each figure was constructed in the same manner using the same data; the second two graphs progressively zoom in on the first, giving more detail around the origin. The choice of 155 s for the third trace is designed to lead to a visible resolution of 1 s lags. The full 48 hour trace, placed into 1 s bins, was examined. For each time lag, $t$, plotted on the x-axis, the probability distribution of $|X_i - X_{i+t}|$ for all $i$ is calculated. For a given $t$ and magnitude of change $c$ we then have $P(|X_i - X_{i+t}| > c)$. We scale the magnitude of change, $c$, to be from 0 to 1, where 1 represents a change from 0 bytes per second to the maximum value recorded throughout the 48 hour trace, and 0 represented no change. This scaled $c$ is plotted on the y-axis, with the value of the cumulative probability function indicated by the colour at that point. These values are plotted on a log scale using the colour chart as indicated on the Figure. Thus the point on Figure 3.11 at 4 hours on the x-axis and 0.4 on the y-axis is marked by a colour representing 0.01, so $P(|X_i - X_{i+4hours}| > 0.4\max_i[X_i]) = 0.01$.

Figure 3.11 shows clearly that there is a strong daily cycle by the dropping of the colour bands towards the 24 hour point, indicating that two spot bandwidth measurements taken 24 hours apart are likely to be closer together than two measurements taken with a lag of between 2 and 22 hours apart. The 'm' shape of the top section of the graph indicates that on a 12 hour lag there are no huge jumps in traffic; for example looking at Figure 3.1 the largest drop in traffic occurs between 0200 and 0700 each night; there is a corresponding drop between peak traffic levels at around 1500 and 1800. However this dip is not maintained beyond around the largest 10% of traffic shifts.

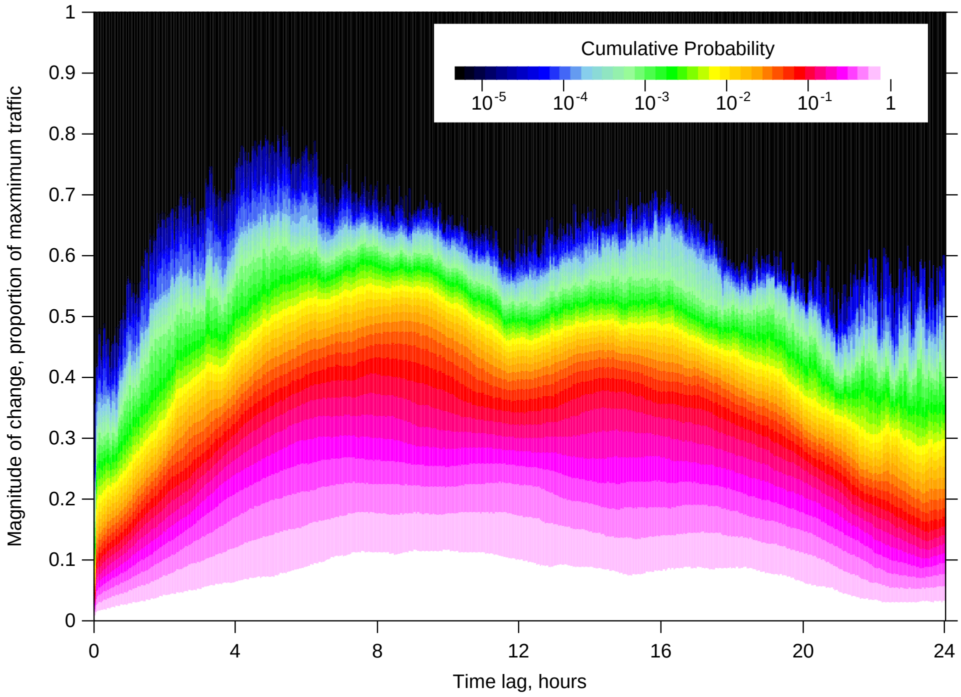

FIGURE 3.11: 3D PLOT OF LAG DISTRIBUTION UP TO 24 HOUR TIME LAG. FOR EACH TIME LAG, THE CUMULATIVE PROBABILITY OF THE DIFFERENCE IN MEASUREMENTS AT THAT LAG IS PLOTTED AS A VERTICAL LINE. THE COLOUR INDICATES THAT THE PROBABILITY OF A DIFFERENCE IN MEASUREMENT IS GREATER THAN THE Y-AXIS VALUE

One key area is the ramp up at low time lags. This is expanded in Figure 3.12 and further still in Figure 3.13, to a maximum time lag of 90 minutes and 155 s respectively. This region shows a very close correlation between adjacent values; at a 1 s time lag — adjacent measurements — 99% of differences are within around 7% of the maximum traffic level. However, within a few seconds this increases to around 12%. From a 10 s time lag, the increase for the 99% mark is almost linear until 4 hours. At this point the rate of increase slows; the 99% mark reaches a maximum at 8 hours.

This has several important consequences for any scheme where we wish to allocate bandwidth for this aggregate trace, or more generally any similar set of traffic with a similar profile. Firstly, if we cannot allocate bandwidth more frequently than every 4–8 hours, then there is little point allocating bandwidth more frequently than is necessary to cope with the gradual increase in traffic levels over a time period of weeks or months. At this 4–8 hour mark, the variability in traffic has reached its maximum. From Figure 3.1 this looks reasonable; if we wish to allocate efficiently we have to be able to cope with the daily cycle, of which the slackest period lasts around 4–8 hours, from 0300 to 1000 each day. If we cannot allocate this frequently, then to avoid major congestion at peak periods we have to allocate for those peak periods, and keep this allocation throughout the day. A weekly cycle of traffic demand has also been observed [Thompson97, Papagiannaki03], so allocating at peak rate every one or two days might be useful, although this is not possible to deduce from our data.

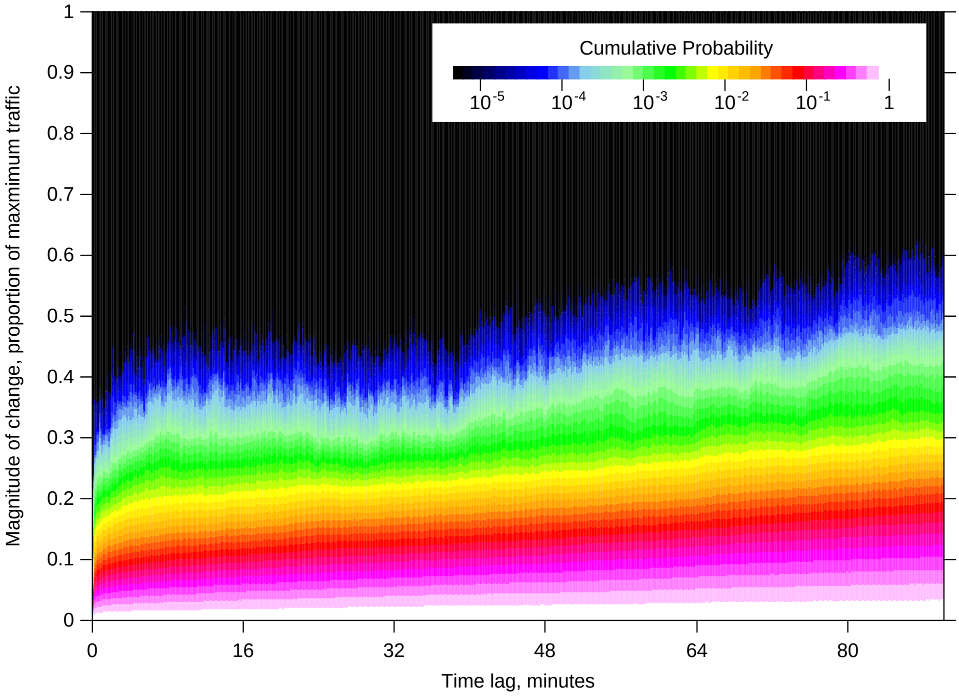

FIGURE 3.12: 3D PLOT OF LAG DISTRIBUTION UP TO 90 MINUTE TIME LAG. AS FIGURE 3.11

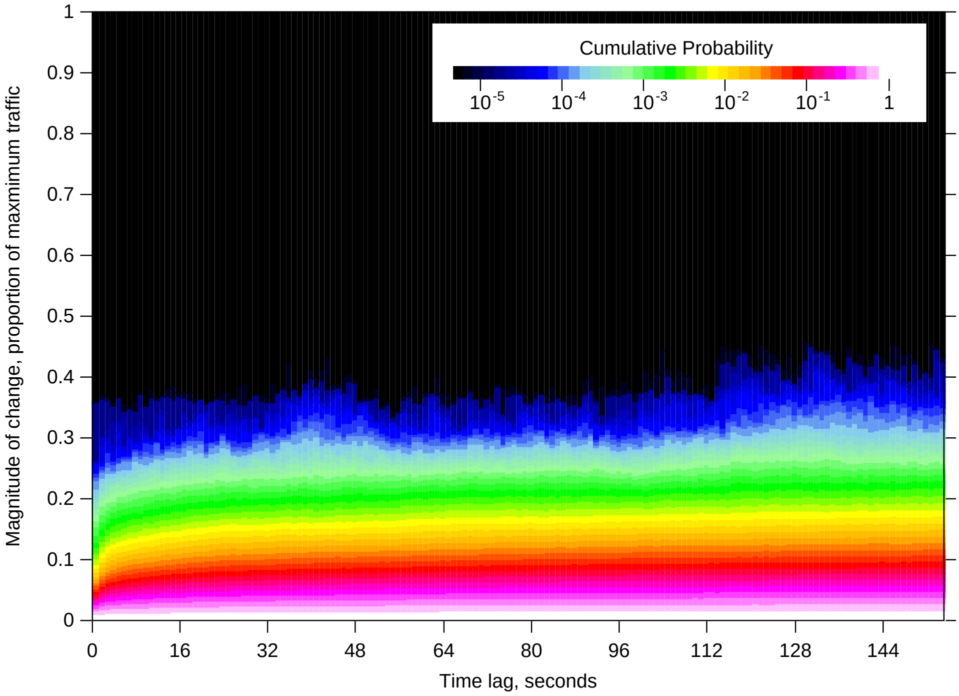

FIGURE 3.13: 3D PLOT OF LAG DISTRIBUTION UP TO 155 S TIME LAG. AS FIGURE 3.11

If we are to allocate bandwidth more frequently than every 4 hours, then is there any better timescale at which to do it? Apart from the first 10 s, the increase in the spread of lags is nearly linear with the time lag. This implies that the relationship between the granularity of allocation and the necessary timescale between allocations is also linear. If this curve had a marked dip — maybe an exponential rather than linear relationship — then we could select the 'knee' of the dip as the optimum point. For example in the extreme case that bandwidth was always constant for periods of 1 minute, then the graph would have an obvious inflection point at 1 minute. However the positive intercept of the linear region at the origin indicates that it is relatively more inefficient to allocate more frequently and at a finer granularity. It is unlikely that with the possible delay imposing by creating a new wavelength route we would be able to reconfigure often enough to be in the non-linear region near the origin.

## 3.6 Bandwidth Allocation Algorithms

Whilst the analysis in Section 3.5.3 indicates that we may on average allocate a level of bandwidth for between a few minutes and a few hours, it does not indicate how quickly we have to be able to respond to changes in demand. The simplest way to understand this is to perform the bandwidth allocation over a 48 hour period, and for each period of constant bandwidth analyse at the end of that period. At some point in time we would make the decision to increase the bandwidth; for example the packet loss reaches a predetermined level. We would like to know the impact of the latency between that decision being made and the new bandwidth coming available. We approach this problem in two ways. The first is to construct an offline algorithm that has full knowledge of required bandwidth, contained in Section 3.6.1. The second is an online algorithm that only has knowledge of past and present bandwidth levels; this is presented in Section 3.6.2.

Given the elastic nature of most Internet traffic it is likely the traffic would partially adapt to the bandwidth limits, thus the traffic trace would change under these conditions. However since the packet loss scales back traffic it is likely that the predicted loss and actual loss would be fairly close, especially with a small target loss where few packets are lost in a single burst. This is due to the high level of aggregation present: if we assume that we have many adaptive flows then each will only take a small proportion of the total bandwidth, so if only a small number of packets are lost then it will be only the affected flows that adapt and decrease transmission rate. As the target loss increases, this assumption becomes less valid, and relating the actual loss to throughput of TCP connections is known to be a very hard problem [Sahu99].

Further issues with the assumptions made in this section regarding elastic traffic are discussed in Chapter 6.

### 3.6.1 Offline Allocation Algorithm

We first construct a bandwidth allocation graph using an offline algorithm. This is not realistic of a network in operation, but it will also help us determine quantitatively a bound on the relationship between granularity of allocation and frequency of reallocation discussed qualitatively in Section 3.5.3. Our algorithm has a target data loss which is incurred by allocating bandwidth to our trace at a given level of granularity. It attempts to meet that target loss on average over the full 48 hour period, and jointly to minimise both the number of changes in allocation needed and the average bandwidth allocated. For practical application to an optical network we are more concerned with the maximum rate of allocation changes — the minimum length of time between two changes — rather than the average rate of allocation change, however this difference will be addressed in the online algorithm.

The algorithm works in three main stages. The first stage allocates bandwidth optimally given the granularity: each time period is allocated the lowest multiple of the allocation granularity necessary to achieve no loss. Due to the bursty nature of traffic, this allocation trace will contain many changes of bandwidth allocation.

The second and third stages try to reduce the number of allocation changes whilst keeping the total loss below the target loss and limiting the increase in allocated bandwidth. The second stage removes short 'dips' — where the allocated bandwidth decreases then increases back after a short period of time. These dips are identified, and then ordered according to how much extra allocation is needed to remove the dip — the area of the dip. The smallest is removed, and then this process is repeated. This continues until the average allocated bandwidth has increased by a set ratio — $\epsilon$, the expansion ratio. Since we are increasing bandwidth no data loss has occurred by the end of this stage. The value of this ratio $\epsilon$ has no direct meaning in the network; it is just an internal parameter of the algorithm that controls the trade-off between the frequency of bandwidth allocation changes and the bandwidth efficiency of the solution.

The third stage takes advantage of the first two no-loss stages by removing unnecessary 'humps' — where the allocated bandwidth increases then decreases after a short period of time. This has the effect of reducing the number of changes, reducing the average allocated bandwidth, and increasing the loss. Since each hump removed decreases the number of changes by at least two, we wish to remove as many humps as possible before reaching our target loss. Therefore we order each hump found by the contribution to data loss from that hump being removed, and remove the smallest. This process continues until the loss reaches the target loss.

It is likely that our solution is not optimal, since the second two stages are both greedy — removing the dips or humps ordered by the smallest bandwidth or loss gain respectively — and separated into two stages. However, the exhaustive search takes an impractical length of time to run, especially for a large number of bandwidth changes, so we assume that this algorithm produces the best results relative to a practical computation time.

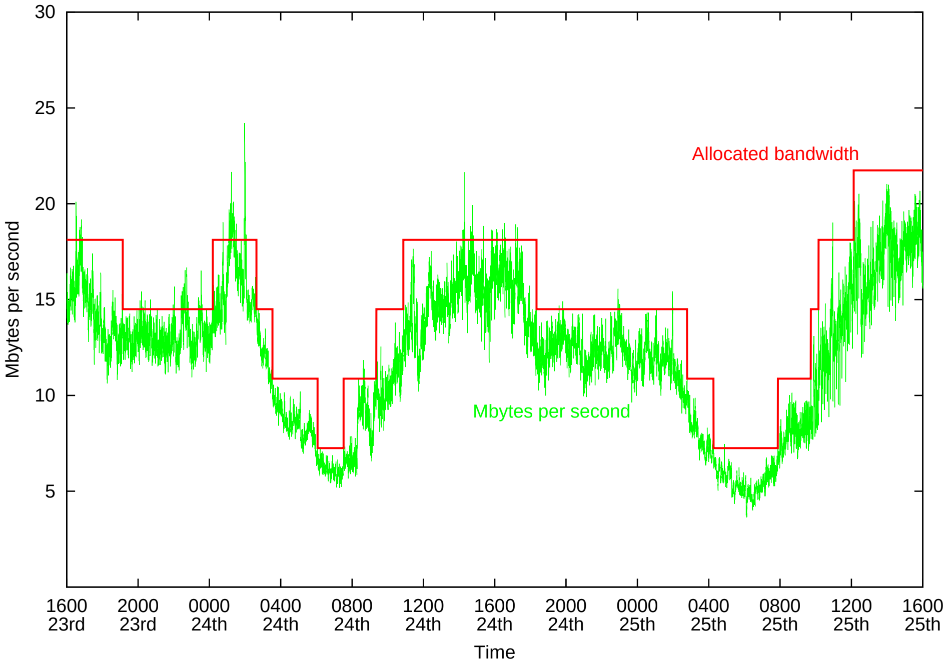

FIGURE 3.14: EXAMPLE OF BANDWIDTH ALLOCATION, TARGET LOSS 0.001 AND GRANULARITY OF ALLOCATION 15% OF MAXIMUM OBSERVED TRAFFIC

An example of this shown in Figure 3.14. Here the granularity of allocation was 15% of the peak traffic level, and the target loss was 0.001; the actual loss after the third stage of the algorithm was 0.00111. The expansion ratio, $\epsilon$, was 1.3; the increase in allocated bandwidth of 30% during the second stage was compensated by the third stage — the net increase in average allocated bandwidth over the last two stages was around 6.8%. There were 16 changes of bandwidth, reduced from 1669 after the first stage. Over 48 hours this corresponds to an average of one change every 180 minutes. However, as can be seen, there is a wide distribution of the time between changes. Without a rigorous statistical analysis, it seems that in this example allocation changes to the allocated bandwidth are the results of diurnal non-stationarities. A different point on the trade-off between average allocated bandwidth and the number of changes shows that changes are made during apparently stationary periods; the lower average allocated bandwidth is achieved by reacting to the long range dependent nature of stationary traffic.

The effect of the expansion ratio is to trade off average bandwidth allocated and the number of changes, for a given target loss and granularity of allocation. This effect is shown in Figure 3.15, where the target loss and bandwidth granularity were kept to the previous figures of 0.001% and 15% respectively. Here $\epsilon$ was varied, creating multiple solution points. These points are indicated on the graph, with the solution shown in Figure 3.14 indicated. The actual values of $\epsilon$ used are not relevant, since it is just an internal tuning parameter for our algorithm. Changing its value allows us to alter the trade-off between the frequency of allocation change and the average bandwidth allocation, to see which combination of values is achievable. In Figure 3.15

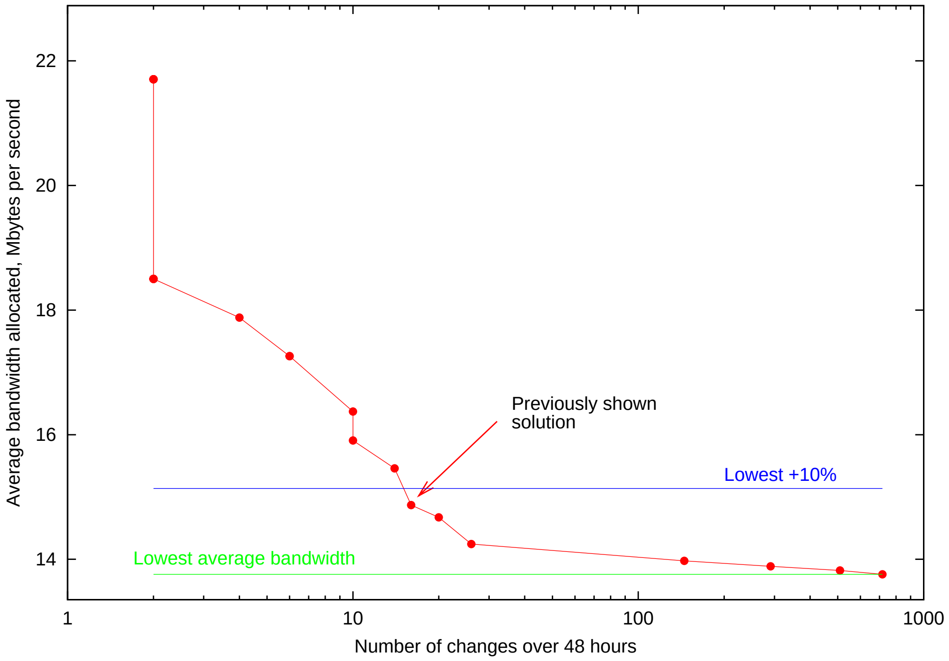

FIGURE 3.15: THE TRADE-OFF BETWEEN THE NUMBER OF CHANGES OVER A 48 HOUR PERIOD AND THE AVERAGE BANDWIDTH ALLOCATED

the bottom left hand corner is the optimal region, a lower average bandwidth with minimum allocation changes. The results presented show that we cannot achieve the combinations below the curved line.

The point on this curve that we pick would depend on the comparison in cost between bandwidth and implementing a network capable of altering allocated demand frequently. We note that the right hand side of the graph indicates that as we try to decrease the average bandwidth allocation closer to the minimum possible value the number of changes increases rapidly. Intuitively this is due to having to allocate on the timescale of bursty stationary traffic, rather than just picking out non-stationary events.

The effect of a larger allocated bandwidth is to reduce the utilisation of the link; an allocation of 21 Mbyte/s corresponds to an average utilisation of 57%, while 13 Mbyte/s corresponds to 90% utilisation. A current IP backbone provider has only 10% of links with a peak utilisation of over 50%, calculated over a 5 minute time period [Fraleigh03]. This limit is equivalent to an allocation of 19 Mbyte/s.

The time period used to calculate peak utilisation is critical, since a long period will underestimate utilisation since the short timescale bursts which cause packet loss with be averaged out. To see the effects of changing the averaging period we recalculate the results shown in Figure 3.15 using three different measurement periods: 100 ms, 1 s, and 10 s. The results are shown in Figure 3.16. As the period decreases more loss is observed since the traffic appears more bursty, so the average bandwidth rises to

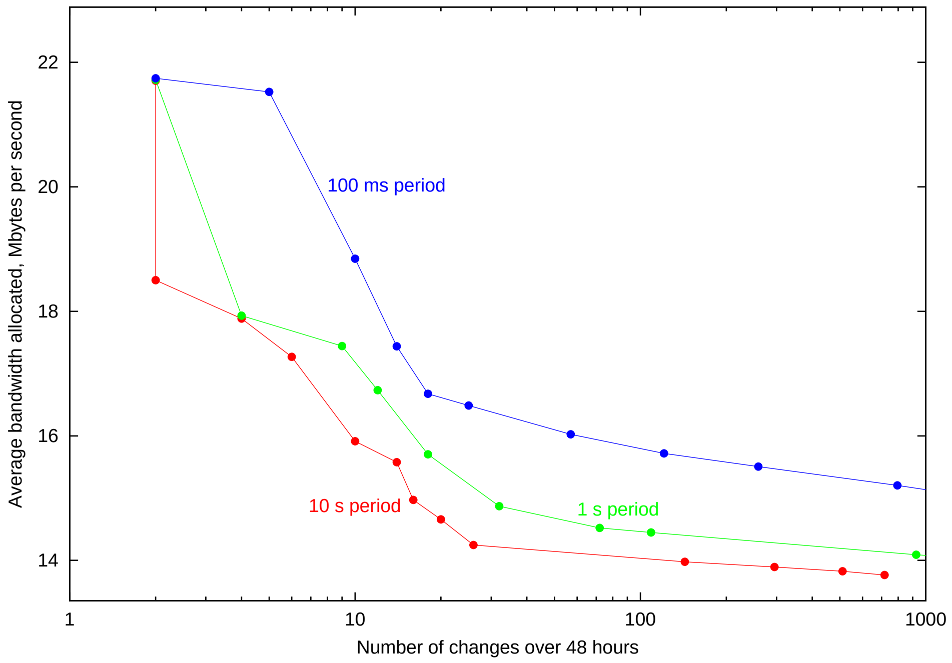

FIGURE 3.16: THE TRADE-OFF BETWEEN THE NUMBER OF CHANGES OVER A 48 HOUR PERIOD AND THE AVERAGE BANDWIDTH ALLOCATED, CALCULATED USING THREE DIFFERENT AVERAGING PERIODS

achieve the same loss target. To make the target loss equate to actual performance the correct timescale to compute over would be close the time taken to completely fill the packet buffer at the ingress of the link, likely to be much less than one second. In a practical implementation it is likely that actual loss statistics will be available so this difference is not a problem.

For the following analysis we use a 10 s averaging period, partly since with fewer records, collecting many results is more efficient. However it is not expected that this will substantially affect the nature of our findings; this is confirmed by initial results not reported here.

We now investigate the effect that changing the granularity of allocation has on the frequency of allocation. To select the point on the previous trade-off between utilisation and frequency of allocation we pick the point with the lowest number of changes which increases bandwidth allocated by less than 10% from the lowest possible utilisation. This lowest utilisation point is the most bandwidth efficient solution found, where the second stage of our algorithm is skipped. The chosen solution corresponds to the indicated position expansion ratio of 1.3 on Figure 3.15, as shown by the two lines drawn, the first at the minimum possible level, and the second at a 10% increase.

Figure 3.17 shows how the number of changes is linked to the granularity of allocation and the target loss. The number of changes corresponding to a change every 20 minutes and a change every 8 hours are also shown as horizontal lines. The average band-

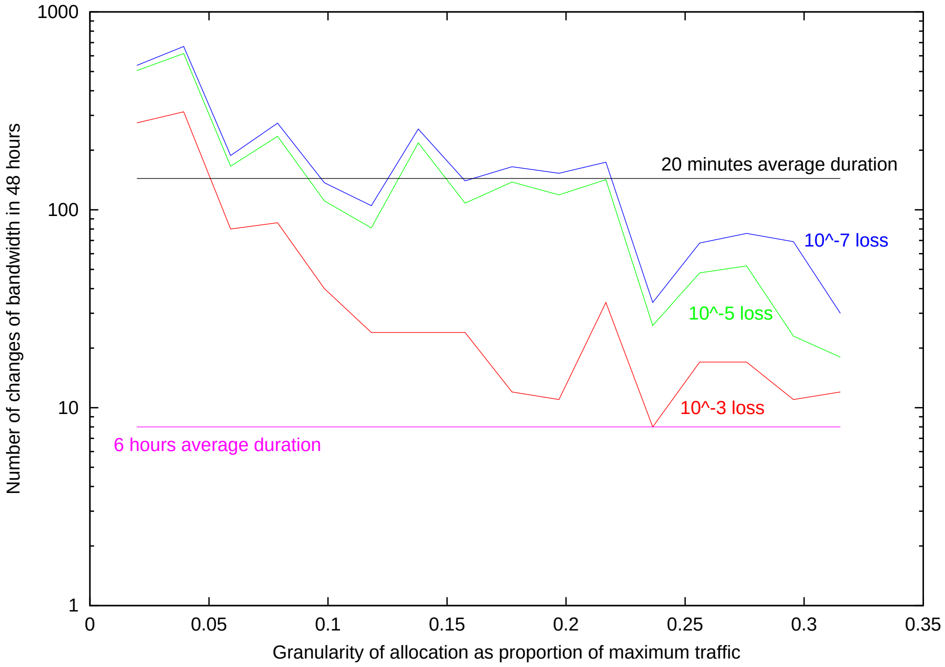

FIGURE 3.17: THE NUMBER OF ALLOCATION CHANGES DURING A 48 HOUR PERIOD DUE TO CHANGING THE GRANULARITY OF ALLOCATION AND THE TARGET LOSS

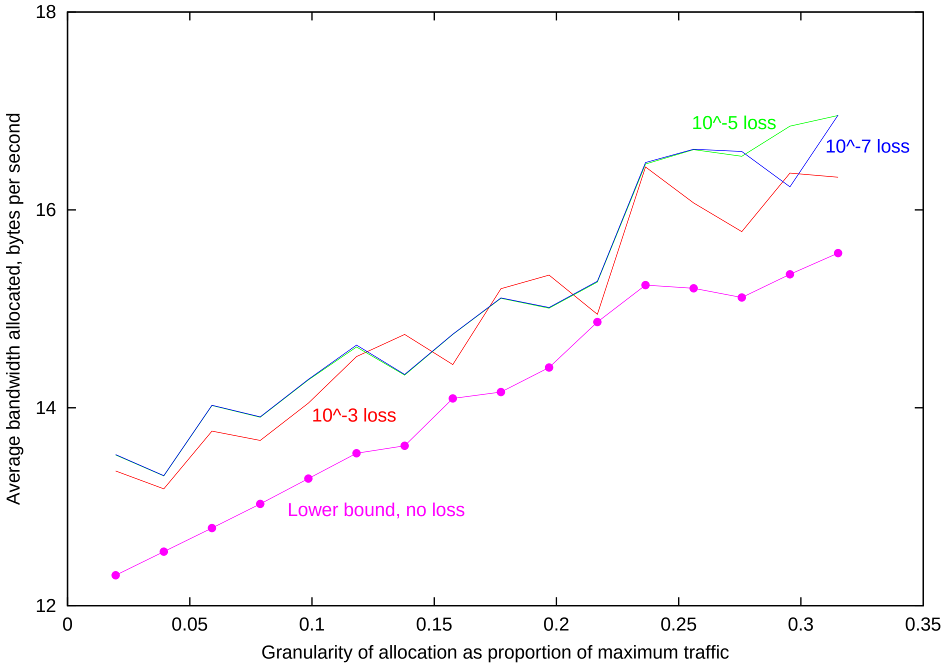

FIGURE 3.18: THE AVERAGE BANDWIDTH ALLOCATED OVER THE SAME 48 HOUR PERIOD DUE TO CHANGING THE GRANULARITY OF ALLOCATION AND THE TARGET LOSS

width needed by the same set of experiments is shown in Figure 3.18. This figure also shows the lower bound on the average bandwidth, computed by taking each bandwidth figure and allocating the next higher available bandwidth for that period. This corresponds to up to 17,000 changes of bandwidth, since these experiments used the measured bandwidth trace averaged over 10 s periods.

Due to the selection of the expansion ratio for each combination of granularity and loss ratio the bandwidth required is approximately 10% more than the minimum, reduced by a factor that increases with the loss ratio. As the granularity becomes coarser this minimum level increases approximately linearly. This is intuitively reasonable, since over the region investigated we waste on average half of the last granularity's worth of allocation; as the granularity increases the magnitude of this fraction grows proportionally.

With our choice of near minimal bandwidth, as the granularity becomes coarser we need to change less frequently to achieve the same loss ratio. However, the rate of decrease slows dramatically as the granularity increases; note that Figure 3.17 is log scaled on the y-axis. The number of changes can be approximated by the inverse of the granularity. This inverse relationship is possibly related to the shape of Figure 3.12, where the time lag is linear with the granularity of measurement; the number of changes is 48 hours divided by the time between changes giving rise to the inverse relationship.

All the experiments show that on average a network operator would change bandwidth allocation more frequently than every 6 hours, which supports earlier evidence. At the other end of the scale, it is likely that the benefits of reducing the required bandwidth by a linear amount are outweighed by the large increase in the frequency of change. The line marking 20 minute average duration is likely to be near the boundary of this point for low target losses.

To achieve a low loss target the average utilisation decreases to around 50%. To achieve similar utilisations to those found in current backbone networks this algorithm would have to change to be based on a target utilisation rather than target loss, since the degree of loss would be insufficient. Since this dissertation is primarily interested in maximising use of limited capacity, this utilisation-based algorithm is not considered further.

### 3.6.2 Online Allocation Algorithm

While the algorithm developed in Section 3.6 may be a bound on the trade-off between the number of changes and the average allocated bandwidth, it is an offline algorithm where all future traffic is known. For practical networks an online algorithm that can react to currently observed traffic is needed. This might operate by tracking the current traffic with some form of damping to reduce the number of temporary oscillations. With a steadily increasing volume of traffic, once the decision has been made to increase the allocated bandwidth the latency of allocation is important. If setting up a new path takes a considerable length of time then we risk a high level of data loss during this waiting period. This could be compensated by having a more sensitive trigger to increasing bandwidth; faced with the same increasing volume of traffic we would then request an increase in allocated bandwidth sooner. However this would then increase the number of changes as we react to shorter timescale bursts.

Both the maximum rate of allocation changes and the average rate of allocation change are important here. The maximum rate of change is controlled by our latency figure; we will not be able to try to reconfigure the optical network faster than this latency allows. However since we will have several of these algorithms running simultaneously, one for each pair of IP-aware nodes, the average rate of change will help identify the speed at which the control layer has to reconfigure across the whole network. This is because the scope for many reconfigurations happening concurrently may be limited. This is discussed in more detail in Section 6.3.3.

We present here an online algorithm for performing bandwidth allocation where no future knowledge of traffic is known. Given a granularity of allocation and a target data loss ratio, $r$, we proceed as follows. For each possible bandwidth allocation we simultaneously calculate the exponentially weighted moving average of the loss incurred in each time period at that service rate, using a filter constant $f$. Values of $f$ close to 1 indicate that our loss estimate moves slowly, maintaining past history, whilst values of $f$ close to 0 indicate that our loss estimate closely follows the measured loss for each time period. We introduce two constants, $\alpha$ and $\beta$ which control the trigger points for changing bandwidth allocation. If the loss for the current allocated bandwidth is $\phi$ then we decide to increase allocation if $\phi > \alpha r$, or decrease allocation if $\phi < \beta r$, where $\alpha > 1, \beta < 1$. When changing to a new allocation we pick the smallest allocation that has an exponentially weighted loss smaller than our target loss.

The values for $f$, $\alpha$, and $\beta$ will affect the trade-off between achieving a high utilisation and reducing the number of bandwidth changes required. We give suitable ranges for these parameters below.

Since the algorithm presented above tends to yield a conservative actual loss ratio, we amend the algorithm by replacing the target loss, $r$, with $r'$. We calculate the difference between the target loss with the actual loss observed so far; calculating our new target loss, $r'$, to compensate for this difference over a fixed time period. The compensating time period used in these experiments was 100 minutes; if $r'$ loss was achieved over the next 100 minutes, the total loss from the start of the experiment to that point would be $r$.

There is also an allocation delay where any decision to change the allocated bandwidth takes a certain amount of time to process. During this time we continue to monitor and update our estimated loss, but cannot schedule another request to change our allocation. The model used here is that our analysis engine has a network control interface which it can request changes to allocation, but has to make synchronous calls to that interface. Due to the power equalisation effects mentioned in Section 2.3.2 this

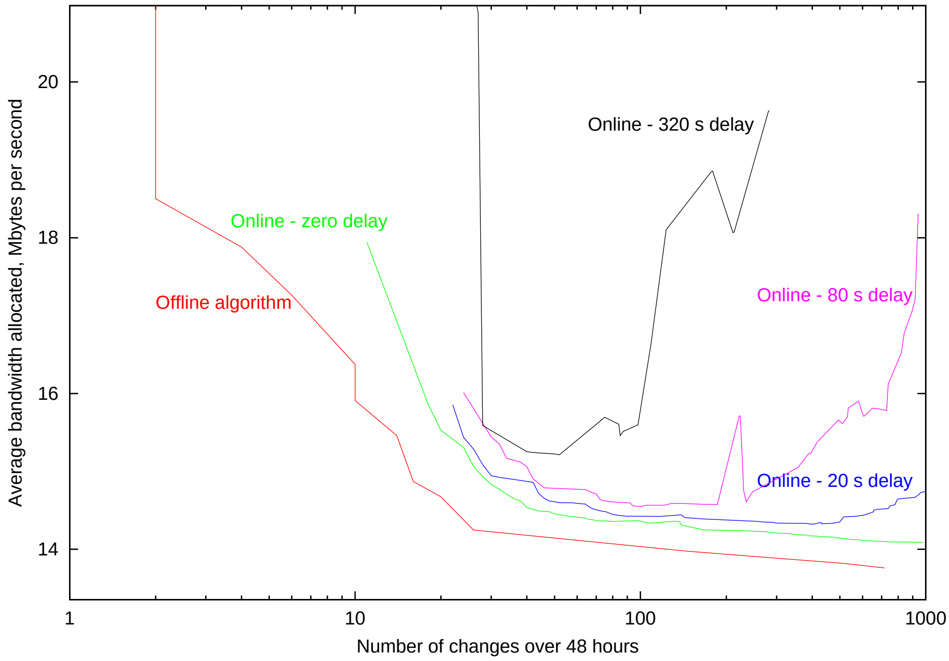

FIGURE 3.19: THE TRADE-OFF BETWEEN THE NUMBER OF CHANGES OVER A 48 HOUR PERIOD AND THE AVERAGE BANDWIDTH ALLOCATED, SHOWING THE DIFFERENCE BETWEEN OFFLINE AND ONLINE ALGORITHMS WITH A RANGE OF ALLOCATION DELAYS

allocation delay may be considerable, perhaps a few minutes.

Using the same granularity of allocation and target loss parameters as presented in Figure 3.15, for each allocation delay we experiment with our three parameters, $f$, $\alpha$, and $\beta$. In Figure 3.19 we show the previous results from the offline algorithm, together with the equivalent curves for several different allocation delays produced by the online algorithm presented above. For each allocation delay we present the results with the lowest average allocation bandwidth for a given number of allocation changes in 48 hours.

To achieve a low number of changes we make our algorithm more stable by selecting high values for $f$ of around 0.999; this produces the leftmost points on each curve. To achieve a better bandwidth efficiency using more changes we select a low value of $f$ of between 0.2 and 0.5; this produces the rightmost points on each curve. Adjusting $\alpha$ and $\beta$ has much less effect than changing $f$; in the results presented $\beta$ was fixed at 0.95, and $\alpha$ was either 1.5, 3, or 5. The larger values for $\alpha$ were used more towards the left hand end of the curve, the smaller values more towards the right. Almost identical results can be gained using a fixed value of $\alpha = 3$, but we lose a few of the leftmost points on each curve. When used with unknown traffic in a real network, the correct value for $f$ for the frequency of changing allocation will depend on the exact characteristics of the traffic — how bursty the traffic is. With a target frequency of change an adaptive algorithm to vary $f$ depending on past performance would be necessary. The experience gained in these experiments would indicate that $\alpha$ and $\beta$

would not need to change from the values of $\alpha = 3, \beta = 0.95$ used here.

Figure 3.19 shows that with zero allocation delay we can obtain results that require 3-5% more bandwidth for the same number of allocation changes at the right hand side of the graph, and 25-50% more changes at the left hand side of the graph. The major difference in moving from offline to online analysis is when trying for only a few changes; there are large bandwidth efficiency gains in knowing the future bandwidth requirements, given the limit on data loss ratio. With an online algorithm we have to guess and set a level either too high giving poor efficiency, or too low and hence have to change allocation frequently to avoid large levels of loss.

As we increase the allocation delay — the time between deciding to change allocation bandwidth and that bandwidth becoming available — we notice two effects. The major effect is that the right hand end of the graph curves upward, as it becomes progressively harder to find efficient solutions that change allocation frequently. Bandwidth efficient solutions must be able to react to small timescale changes in observed bandwidth, but with a high time lag this becomes increasingly impossible since the traffic will have changed to a significant degree over the delay period. The smaller effect is that at a very high delay the left hand end also shifts upwards, indicating that it becomes harder still to find efficient allocations using few changes.

The area affected least by imposing an allocation delay is the region investigated previously, at the knee of the curve with a few tens of changes and moderate bandwidth efficiency. This region would be the only efficient operating point for a high delay system. In contrast, if we wished to operate at any other point we would require minimal allocation delay.

### 3.6.3 Discussion

To operate in an optical network we would need to monitor the aggregate traffic from the ingress point to every egress point: between every pair of IPONs in Figure 1.2. This functionality is already standard in most high-capacity routers: the most modern has more than 250,000 fully programmable counters which can classify packets based on any header field [Procket03]. As outlined in Section 1.7 each OXC connected to an IPON will export a virtual interface to that IPON for each destination point in the optical network. The control layer in that OXC will then keep the necessary statistics to operate the bandwidth allocation algorithm for that virtual interface. Over time that interface may create several wavelength routes to cope with the traffic levels. This is covered in more detail in Chapter 6.

Other methods to find the aggregate cross-network traffic matrix have been developed which correlate traffic measurements throughout the network with synchronised routeing and topology data [Medina02, Papagiannaki03]. Both currently use off-line analysis and may not provide enough information about current loss, although for an utilisation-based algorithm they might be suitable.

Whilst some regions of the results of the online algorithm presented are close to the results for the offline algorithm, it is not clear that the algorithm presented is necessarily the best online algorithm. For example knowing that this particular traffic follows a daily cycle could enable an online algorithm to bet against the future, knowing probable future movements. Alternatively, gathering more statistics about past history, including mean, variance, and self-similar categorisation, would allow more informed decisions about how to change bandwidth allocation. Another area of investigation would be to use some form of machine learning to remove any bias from imposing an overly simplified model on traffic aggregate measurements.

The major advantages of the algorithm presented are that it requires only the aggregate offered load for this link, it stores very little state, and it is computationally simple. As a minimum it is a proof of concept that online bandwidth allocation is feasible and in the correct region can operate successfully under considerable allocation delay.

## 3.7 Traffic Multiplexing

As traffic streams multiplex together towards the core of the Internet, we need to know the characteristics of these large streams. The analysis contained in this chapter focuses on one set of measurements taken at the uplink of the University of Cambridge. However, future networks will see streams such as this multiplexed many times over. If results of the analysis performed are to be valid for the future core network, we need qualitatively to assess the difference between one such stream and the multiplexed traffic.

We broadly divide this into three separate timescales: stationary periods of traffic, non-stationary period, and routeing changes. On the smallest timescale all the component streams in our multiplexed traffic maintain the same statistical characteristics. On a long timescale these component streams may change, typically exhibiting a diurnal cycle. Finally, the set of component streams may change as some follow an alternate route.

### 3.7.1 Stationary Traffic

Multiplexing together streams that exhibit long range dependence maintains the long range dependence present, but reduces the standard deviation to mean ratio by one over the square root of the multiplexing factor [Cao01a, Cao02]. These studies also show that this is maintained in the presence of link saturation. Thus the analysis of data in this section is applicable to this traffic multiplex. This is broadly seen on the smaller scale where we split the traffic observed by looking at possible route aggregation in Section 3.4. Each smaller trace is typically more bursty than the multiplex, but long range dependence is preserved.

### 3.7.2 Non-stationary Traffic

Moving on to timescales that exhibit non-stationarity, the major question is whether or not these non-stationarities are independent amongst the component traffic streams. To understand this we need to know the causes of non-stationarity. We can divide traffic into two broad classes: user-driven and automated. User-driven traffic, for example web browsing, will be driven directly by user profiles; the daily cycle of people being awake, at work, or at home, which drives the diurnal, weekly, and even yearly cycles. Automated traffic is likely either to be fairly constant or deliberately shifted away from busy periods to take advantage of lower network congestion, so it can be ignored for the purposes of non-stationary analysis as it is also driven implicitly by user behaviour.

For example if we were to multiplex traffic from all UK universities it is likely that, since the users generating the traffic have the similar working profiles, non-stationary events will be correlated; for example the rise in traffic during the morning and a sustained peak towards midnight. However smaller differences may still emerge between different universities depending on, for example, their timetabling of lectures. Moving to a more diverse mix, we might find that universities and businesses in the UK have some correlation since they are in the same time zone, but that businesses are likely to have an earlier peak in the morning and lower traffic from 6pm onwards. Moving wider again, we may find that comparing traces of demand originating in the UK and California that both are similar in form but one is shifted in time due to the eight hour time zone difference.

If streams are correlated then we can expect a similar, but enlarged, form for the trace analysed in this chapter. If streams are completely uncorrelated then we could assume that if one trace experiences an increase in activity then others might be increasing or decreasing: in one case the changes cancel and in the other they reinforce each other. As the number of traces in the multiplex increases then on average the number of significant non-stationary events will decrease. However this is unlikely, since any multiplex of traffic is unlikely to be uncorrelated. Since we are multiplexing this traffic this implies that these traffic streams have some portion of their path in common, implying that they could be closely related. For instance we might group traffic from UK universities to California, which would exhibit clear correlation patterns.

### 3.7.3 Routeing Changes

While the majority of routes through the Internet have a prevalent or most common path, it has been shown that changes in the path used happen over several timescales [Paxson97a]. More recent work shows that the number of BGP updates which change the AS paths for a route was measured at 13,586[8] per day [Labovitz99], and a significant number of announcements involve a change of entry point into an AS.

[8] This figure is a small proportion of the total number of BGP updates measured, reported as 'several hundred thousand per day', since many updates are pathological.

The first of these would indicate that whole traffic streams will appear and disappear from the intermediate AS that is being added or removed from the AS path; the second indicates that a traffic stream travelling through a single AS will have the ingress into or egress out of that AS changed.

These events are most likely to be caused by the failure in connectivity of another AS, causing traffic to be diverted through this AS as BGP recalculates the best AS-path for this traffic stream. These events may be less frequent than the stationary and non-stationary patterns mentioned above, but may cause a larger individual effect.

## 3.8 Conclusion

In this chapter we have analysed a 48 hour trace of Internet data arriving and leaving a single university that has around 10,000 staff and 10,000 students. We have shown this data to have long range dependence and possible self-similarity at longer timescales. Whether the exact statistical model behind shifts in traffic is due to self-similarity or non-stationarity is less important than the fact that traffic levels do shift a significant amount over timescales of several hours.

An offline algorithm for finding a bandwidth allocation graph was described, allowing a target loss to be set, and the trade-off between average allocated bandwidth and the number of allocation changes to be made. The 48 hour trace was investigated and when limiting the allocated bandwidth to the minimum possible value for a given granularity plus 10%, we obtained a requirement to change the allocated bandwidth between every 20 minutes and every 6 hours. This figure is affected by the data loss incurred, but mainly dependent on the granularity of allocation: how precisely we can select bandwidth. As we move to a coarse grained allocation the number of changes decreases, but at the cost of average bandwidth increasing. For example with a low loss rate of $10^{-7}$ if we move from being able to allocate a tenth of peak traffic rate to allocating a third of peak rate, the number of changes decreases by a factor of four, but the average bandwidth increases by 20%.

We have developed an online algorithm to perform the same task, and investigated the effect of allocation delay — the delay setting up or tearing down existing routes in our network. As delay increases, solutions with many changes become much less bandwidth efficient, and the minimum number of changes required increases. For longer delays an operating region of around 10% bandwidth inefficiency and changes on average every 1-2 hours was identified. For other operating regions, that are either more bandwidth efficient or require fewer changes, the allocation delay would need to be reduced, to less than a few tens of seconds.

The analysis conducted in this chapter was performed on a single measurement trace. It would be essential to replicate this analysis on traces taken from a different time or point in the network to be able to assess how universal the results gained in this chapter are.

The point chosen on this trade-off between granularity, bandwidth efficiency, loss, and frequency of change depends on many factors. The capability and cost of technological solutions both for the data transfer and the control plane set the granularity and frequency of network updates respectively. The network operator might offer differentiated services, with customers paying more for a lower loss rate, which requires either more average bandwidth or more frequent allocation changes.

We have argued that features of the traffic sample used may be representative of future network traffic. This is on the basis that over short stationary time periods the self similar nature of traffic will be maintained. For longer time periods it is quite likely that some traffic multiplexes will continue to exhibit non-stationary behaviour, which vary the total observed traffic on a similar time scale to that identified as a valid operating region for the online bandwidth allocation algorithm.

This chapter has shown that when considering a single traffic stream we can achieve a similar loss ratio whilst having a lower average bandwidth allocation by adjusting the bandwidth available. The assumption is that with multiple streams crossing the network, traffic shifts will emerge between different streams, allowing resources used by the falling stream to be used to cope with demand from the rising stream.

The next chapter identifies the possible network architecture which would support this reconfiguring of bandwidth amongst competing streams, together with a suitable routeing algorithm. We also discuss the setup of the simulator used to experiment with these proposals.

# Chapter 4

# Architecture and Experimental Setup

In the previous chapter real network traffic was examined, and it was shown that an online algorithm is capable of indicating when the bandwidth allocated to a traffic stream should be changed. This algorithm can control the trade-off between the frequency of changes in allocation and the bandwidth utilisation efficiency of the traffic stream. In the reconfigurable optical network changing allocation corresponds to adding or removing a wavelength route across the network. This can lead to substantial efficiency gains over peak rate allocation depending on the frequency of reconfiguration and the magnitude of traffic change relative to the bandwidth of a wavelength.

In this chapter we look at the optical network layer, and how that network could support reconfiguration. We also describe the design of a network simulator which is able to provide quantitative analysis of different network architectures.

## 4.1 Introduction

The current approach of most network operators who run WDM networks using optical switching is to have a fixed set of virtual paths through the network. Revenue is made from either offering Virtual Private Networks (VPNs) across the network to companies or resellers, or by peering with other networks and agreeing to exchange a given level of traffic. Since both are led by human negotiation these tend to change network configuration over a generally slow timescale of days, weeks or even months. It is common practice that optical paths, once placed in the network, are fixed and are not removed [Strand01]. Routes are fixed since with currently deployed technologies altering a route causes a drop in connectivity, which at high line rates gives high data loss. The main focus of network design is a long term periodic upgrading of network resources to meet predicted traffic growth for a number of months or years. The problem here is to minimise the cost of supporting a fixed known set of traffic requirements. We call this the *fixed demand* problem. The process of solving this problem by some method can be called *provisioning*.

Future optical networks are likely to reconfigure over a shorter timescale than at present. A network that frequently alters the bandwidth available between different source-destination pairs can then track traffic shifts and thus give better performance. The focus shifts here from a provisioning stage with fixed demand to one where we have a fixed network topology and wish to place as many advantageous routes as possible. We call this the *fixed topology* problem. The set of routes that we wish our network to support changes over time due to changing traffic conditions or other stimuli. We also refer to this process as *dynamic demand*.

Considering the approach of a network operator, it is clear that there are three major timescales of change involved. The shortest involves some form of packet based multiplexing and routeing of data: a timescale close to transmission time and line bandwidth rates — less than one millisecond. This takes place either hop-by-hop for a packet switched network, or at the network ingress for a circuit switched network. The next timescale involves some form of load balancing between physical links, aimed at reducing the network loss through overload present at the previous timescale. With modern technology this is likely to be done over a shorter timescale than that at which we can add new physical resources to the network, so it could involve changing some default paths inside routeing tables for a packet switched network or altering the existing set of circuits in a circuit switched network. As shown in the previous chapter this can occur over a range of timescales, depending on the desired bandwidth utilisation efficiency. This is an example of the fixed topology problem. The final timescale involved would be where the network could be physically altered — for example new fibres being added to perennially congested links, or migration to a new network technology. Here we are likely to have details of current network usage and predicted usage over the next time period: this is the fixed demand problem.

We are interested primarily in the fixed topology problem. However to investigate this practically via simulation requires a method to create our fixed topology with which to experiment. One method commonly used is to set a fixed demand problem and use the resulting network topology. There are two issues inherent in this approach. Firstly there is the question of the demand set used to provision the network. Possibly the most realistic technique is to use the predicted demand that will be experienced during the dynamic demand phase of the experiment. This would correspond to a network being built using a completely accurate prediction of future traffic. However, as seen in Chapter 3, traffic levels alter considerably over time, both as random samples from a statistical distribution and also in reality since that distribution will be non-stationary. It is also likely that predictions of future traffic will not always be accurate; while accurate measurements of link bandwidths can give estimates of future bandwidth, changes in the general trend can lead to large errors [Papagiannaki03]. Thus we should take care to evaluate our fixed topology solution with a variety of traffic conditions that are not just the same as those used for provisioning.

The second issue is that the most efficient solution to the fixed demand problem might not be the most effective provisioned network for the fixed topology problem. For the fixed demand problem we emphasise maximising utilisation and keeping routes short; for the fixed topology problem we need to have a flexible network that can handle a large range of different sets of demand.[9]

This chapter will consider the architecture of an optical network and outline areas of interest in designing and managing the network. The concept of *wavebands*, a method of partial route aggregation, is introduced in Section 4.2.1. Different methods of protecting against failure are covered in Section 4.2.2, together with the implications on these methods of using wavebands. Section 4.2.3 considers the high level design of a network node by focusing on the flexibility of routeing allowed in a dynamic network. Possible routeing algorithms suitable for a network reconfiguring to dynamic demand are examined in Section 4.2.4. Finally in this chapter, Section 4.3 looks at how to model optical networks at this level so as to simulate network operation.

## 4.2 Network Framework

We assume that we have a network consisting of nodes, a subset of which source and sink data. This subset may include all the nodes, or only a partial selection. Nodes are connected together in a fixed mesh-like pattern, where if two nodes are connected then they may have a number of fibres physically running from node to node. We assume that the connections form a single fully connected network; a subset of the connections between nodes forms a spanning tree. Moreover to allow route-based protection the network must be biconnected: the network must contain two disjoint paths connecting every pair of nodes.

Traffic has been observed be non-symmetric on a single link due to the inherent non-symmetric nature of many applications and the occurrence of hot-potato routeing, where the end-to-end routes used by the forward and reverse traffic flows are often different [Fraleigh03]. It is not clear for a major AS backbone network how asymmetric cross-network traffic will be, and how this may change as the size and external connectivity of that network increases. For simplicity we assume here that traffic demands are symmetric, and that reverse routes will be routed symmetrically. This is due to the lack of proper motivation for a non-symmetric traffic matrix, and not through any assumptions in the architecture or routeing algorithm design, or shortcomings in the simulator implemented as part of this work.

We assume that each fibre contains the same set of wavelengths. This is also for simplicity but in addition because the network is likely to be dealing with a consistently high level of aggregation and hence using the same optical transmission technology throughout the network. We also assume that all wavelengths can be used equally without any restriction from crosstalk imposed by optical switches or other transmission effects. Logically we therefore assume that lightpaths formed from several fibres can also be used equally, apart from where the propagation delay determined by total

[9] The continuous process of dynamic demand could be viewed as a series of fixed points, each of which has a constant set of traffic demands for some period of time

| ATM | Optical |
|---|---|
| virtual path | waveband |
| virtual path connection | (waveband) path |
| virtual channel | wavelength |
| virtual channel connection | (wavelength) route |

TABLE 4.1: ANALOGY OF TERMINOLOGY

lightpath length is critical.

### 4.2.1 Wavebands

On a single fibre we have a number of wavelengths each carrying the same bandwidth; wavelengths are identical across all fibres. We would like to measure the possible benefits from using a two-stage multiplexing process at every node, as first discussed in Section 2.3.3. We divide the set of wavelengths into disjoint sets, each of which is called a *waveband*, to correspond with the output from the first multiplexing stage. This splits a fibre into wavebands; the second stage will convert each waveband into wavelengths. Each set is typically contiguous, although this is not assumed unless otherwise stated.

This approach allows potential benefits since if we can place routes that have a common subsection in a single waveband we can switch this waveband throughout this subsection without having to split it into wavelengths. This will reduce the number of elements that need to be switched at the intermediate nodes. Whilst work has been done to develop a heuristic algorithm using shortest path routeing for the fixed-demand problem [Lee02], the effects of wavebands on using alternative path routeing or the fixed-topology problem have not been previously addressed.

We use terms analogous to ATM's virtual circuits and virtual paths [Kalmanek02]; these terms are shown in Table 4.1. At a network node, we demultiplex incoming fibres into wavebands. We may switch these directly to form a *waveband path*: a series of physical hops through which we switch this waveband without further demultiplexing into wavelengths. This waveband path is analogous to a virtual path connection. At the endpoint of this waveband path we demultiplex it further into its constituent wavelengths. These wavelengths may either terminate here, or be switched to the start of an outgoing waveband path. Hence multiple waveband paths may be traversed to form an end-to-end *wavelength route* through the network; this is our virtual channel connection analogy. We use *path* to refer to a waveband path and *route* to refer to a wavelength route.

### 4.2.2 Protection

When routes are set up they may have a protection parameter that describes what form of protection, such as dedicated or shared protection, should be employed. In the event of failure of a network element, to regain connectivity the affected routes will have to be re-routed across the new network topology. There are three main options as to how to achieve this:

**dedicated protection:** alternative routes are allocated in advance along disjoint paths and relevant switches configured for these backup routes. When failure is discovered traffic can be transferred immediately to this new route, achieving total recovery times in the order of tens of milliseconds at the cost of doubling the effective bandwidth requirement for each protected route. Alternatively the signal may always be dual-fed through both the primary and protection routes, and the signal to be used selected at the end point of the route.

**shared protection:** alternative routes are calculated in advance and the bandwidth for those backup routes is reserved. However, multiple alternative routes may share the same reserved bandwidth if their primary routes are also disjoint. Thus for the failure of any given network element, only one of the primary paths that share this protected bandwidth needs to use its protection path. As the protection route is shared, when failure is discovered the nodes along the protection route need to set up the relevant switching configuration to enable the protection route. Estimated time to recover connectivity may approach the order of hundreds of milliseconds depending on the signalling mechanisms used. However the required bandwidth overhead is reduced from that required by dedicated protection.

**no protection:** in the event of failure alternative routes will have to be found across the network and set up, possibly re-routeing or re-sizing competing paths. This would either require a manual reconfiguration or an automatic system to find a new path.

Thus a network operator may trade off the speed of recovery following a network failure against the total available bandwidth. It is entirely possible that different routes across the network may have different levels of protection. One possible scheme could be to create multiple disjoint routes for every path, depending on the relationship between the network topology and the minimum size of a channel. In the event of a single failure we would experience a decrease in total bandwidth along that path, but not a complete failure. This may be desirable given the elastic nature of TCP traffic, one large probable component of carried traffic. This is a similar concept to that proposed for wireless networks, where the routeing protocol used for each hop uses multiple paths available for use in load balancing and failure recovery [Zaumen98, Vutukury01].

Using shared protection with wavebands, the natural granularity at which to share protection bandwidth is at the waveband level, since only at the end point of each

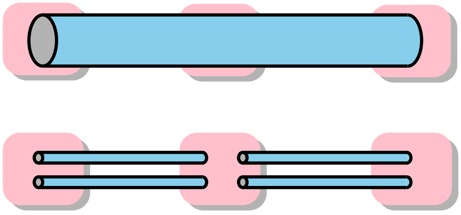

FIGURE 4.1: EXAMPLE OF LIMITED ROUTEING FLEXIBILITY. THE TOP SOLUTION REDUCES NETWORK COST, USING A SINGLE WAVEBAND. THE BOTTOM SOLUTION, USING WAVELENGTHS, COSTS MORE, BUT HAS MORE ROUTEING FLEXIBILITY

waveband path are the individual wavelengths multiplexed out. Therefore if we have a waveband path of several hops, if two protection paths share a wavelength inside that waveband path then they do so for the whole length of the path. This is coarser grained than some sharing schemes, where we allow routes to share a wavelength on a physical hop-by-hop basis, but allows for cost gains as outlined in Section 4.2.3. Given the possible difference in recovery time between dedicated and shared protection, it would be instructive to know the extent of the capacity benefits from using shared protection. Choosing dedicated protection could be warranted if, in a dynamic demand environment using waveband routeing, the benefit of shared protection is small.

### 4.2.3 Waveband Routeing Flexibility

If we want to reduce the installation cost of a network by utilising waveband routeing, then we do so at the penalty of reducing the flexibility of the network. For example if we provision a network with long waveband paths we only require the multiplexing and switching equipment necessary to deal with wavelengths at the ends of each waveband path, reducing the equipment cost. However, we are then limited to routeing with this set of paths, since we cannot access individual wavelengths at arbitrary points along a path. This can be seen in Figure 4.1 where we show two alternatives. The first uses a waveband path to route between two points, bypassing the intermediate node, whereas the second uses wavelength routeing on both fibres. The first will be cheaper, since we remove the switch and second stage multiplexing equipment at the intermediate node, but is less flexible — indeed as shown we cannot reach the intermediate node at all.

Since it is hard to compare a reduction in cost to a reduction in performance, we would like to compare the performance of multiple architecture designs at equal costs. To show that using wavebands in this way is beneficial, our hypothesis is that with appropriate provisioning and routeing algorithms we can overcome the drop in flexibility from waveband routeing to yield better network performance at an equal cost.

We present here examples of several alternative node architectures, each of which corresponds to a different trade-off between dynamic routeing flexibility and installation cost. These correspond to the different cases to be presented in the results of Section 5.2.2. They are presented in order of decreasing cost and decreasing reconfiguration flexibility, starting with the most expensive and flexible network. In Figures 4.2, 4.3, 4.4 and 4.5 traffic flow is depicted left to right; the outer containing box depicts the all-optical network node. The traffic flow into the node is shown on the left hand side of the box, with *electrical input* corresponding to the ingress point into our network at this node, shown by the dotted lines, and *optical fibre input* corresponding to traffic flow from within our optical network arriving at this node. Output is similarly shown on the right hand side; here *electrical output* is the egress from our network, *optical fibre output* carries traffic destined for other network nodes within our optical network. For clarity only data flow paths are shown, with control flow paths omitted.

The solid labelled boxes on the boundary of the node connected to the dotted lines represent opto-electrical conversion units; transmitters on the left and receivers on the right. They are labelled with the wavelength they are tuned to, with wavelengths being called A, B, C and D. In Figures 4.3, 4.4 and 4.5 the node has two sections; in the top section the thin lines carry individual wavelengths, in the bottom section the thick lines carry wavebands. In Figure 4.2 there is no waveband section.

The small filled wedges represent multiplexers and demultiplexers. Incoming and outgoing fibres carry all four wavelengths {A,B,C,D} as marked. In Figures 4.3, 4.4 and 4.5 we see two types of multiplexer. Fibre to waveband demultiplexers split a full fibre into two wavebands, the first containing wavelengths {A,B}, the second containing wavelengths {C,D}. These are seen on the far left and right walls of the node. Waveband to wavelength multiplexers split a waveband into individual wavelengths; these are seen on the internal horizontal wall of the node. In Figure 4.2 the multiplexers shown split full fibres into wavelengths, since this node topology has only single stage multiplexing.

Switches are labelled with the set of wavelengths they switch inside braces. Thus for a wavelength switch marked {A} all ports on the switch contain the single wavelength {A}. In the waveband switch marked {C,D} in Figure 4.3, all ports will contain the waveband {C,D}. Note that both types of switch may be physically identical, since these switches are wavelength agnostic; they are labelled 'wavelength' or 'waveband' switches depending on whether they switch a single wavelength or a set of wavelengths.

The remainder of this section describes and discusses the four architectures shown in Figures 4.3, 4.4 and 4.5, following this brief summary of each architecture:

**FULL architecture:** shown in Figure 4.2, at each node all incoming fibres are demultiplexed into wavelengths, which are then switched, before multiplexing back into outgoing fibres. All incoming wavelengths may be received into electrical form to leave the optical network here; all outgoing bandwidth may originate from transmitters at this node.

**WAVEBAND architecture:** shown in Figure 4.3, at each node incoming fibres are de-

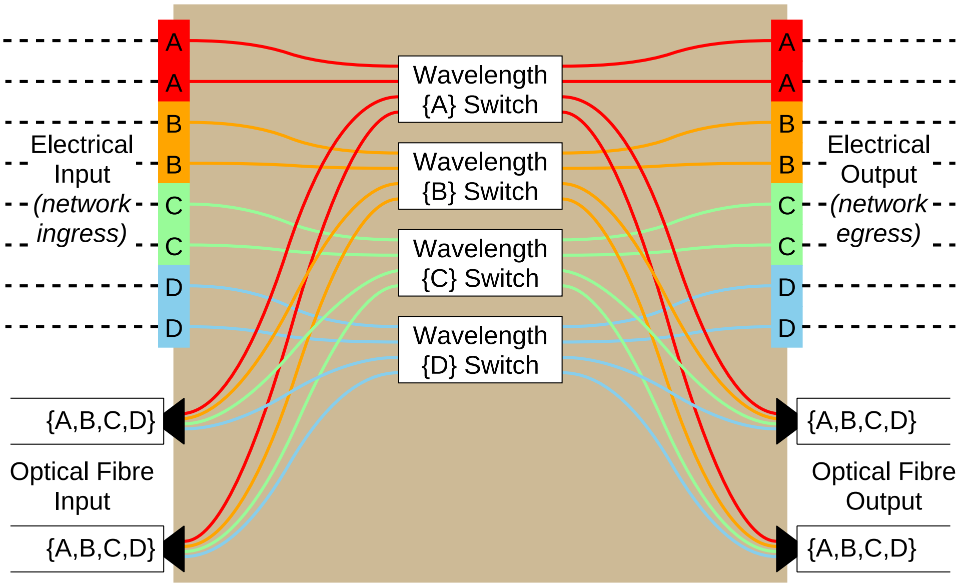

FIGURE 4.2: WAVELENGTH ROUTED TOPOLOGY — THE FULL ARCHITECTURE

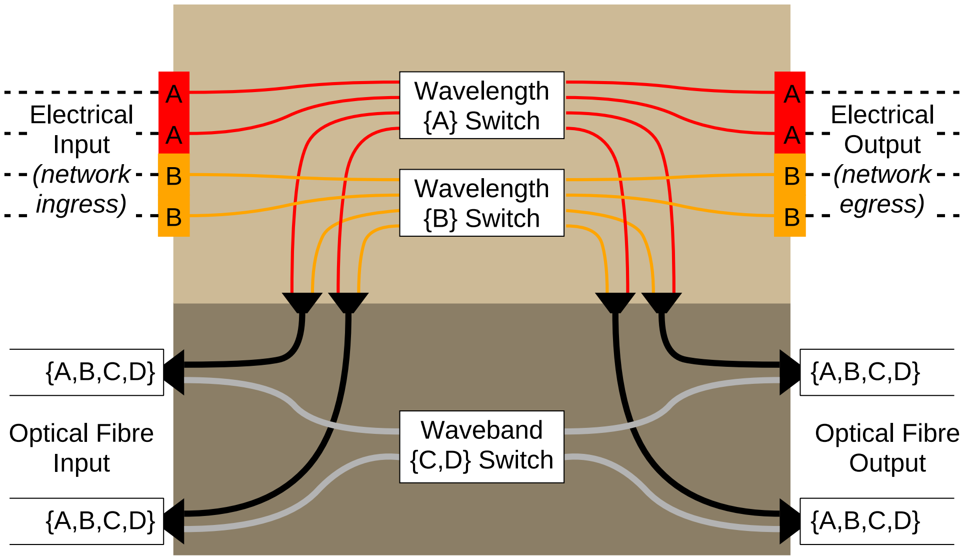

FIGURE 4.3: FIXED WAVEBAND PATH ENDPOINTS — THE WAVEBAND ARCHITECTURE

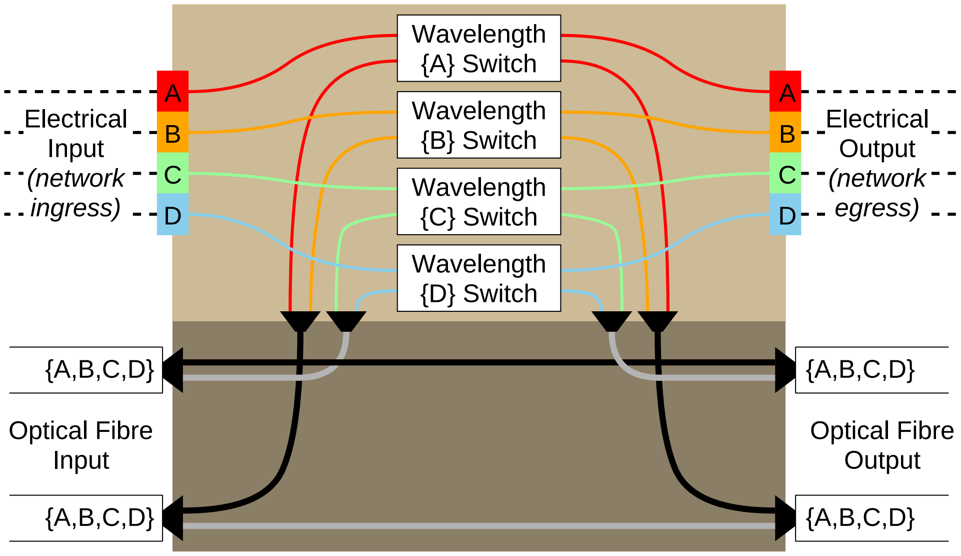

FIGURE 4.4: FIXED WAVEBAND PATHS — THE PATH ARCHITECTURE

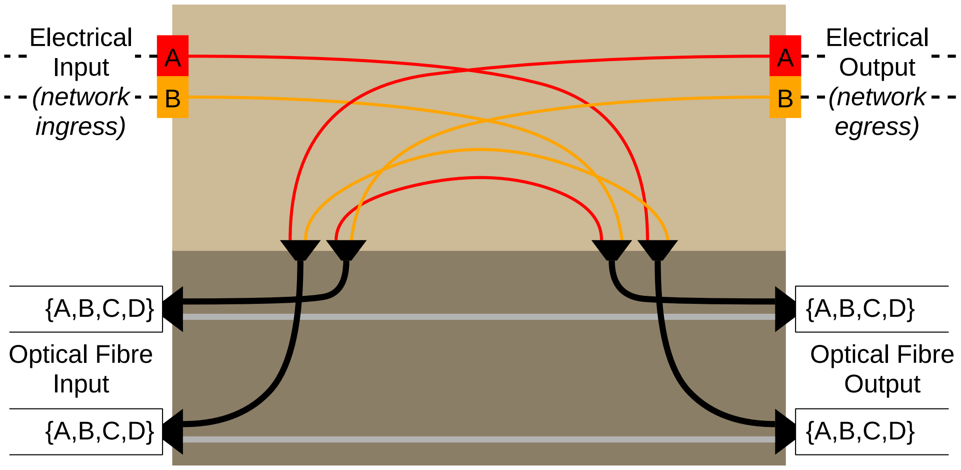

FIGURE 4.5: FIXED WAVELENGTH ROUTES — THE FIXED ARCHITECTURE

multiplexed into wavebands. These may be switched as entire wavebands, before being multiplexed into outgoing fibres, or they may be demultiplexed further into wavelengths. All these wavelengths may either be received into electrical form, or switched before being multiplexed into the start of a new waveband path. Since waveband paths are switched, new waveband paths may be created dynamically.

**PATH architecture:** shown in Figure 4.4, similar to the WAVEBAND architecture, but the set of waveband paths in the network is fixed at provisioning time. Wavelength routes are created by joining-up waveband paths in the network that have spare capacity.

**FIXED architecture:** shown in Figure 4.5, all wavelength routes are fixed in the network at provisioning time; each node will demultiplex incoming bandwidth into wavelengths and have a patch panel to either receive the bandwidth into electrical form or to multiplex into an outgoing fibre.

Figure 4.2 shows an example of a fully flexible wavelength routed node architecture — the FULL architecture. Each wavelength has a separate switch, which has twice the number of input ports as the number of incoming fibres. This is so the full outgoing optical bandwidth can originate here. A similar argument exists for the output ports. We have full reconfigurability, but high cost.

Figure 4.3 shows an example of the architecture for a waveband routed network node — the WAVEBAND architecture. The fibres are subdivided into 2 wavebands, each of which contains two wavelengths. These may either be further demultiplexed into their constituent wavelengths, or be switched at the waveband level.

Part of the provisioning algorithm picks a set of endpoint capacities for each (node, waveband) pair; these say which wavebands at which nodes can be demultiplexed into wavelengths and switched at the wavelength level. Such demultiplexing may be independent for incoming and outgoing fibres, although all the results presented here use bidirectional demands and routeing, so the incoming endpoint capacity and outgoing endpoint capacity for any (node, waveband) pair are equal.

This means that we can dynamically set up a waveband routed path, so long as the nodes where the path originates and terminates have sufficient capacity. In Figure 4.3 waveband {A,B} has a termination capacity of two paths, since both incoming wavebands are terminated; demultiplexed into wavelengths and switched at that level. Waveband {C,D} has no ability to terminate or originate waveband paths, since both incoming paths are switched at the waveband level then multiplexed back onto the outgoing fibres.

Figure 4.6 shows, for a single waveband, two alternative configurations of waveband paths given the same topology and endpoint capacities. The endpoint capacity is one for all three nodes: since we are assuming bidirectional routeing one path may terminate each node, and one path may originate from each node. If we assume there is only

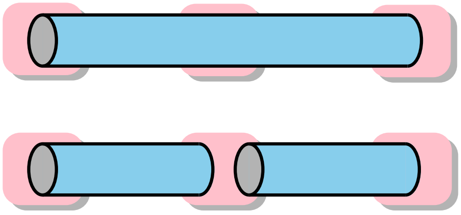

FIGURE 4.6: EXAMPLE OF TWO DIFFERENT WAVEBAND CONFIGURATIONS FOR THE SAME SET OF ENDPOINT CAPACITIES

a single fibre on each link, then in the first diagram we lose the ability to terminate a path in the centre node. This means that while the long path exists, we do not have access to the wavelength switch, receivers, and transmitters at the centre node. The second diagram, with two shorter paths, leads to a more flexible set of waveband paths. Whilst this is a trivial example, the many fibres on each link and different endpoint capacities for different wavebands complicate the routeing process.

The WAVEBAND architecture has the advantage of reduced cost over the FULL architecture: since we do not allow all the bandwidth to originate or terminate at the node shown, we need fewer transmitters and receivers. We also need fewer and smaller wavelength switches since some bandwidth is switched at the waveband level. The disadvantage of this is the loss of flexibility: originally we could add eight wavelengths of traffic onto the network at this point —- now we are limited to four. This may lead to blocking if we have a high demand for traffic originating or terminating here. Since we are dynamically routeing waveband paths, we also have a problem of optimising capacity as shown in Figure 4.6.

We note that in general each waveband switch has to have a number of ports equal to the sum of the number of fibres and the number of wavebands to be demultiplexed into wavelengths. This allows any combination of our incoming wavebands to be demultiplexed whilst the rest are switched directly. This is shown in Figure 4.7, which shows the required paths for a single waveband {A,B} if we have four incoming fibres and wish to demultiplex any two wavebands. Note that for clarity in this diagram the other waveband is omitted, as is the wavelength routeing section of the node.

Figure 4.4 shows an example where at the provisioning stage we fix the topology of all waveband paths in the network — the PATH architecture. This is similar to WAVEBAND, but is less flexible since we cannot alter the route of any given path after provisioning the network. Since routes generally consist of multiple paths we can still place wavelength routes dynamically, which may give sufficient flexibility. We also remove the need for a dynamic path routeing algorithm that takes into consideration the joint optimisation of endpoint and fibre utilisation, since this is done once during provisioning. For the remainder of this work we will consider this at the expense of WAVEBAND.

Figure 4.5 shows the final example of the trade-off between routeing flexibility and cost — the FIXED architecture. Here we place wavelength routes and remove all unneces-

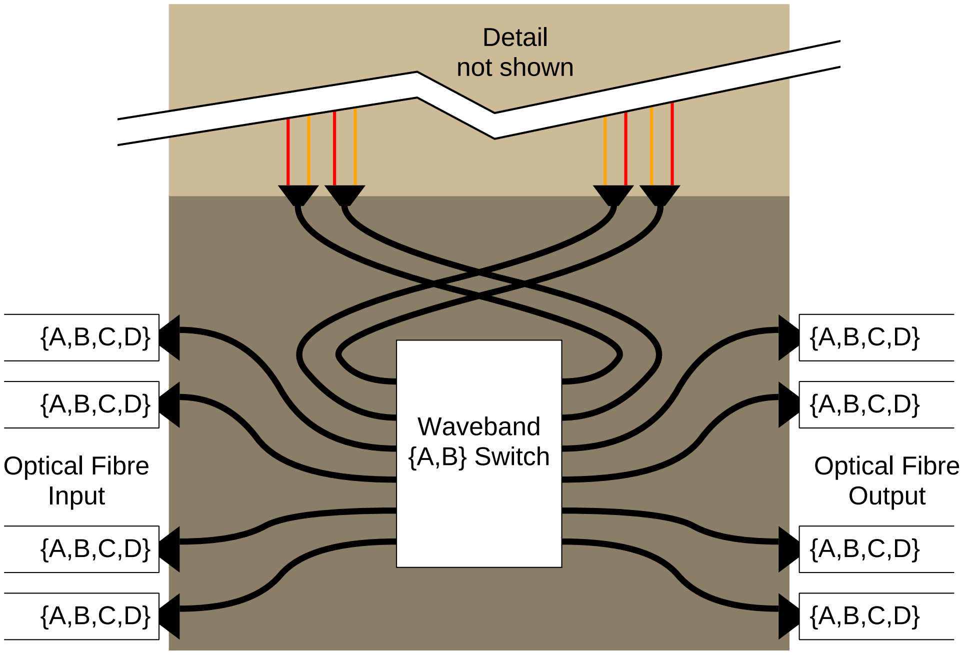

FIGURE 4.7: PARTIAL DETAIL OF A WAVEBAND SWITCH WITH PARTIAL DEMULTIPLEXING OF INCOMING WAVEBANDS. THE SECOND WAVEBAND AND WAVELENGTH ROUTEING DETAILS ARE NOT SHOWN.

sary equipment, our network node becoming a static optical patch panel. We have no flexibility in the wavelength routes after provisioning the network. However, since we minimise the optical components needed to satisfy a given demand, we may be able to over-provision enough to yield a comparative level of performance.

We contend that using waveband switching leads to a net gain over the two extreme alternatives of the most and least flexible networks. Results of testing these architectures are reported in Section 5.2.

### 4.2.4 Routeing Algorithm

In using these network architectures we have two timescales in consideration, as explained in Section 4.1. The timescale of dynamic network reconfiguration leads to the *fixed topology* problem, where our network topology is fixed and traffic demands varies. The longer timescale of network installation can be modelled by the *fixed demand* problem, where the traffic demand is fixed and we build our network to support that traffic load.

When considering the fixed demand problem we are not trying to optimise the final network cost, but to optimise the performance of that network under the fixed topology problem. We decide to design a single routeing algorithm which can solve either problem. This is partly to reduce the overhead of designing two routeing algorithms, but also since allowing the routeing algorithm to solve the fixed topology problem using a topology that was designed by the same algorithm may lead to better performance.

We consider the operation of a dynamically reconfiguring network. Through a range of management and traffic led stimuli we would like to either add a new route, delete an existing route, or combinations of the two.

One design choice is whether to batch changes together or to process each individually. Batching changes together would increase the latency of change occurring which is undesirable especially if the changes are traffic led. It also involves some centralisation, whereas we might like to maintain the ability to choose between a centralised or distributed approach for other performance or scalability reasons. For the remainder of this work we will assume that changes are not batched in groups, however with the existence of some centralised scheme batching could be achieved if desirable for technical reasons.

Another consideration is the scope of each change. For each small change, such as adding a new route, we have the option of re-routeing any existing routes in the network that are not the subject of the changes.

At one extreme we might perform a full optimisation of all current routes, which could mean re-routeing a large number of existing routes. At the other extreme we could leave all existing routes unchanged. Firstly the complexity of routeing the full set of current demands for a network of any reasonable size might to be too great to be able to perform this very frequently. This would inhibit our ability to track traffic shifts over the timescales that were investigated in Section 3.6, or require batching together of pending changes followed by a full optimisation.

The middle ground between these extremes is to re-route only some of the existing routes, leaving the majority intact, to allow new routes to be added to the network. In some cases where routeing fails initially, existing routes may have alternative paths through the network that would then allow our new route to be added. It is not clear if the average blocking would decrease however; if we assume that routes generally follow short paths, then rearranging existing routes might make those routes consume more resources. In the longer term we might find that on average routes become more inefficient. This issue is explored further in Section 5.3.2. As discussed in Section 3.6.2, we would like to minimise the time taken to add new routes; performing an analysis to decide which existing routes could be moved to pack our new route would certainly increase this latency. With a distributed implementation we may have multiple routeing attempts taking place in parallel, with a form of locking on resource allocation to prevent two routeing attempts both thinking they can use the same resource: the same wavelength on a fibre for example. Using a routeing method that affects multiple routes increases the chances of collision in concurrent routeing attempts.

For these reasons we shall concentrate on schemes that perform only a local optimisation; we add our new streams regarding all existing routes as fixed, optimising only the path of our new streams. We regard changes as individual events to minimise the latency between the change being requested and acted upon.

This local optimisation is carried out in the knowledge that other routes will be added in the future. This is important when we consider the effect of modularity of allocation, since optical networks are inherently modular. For example, when provisioning a network we have to add entire fibres at a time. During provisioning when adding a new route we might have two choices. The first uses the shortest path route, but would require another fibre to be added since there is no spare capacity on one of the links; the second takes a longer path but uses existing fibre capacity. The second requires less addition to the actual network cost, since we are using spare capacity on our existing modular resources which increases utilisation but is relatively less efficient than the former solution — our route itself is using more resources than if it followed the shortest path. So whilst in general the first route would be better, leading to a more efficient network — in the sense that all routes follow their shortest path — it does not promote good utilisation of modular resources, especially if there are relatively few future routes to add. The following section presents an approach that balances these two issues.

### 4.2.5 Routeing Metric

A method that can trade off the advantages of short efficient routes and high network utilisation is now described. For each edge we adjust the weights which are used to calculate the route using Dijkstra's shortest path algorithm [Dijkstra59]. Weights are determined by how the new route will be constructed and the distance travelled through the network. We can either operate the flooding algorithm on the entire network, or restrict the possible routes chosen to a predetermined set of alternatives: the k-shortest paths, for example.

If routes are limited to a predetermined set, the choice of allowed routes is not trivial, since we wish to be able to create disjoint pairs of routes for protection purposes. Restricting the route choice might lead to a route being blocked when it is possible to construct a non-allowed route. If our set excludes long routes then this may be of benefit, since we do not allow routes that use excessive resources. It also has the consequence of speeding up the search algorithm since we operate it over the restricted topology containing only the allowed routes. However with the open question of route-set selection, results reported in this dissertation use a complete flood of the network.

At each node we have a set of possible edges to choose from to continue our new route. These edges will have a number of fibres on them, each of which will have a number of wavebands in use. Each waveband in use forms part of a waveband path; this path is fixed with a given start and end point, and may have some wavelengths in use and some free within the waveband. We can therefore create a set of possible virtual hops. These will include creating a new waveband path inside an existing fibre, creating a new waveband path inside a new fibre added to an edge, or following an existing

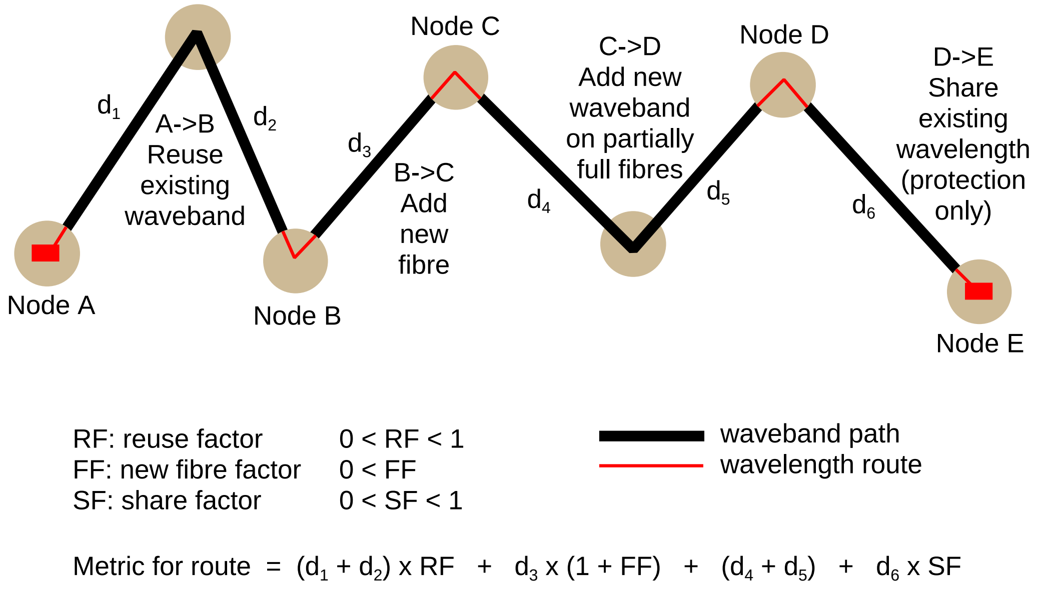

FIGURE 4.8: EXAMPLE USE OF THE METRIC FOR PLACING A NEW ROUTE

waveband path using a spare wavelength inside the waveband — this virtual hop will take us to the destination of the path. Finally if we are creating a shared protection path we can share an existing wavelength on an existing waveband path. This is allowed if the existing wavelength on this path is part of a shared protection route, and that the two corresponding primary paths are disjoint.

Some of these options will depend on the scenario: we can only add a new fibre if we are provisioning a network, rather than doing dynamic demand rearranging. The flexibility of the node architecture is also important. Using the model of routeing flexibility presented in Figure 4.4 we are not able to create new waveband paths; when using Figure 4.3 we have to check the available path endpoint capacity before creating or terminating a new path.

The routeing metric is shown in Figure 4.8, as a function of the physical distances between nodes ($d_1$, $d_2$, etc.). This distance function could be altered to reflect other aims, such as letting $d_i = 1$ to minimise hops. The route, shown as the thin red line, is built from multiple paths, shown as thick black lines. A path is one of four possible types, as explained above. The first, (A → B), is where a path with a spare wavelength exists along these edges. The second, (B → C), requires a new fibre to be added, upon which one waveband path will be fixed and one wavelength from that waveband used. This type of path is available only during the provisioning stage. The third, (C → D), uses existing fibres with spare capacity and adds a new waveband path along this route, switched at the waveband level through the intermediate node. The final path, (D → E), is only allowed for protection routes and shares an existing wavelength along an existing path. The factors $RF$, $SF$ and $FF$ can all be adjusted to affect the end result, and should be set to minimise the real cost of the end solution; they balance the goals of utilisation and efficiency mentioned above.

Adding different weights based on the type of path used effectively adds many virtual edges to our physical topology. Each waveband path is a new virtual edge from its source to destination, every wavelength used as a shared protection path through a given waveband is a virtual edge, each fibre with spare waveband capacity is a virtual edge, together with a virtual edge for each actual edge corresponding to a new fibre. The lengths of these virtual edges are linearly related to the length of their physical path by the factors given above. We run a search algorithm over this set of virtual edges.

Other possible extensions to this routeing metric could add further bias. For example, if a limited wavelength conversion facility were available we could add a penalty to the metric for routes using this sparse resource. This would prevent unnecessary use of conversion facilities, which would be used only when no other alternatives exist, or where alternative routes are significantly longer than the route using conversion.

### 4.2.6 Discussion

A problem with proposing this scheme is that we do not know what the parameters should be to yield the best performance. Different scenarios might even require different parameter values, or there might exist multiple minima so the correct parameter values do not converge to a single set of values. One option for a given experimental scenario would be to evaluate several different sets of parameters. Although our parameter space is multi-dimensional, we are limited by the constraints listed in Figure 4.8. In Chapter 5 we explore this parameter space and find that there are values for these parameters which typically yield good results.

However, for some experiments it may not be of critical importance whether we select the optimal combination of parameters. To assess the impact of a change in design of the network, we need to ensure that our choice of parameters gives results equally close to optimal both before and after the change. Alternatively if we can estimate rough bounds on how close to an optimal parameter set we have, and as long as the change in performance during our experiment exceeds our optimality bound then we can say that the measured difference is significant.

When investigating the fixed topology problem we may have two parameter sets to choose from: the first used to construct our topology, the second used to route the dynamically occurring demand. For the provisioning stage we are not seeking to find the optimum network cost for a fixed set of demands: we are seeking a flexible network costing a fixed amount. We need the network cost to be fixed to compare different network architectures, and can freely vary the demand set used to provision our network, so this stage need not be optimal. The second stage, during dynamic demand, is an easier optimisation problem since we remove $FF$, the new fibre factor, reducing the dimensionality of our search space.

The motivation behind this metric is that it can be used in a distributed fashion, and is able to add a single route rather than having to perform some global routeing algorithm over all routes. This is desirable both since we wish to reduce the latency between requesting a new route and that route being operational, and that existing routes in the network should not be affected as we are adding new routes. Using the different parameters allows the algorithm to cope with choosing between the large virtual topology created by partially used fibres and waveband paths.

In the next chapter we will look first at what the best values for our routeing parameters are, and the effectiveness of using this algorithm for the fixed demand problem. We can then look at network design parameters such as the size of a waveband and the best node architecture to allow dynamic network reconfiguration.

## 4.3 Experimental Setup

Since we are interested only in the broad timescale of network reconfiguration under a circuit switched environment, it is not necessary to deal with individual packets or packet flows. Instead we shall restrict ourselves to dealing with reconfiguration events: adding a new path through the network or removing an existing redundant path. Simulating packet flows across the network at the intended level of aggregation would be very time consuming. In Chapter 3 we examined how packet-based information such as delay or loss corresponds to large scale bandwidth allocation changes, so that level of detail is unnecessary here.

### 4.3.1 Topology Generation

A large body of work exists on accurately generating realistic Internet-like topologies. However, what we are engaged in is simulating the core of the future Internet, where the multi-hierarchy access networks provide the aggregate traffic streams that we route across our optical network. Our approach is motivated by the method used by Tiers [Doar96] to generate a single-level network, amended to create a biconnected network.

As our model topology we use a pan-US network provided by a major network equipment designer as one of their major client's networks, shown in Figure 4.9. Access points are at 21 major cities; the 32 links give an average node degree of just over 3.

To generate network topologies of this type we place $n$ nodes randomly in a square area with sides of length $s$, with a minimum separation between nodes of $\alpha s/\sqrt{n}$. This is designed to simulate the natural placement of population centres, and therefore desirable access points, for our network. The *proximity factor*, $\alpha$, determines how clustered our nodes can be. The current algorithm places each node, and then tests whether or not our minimum separation constraint has been violated. If it has, the node is placed again at random. Although there is a theoretical upper bound on $\alpha$ from the circle

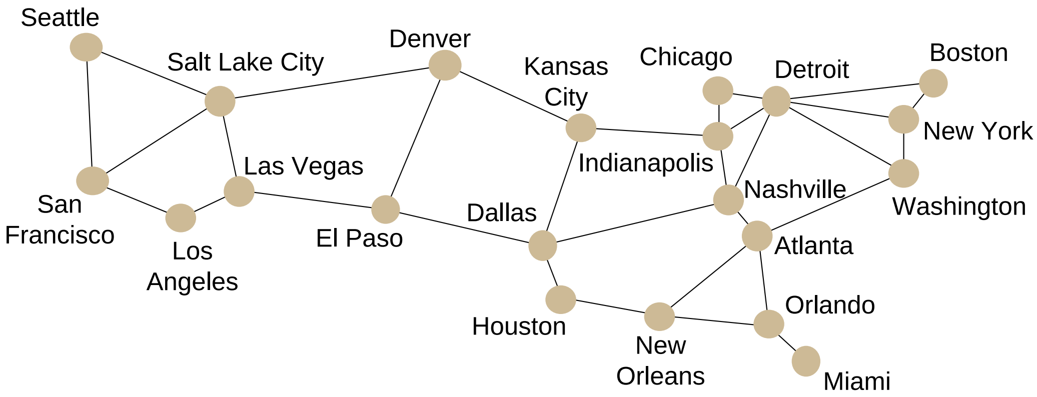

FIGURE 4.9: A US BACKBONE NETWORK

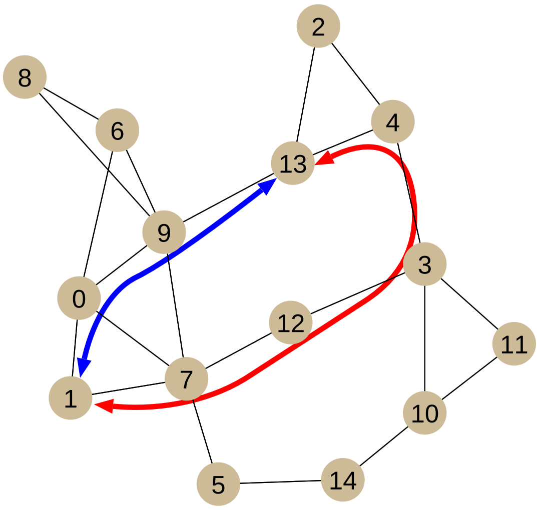

FIGURE 4.10: EXAMPLE TOPOLOGY WITH SHORTEST DISJOINT PATHS BETWEEN NODES 1 AND 13 MARKED

packing problem, it is hard to compute and depends on $n$ [Boll00]. Practical testing has shown that the naive algorithm works for values of $\alpha < 0.9$, high enough to produce a near regular mesh network.

We allow fibre routes between nodes to be added only between given pairs. These channels are mono-directional, but added in pairs to make bidirectional edges. Each channel can carry multiple fibres; each fibre has the same set of wavelengths.

Having a node disjoint primary and protection path between each pair of node requires a biconnected network. Finding the minimum biconnected graph is an NP-hard problem [Czumaj99], and in general solutions result in a ring network. Therefore we present a simple scheme for generating biconnected mesh-like topologies.

The pairs of nodes to be joined with an edge are chosen first to include a minimum spanning tree between the nodes using Kruskal's algorithm [Kruskal56]. Secondly we add edges to make our network biconnected. We consider each node in turn, and temporarily remove that node and all edges connected to it. In a biconnected network, the remaining network will be connected so we add the minimum length edges needed to reconnect our network. We add back the removed node and respective edges, and move on to the next node to be removed.

We now have the option to add further redundancy to our network by specifying a minimum number of edges per node, $\rho$, and adding edges to meet this minimum. These edges are added in length order, starting with the closest pair of nodes that are not already connected. This is primarily used to be able to combine results from multiple networks, where each network has a similar topology both in connectedness and in the number of nodes and edges.

An example topology produced by this process, having 15 nodes and 22 bidirectional edges, is shown in Figure 4.10. It has $n = 15, s = 100, \alpha = 0.7, \rho = 3$. This is one of the topologies used for simulations presented in Chapter 5.

It is likely that the major difference between this and future optical networks is that this algorithm assumes a uniform distribution of access points, subject to our proximity constraints. It is likely that local effects such as geographical features will affect this, and also that the topology of existing networks depends on their growth over time. A network that starts small and grows is likely to have a higher interconnectedness around the oldest part of the network. This is partly modelled by the proximity constraint; a lower constraint allows nodes closer together, which will be more richly interconnected due to redundant edges being added in length order.

While networks of this type are easy to visualise and reason about, current network designs tend to be more complex. Temporary failures due to line card and router reboots are a common source of day-to-day failures, as discussed in Section 2.3.6. To protect against this, inside each network node is a small collection of IP routers connected with a full-mesh topology, with multiple links between the same pair of nodes connected via different routers [Iannaccone03]. Thus for a single failure inside the node, traffic can be restored by re-routeing internally in that node, and still use the same inter-node path. Quite often multiple inter-node links will not be sufficiently disjoint to be independent from fibre cuts. To simulate network of this type would require the concept of a Shared Risk Protection Group (SRPG), where we are explicitly able to say which resources are independent from failure. This contrasts with our current notion that a single node constitutes a SRPG. Our simple topology would have to be enlarged, with single nodes split to reflect more accurately current topologies, and multiple links between a single node pair. Apart from the longer running time caused by a large topology it is not expected that these changes would be impractical or substantially change the results presented in this work.

### 4.3.2 Traffic Generation

To assign a fixed traffic demand we set a maximum demand $\beta$, and for each node $i$ we associate a population index $P_i$, uniformly distributed between $0$ and $1$. The number of wavelengths required from node $i$ to $j$ is $\beta P_i P_j$. Each of these routes may also require some form of protection path, depending on the experimental scenario. For the example topology shown in Figure 4.10 setting $\beta = 20$ would give an average of 5 wavelengths of demand each way between a pair of nodes, which with 32 wavelengths per fibre gives a provisioned network of around 7 fibres per mono-directional edge. This method of creating our traffic matrix was chosen after discussion with a major network equipment designer, and is one of their methods for investigating long-term future network design.

To create dynamic demand we use the observation from Chapter 3 that traffic aggregates are likely to exhibit some form of long range dependence and non-stationarity. To render a practical set of experiments we limit the traffic distribution to a stationary long range dependent one. This will model the average behaviour of our traffic that would be observed over periods of time between events that cause non-stationarity. We can model the effects of non-stationarity by changing the statistical distribution of the dynamic traffic mix and restarting the experiment. This would obtain the average behaviour after a non-stationary event, during the next stationary period. Given the complex nature of non-stationarity this is certainly a simplification; however it is a tractable approach.

To create our long range dependent trace we chose to use Fractional Gaussian Noise (fGn) to generate a series of numbers that indicate the quantity of traffic between two nodes. These figures are interpreted as the number of wavelengths required between these two nodes. Fast methods of creating long fGn series exist and, given the increasing self-similar nature of observed traffic over the timescales of interest seen in Figure 3.10, using a self-similar traffic model is appropriate.

Each fGn stream between nodes $i$ and $j$ is scaled to have a mean of $\beta P_i P_j$, and a standard deviation equal to $\tau$ multiplied by the mean. The larger this value $\tau$, the larger the movements away from the mean value, so the harder the test for the network since actual demand can be different from the expected demand by a greater amount. For the full trace investigated in Chapter 3, $\tau$ varies from 0.30 to 0.278 over the estimated timescale of change indicated by Figure 3.17, and for the 13 split traces from Section 3.4, $\tau$ varies from 0.2 to 0.65.

With any reasonable value of $\tau$ some demand values will be under zero, not a valid number of wavelengths. The remedial approach used here was to truncate negative values to zero, but carry forward the values lost to subtract against the next positive values. This was chosen over the other option of strict truncation, where negative values are changed to zero, in order to maintain the mean value of the series.

The value of $\tau$ chosen was 0.5, which is a compromise between wanting the highest value measured in actual traffic and the need to limit the frequency of values below zero that have to be truncated. This will give the most variable traffic conditions, but limit the effect of the change in the lower tail of the distribution.

To translate the number series to a set of events to create or remove extra wavelengths, we assume that for each pair of nodes there is a continuous time series indicating the exact quantity of traffic required. We assume that our discrete fGn time series samples this continuous process at fixed time intervals. Thus our simulator works by moving along these time intervals, and for each pair of nodes calculates how many routes to add or remove by comparing the current sample value to the next value. Considering the case where we have to add or remove multiple routes from one time period to the next, we assume there is a monotonic increase or decrease in average traffic between the two sample points so we randomly place the events required between the current sample time and the next sample time. This approach also removes the problem of deciding which events to process next, which if implemented naively could result in biased results.

A defined translation between our sample time period and the actual time that we are simulating is not required. This is the relationship between the timescale of data transfer and that of network reconfiguration, which depends on many factors as shown in Section 3.6. Whatever the point chosen on the trade-off between frequency of change and bandwidth utilisation, our simulator only needs to monitor the performance of the network between changes.

When running a simulation with high load, some attempts to add a route to the network may fail. The proportion of these failures against all routeing attempts is the *blocking probability*. If we have a routeing failure between a pair of nodes, when the next call comes to delete a route between this pair we treat the routeing failure as the route to be deleted, rather than deleting one of the routes between these nodes that was routed successfully. This ensures that we treat the number of wavelengths requested by our fGn traffic stream as the traffic demand to be met, and the number of wavelengths successfully routed between our pair of nodes is the greatest possible number up to this traffic load.

### 4.3.3 Network Cost Calculation

To be able to compare two network solutions in the fixed demand problem, we could choose a number of evaluation functions. Past work has concentrated mainly on the final cost of the networking infrastructure required [Baroni00, Binetti00]; alternative approaches include setting acceptable limits of signal to noise ratio along optical paths [Sabella98]. In this work we use a measure of the final equipment cost, breaking the optical network into four major categories. This is deliberately a simplification of a real optical network, designed to be applicable to a wide range of proprietary network solutions and network stacks, as previously discussed in Section 2.3.7.

The equipment breakdown and estimated costs are listed in Table 4.2. Breaking down

| Equipment | Cost |
|---|---|
| fibre and amplifiers | 10 per unidirectional unit length |
| transmitters and receivers | 8 each per wavelength |
| optical $n$ port switch | $4(n \cdot log_2(n) - n/2)/5$ |
| multiplexer $1 \rightarrow n$ | $n/4$ |

TABLE 4.2: EQUIPMENT COST MODEL

the costs into four areas was designed to measure the differences in the node architectures listed in Section 4.2.3, and to be at the level of detail of the node descriptions given. Other network costs are likely to be the same across all architectures — for example the cost of the IP level equipment will be the same for a given scenario — or they could be included in one of the four categories listed.

The scale of the fibre and amplifier costs is per unit length, where the mean length for a fibre cable is intended to be around 30 units, giving an average fibre cost of 300. The transmitter and receiver costs are a fixed cost per wavelength. The multiplexing cost is given per multiplexer and is linear to the fan-out or fan-in for that multiplexer. The optical switch cost is based on the Beneš rearrangeable non-blocking switch architecture, similarly to previous work which compares network architectures by cost [Binetti00]. In large optical networks it is more likely that MEMS switches will be used, possibly partitioning a large switch into the required sizes. However the cost model for this is unclear, and because of the modularity of allocation it depends critically on the scale of the network being simulated. For this reason we use the cost model as shown.

The weights between each cost were chosen through consultation with a major manufacturer of telecommunications equipment. They are in arbitrary units and only have relevance in their relative values; they were designed to be representative of major long haul networks. A typical scenario used an $s = 100$ square to place nodes, with fibre hops of between 18 and 42 units long and an average of 27.

Whilst finding the optimum solution for a given topology and traffic load may be critical, it is informative to be able to identify general trends that will affect a range of network scenarios. For example, whilst a specific scenario might have an optimum waveband size of 8 wavelengths, most might have an optimum of 4. To identify trends a number of independent scenarios are used and average results are calculated. These scenarios will be similar; for example, in the number of nodes and edges, and the distribution of edge lengths, but the exact topology and traffic will be generated using different random seeds.

To be able to average results we need to give even weight to each scenario used; however different scenarios might have very different final costs due to differences in topology or traffic matrix. To produce even weights, costs are divided by a *base line cost* for each scenario to form a *cost ratio*. The base line cost for a scenario is a measure of the final network cost given uniform routeing conditions: shortest path routeing for all traffic. To remove the influence of fibre modularity — the fact that wavelengths come in sets of a fibre — the cost of each fibre is multiplied by the utilisation of that fibre. Although this will lead to a lower bound for fibre cost, it is likely either not to be achievable or to be a particularly close lower bound to the optimal network cost. Therefore the cost ratio achieved for each scenario is meaningful only for relative comparisons. Since each scenario can now be given equal weight, these cost ratios can be averaged and used for a fair comparison.

## 4.4 Conclusion

In this chapter we have outlined possible options for the architecture of an optical transport network. We have introduced the idea of a *waveband*, as a way of exploiting the two-stage multiplexing process. This allows us to demultiplex a fibre into wavebands, which can then be optically switched as a single unit, without further demultiplexing into constituent wavelengths.

We have described the high level design of several different all-optical network nodes, each of which trades off the potential benefits of wavebands, reducing the cost of each network node by using waveband switching, against the potential costs of this approach — the loss of routeing flexibility.

A routeing algorithm has been described that is suitable for local optimisation of routes, using a distributed algorithm to find suitable routes. This can also be used for network provisioning. Here we constructed a new network to handle a given set of routes, aiming to create not the minimum cost solution, but a network with enough routeing flexibility to handle changing demand with low blocking rates.

In order to test these ideas we chose to simulate the high-level functionality of a network. We gave a method to create the topology of a biconnected mesh-like network with redundancy, and a method of creating a traffic process that has similar statistical properties to that seen in Chapter 3. We finally provided a way of reducing an optical network design into four simple cost categories, to allow fair comparison of different architectures.

In the next chapter we use the simulator described above to experiment with the ideas contained in this chapter. We firstly explore the routeing algorithm given, seeing whether suitable values for our routeing parameters can be found. Next we look at a critical design choice when using wavebands: how many wavebands should be in each fibre? We then compare the different node architectures outlined in this chapter under a range of traffic conditions and relative network costs, and look at some possible network optimisations.

# Chapter 5

# Simulation Results

In designing the architecture of an optical network we outlined a series of possible options in Chapter 4. We have given a method of placing new routes in a network in Section 4.2.5.

- The routeing metric has a number of parameters — are there suitable values for these parameters?
- We have presented the notion of using wavebands — is there an optimum waveband size for a given scenario?
- What are the effects of trying to incorporate shared protection routes if we use wavebands?
- When we have a changing traffic load, the routes will be added to the network in no particular order — is this a disadvantage over being able to add routes in a given order?

These questions are answered in Section 5.1 by investigating the fixed demand problem, the problem of finding a low cost network to satisfy a fixed set of demands.

In Section 4.2.3 we outlined possible network node designs, which trade off the routeing flexibility of a network and the cost of each node for the same fibre throughput.

- When networks are exposed to dynamic demand, what is the trade-off between routeing flexibility and total network cost?
- Does this trade-off change as we change the dynamic traffic matrix to be supported by the network?
- How is the trade-off altered by changing relative equipment costs?

- Can we improve performance by looking at using tuneable components or other optimisation schemes including route rearranging?

Section 5.2 answers these questions. In Section 5.2.1 we give the methodology used for the experiments which use a fixed topology and a dynamic demand process.

## 5.1 Fixed Demand Problems

In this section we consider the fixed demand provisioning problem. Note that we are not trying to route a set of fixed demands with minimal network installation cost. This can be solved by using integer-linear programming techniques and a typically time-consuming offline algorithm. However, the provisioning problem is closely related to the fixed topology problem which is considered in Section 5.2. Techniques developed in this section may be applied to the fixed topology problem, so long as they are applicable. The biggest difference in the problems is that for the fixed topology problem demand changes by incremental steps — we do not have any idea of the next change in demand while we are adapting the network to the current shift. It is also likely that the fixed topology problem could benefit from a distributed algorithm, since it might scale better for larger networks.

From a practical perspective we need a way of provisioning a network to be used for the fixed topology problem, where we can allocate capacity where predicted demand will be. This is another reason why it is beneficial to understand how our routeing algorithm works when solving a fixed demand problem.

Unless otherwise stated, all scenarios presented below are similar to that shown in Figure 4.10: produced in a grid of size 100, with 15 nodes and an average of 3 bidirectional edges per node placed to create a biconnected network. Fibres contain 32 wavelengths, which are divided into 8 wavebands of 4 wavelengths each. Traffic is scaled with $\beta = 20$, creating an average of 5 wavelengths worth of traffic between each pair of nodes.

### 5.1.1 Example Parameter Optimisation Procedure

There are several parameters in the routeing metric proposed in Section 4.2.5. For a simple provisioning problem using wavebands and no protection, we have the Reuse Factor, $RF$, and the Fibre Factor, $FF$. If we place a new waveband path on an existing fibre, then routeing cost equals distance travelled. If we have to add a new fibre, this cost is multiplied by $1 + FF$; if we are reusing a partially full existing waveband path we multiply by $RF$.

An example of the optimisation procedure to find the setting for $RF$ and $FF$ can be seen in Figure 5.1. Here no protection was used, all fibres contain 32 wavelengths

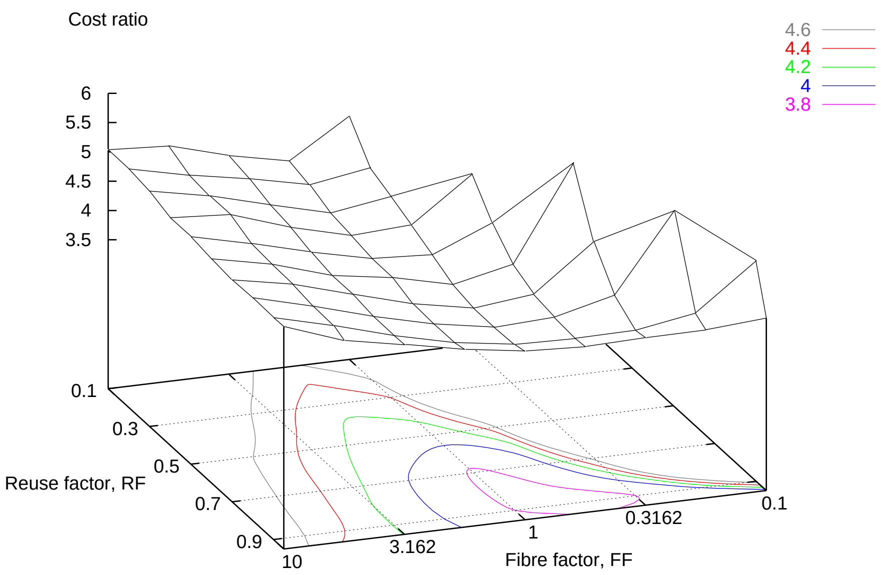

FIGURE 5.1: FINDING THE OPTIMUM VALUE FOR $RF$ AND $FF$ IN ONE SCENARIO

that were divided into 8 wavebands containing 4 wavelengths each, and the network contained 15 nodes and 22 mono-directional links. 20 different random topologies and traffic matrices were used; Figure 4.10 is one of the topologies. Routes were placed one at a time, in a random order. The final network costs are normalised using the procedure from Section 4.3.3 and shown on the vertical axis, presented as a unit-less ratio. Contour lines are shown on the horizontal plane, marking points of equal cost. The Fibre Factor axis is shown as a log scale to improve the clarity of the graph, and the test points are chosen to be evenly separated in a log scale.

Firstly a steep rise is seen when $FF + RF < 1$; to gain resolution on the z-axis for the critical region the very high cost ratios are not plotted. Imagine we have two edges forming part of a route, the first has all of waveband 1 filled, the second has spare capacity in waveband 1. We would like our algorithm to create a new waveband path on both fibres, using waveband 2, rather than add a new fibre on the first edge to take advantage of the spare capacity in the existing waveband on the second edge. This corresponds to $(1 + FF)d_1 + RFd_2 > d_1 + d_2$, if $d_1$ and $d_2$ are the lengths of the edges. If we assume that $d_1$ and $d_2$ are similar in magnitude, we can cancel lengths to get $1 + FF + RF > 2$, or $FF + RF > 1$.

The optimum values for $FF$ and $RF$ can be seen from the contour lines, drawn on the bottom plane of the graph. For this set of experiments the best values found are $FF = 1, RF = 0.8$. The value of $FF = 1$ seems to indicate that this weighted flood fill search is better than either shortest path routeing, which would correspond to $FF = 0$, or routeing based entirely on increasing existing fibre utilisation, $FF = \infty$. The value of $RF = 0.8$ also indicates that a middle ground is preferred; a detour used to increase waveband utilisation would be preferred if it increases distance travelled by less than

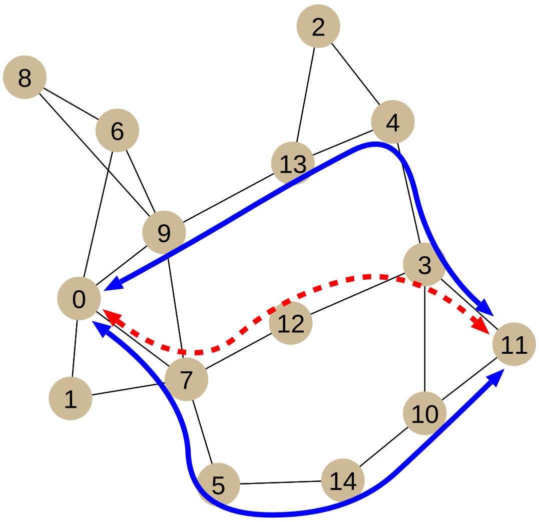

FIGURE 5.2: EXAMPLE OF SHORTEST PATH AND SHORTEST DISJOINT PAIR ROUTES

25%.

### 5.1.2 Routeing with Protection

To route a stream that requires protection, we apply the metric from Section 4.2.5 to potential routes and place the pair of primary and protection routes that are disjoint and have minimum combined score. If we are to allow unconstrained routeing it is essential to find both the primary and protection routes simultaneously, since we may have a situation where if the optimum primary route is placed we can find no disjoint protection route, but there does exist a valid pair of disjoint routes. This is illustrated in Figure 5.2, showing routes between node 0 and node 11. The shortest path follows the dashed red line, $0 \rightarrow 7 \rightarrow 12 \rightarrow 3 \rightarrow 11$, but it has no node-disjoint alternative. The shortest pair of node-disjoint paths is indicated in blue.

Routeing both primary and protection routes simultaneously leads to a complication due to the constraint about protection routes sharing bandwidth. All corresponding primary paths involved in the resource sharing have to be node-disjoint, so that only one protection route requires the shared resource in the event of a single failure. Since we choose both routes together, this check has to be delayed since we do not have a given primary path to check against.

We now have three parameters to optimise; Reuse Factor, $RF$, Fibre Factor, $FF$, and Share Factor, $SF$. When routeing a protection path, if we share a wavelength for the duration of a waveband path then the routeing cost of using that path is the path length multiplied by $SF$.

An example of finding the optimum Share Factor is shown in Figures 5.3, 5.4, 5.5, and 5.6. The same set of network scenarios from Section 5.1.1 is used, but with all traffic

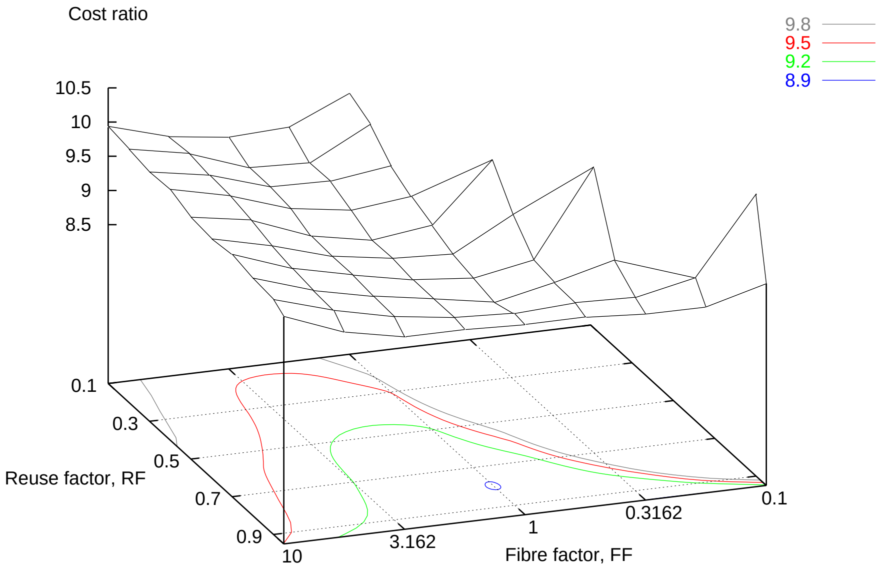

FIGURE 5.3: FINDING THE OPTIMUM VALUE FOR $RF$ AND $FF$, WITH $SF = 0.9$

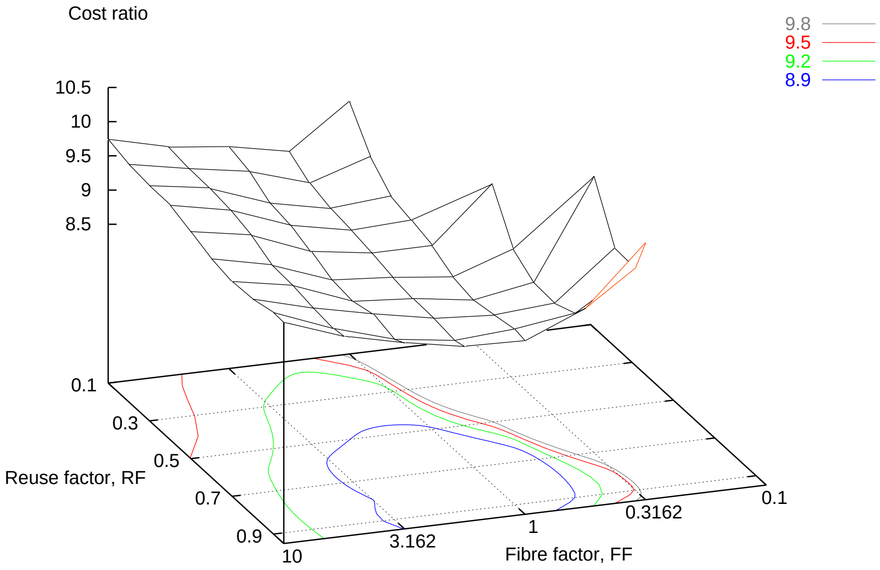

FIGURE 5.4: FINDING THE OPTIMUM VALUE FOR $RF$ AND $FF$, WITH $SF = 0.5$

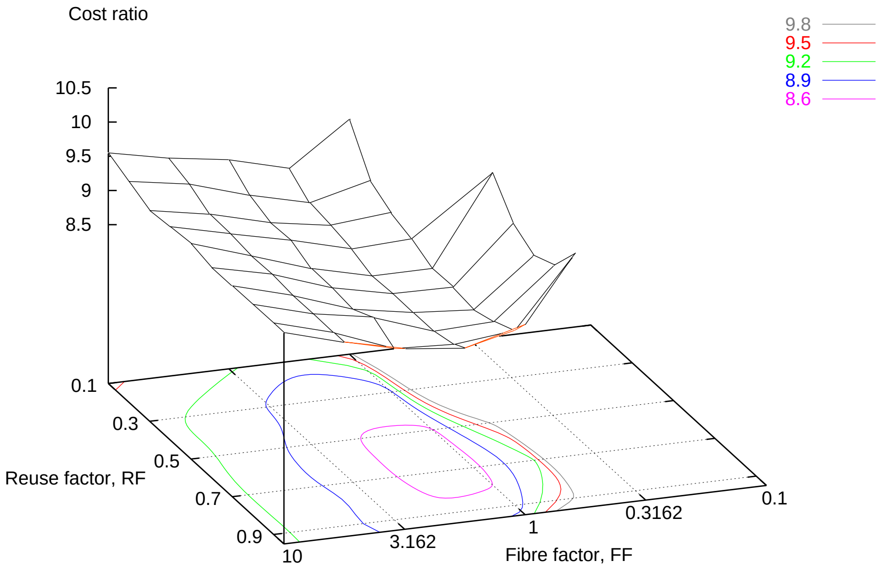

FIGURE 5.5: FINDING THE OPTIMUM VALUE FOR $RF$ AND $FF$, WITH $SF = 0.1$

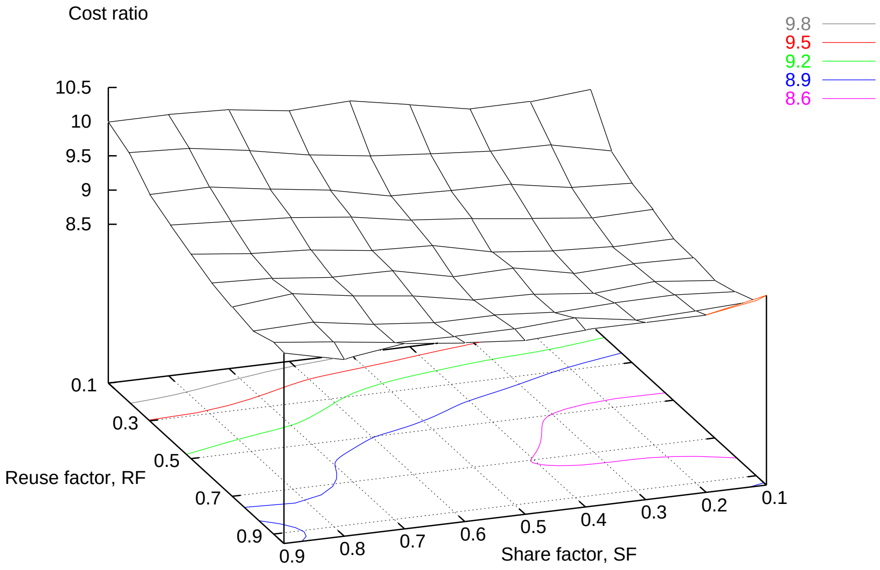

FIGURE 5.6: FINDING THE OPTIMUM VALUE FOR $RF$ AND $SF$, WITH $FF = 1$

demands requiring a shared protection path. With the addition of $SF$ our data set is now four dimensional: $RF$, $FF$, and $SF$ combining to give the cost ratio. To view this data we present a series of three-dimensional graphs where one variable is kept constant. Figure 5.3 fixes $SF$ at 0.9, and varies $RF$ and $FF$ in the same way as for Figure 5.1. With little incentive to share bandwidth, these two graphs are very similar in shape. Figure 5.4 has a $SF$ of 0.5, and Figure 5.5 has a $SF$ of 0.1. Progressively the minimum point in the graph deepens as $SF$ decreases, shown by the enlargement and appearance of another contour line. The minimum point found is with $SF = 0.1, RF = 0.7, FF = 1$. Since on all three graphs the minimum point found is with $FF = 1$, Figure 5.6 varies $SF$ and $RF$, with $FF$ fixed at 1. Here we clearly see from the contour lines that the optimum point is where $SF$ is small and $RF$ is around 0.7.

It is not clear whether $FF = 1$ will always be the best value, or whether it just holds for the set of topologies investigated. This could be examined further by performing more simulations with different network parameters governing the topology and traffic generation. The optimum values of $FF$ will depend on the relative costs given in Table 4.2, since the relative cost of fibre and the other components will affect the desired trade-off between routeing efficiency and utilisation.

### 5.1.3 Comparing Shared and Dedicated Protection

We are inherently limited in the possible degree of sharing, since if two routes share bandwidth for their protection paths, their primary paths must be disjoint. Given $n$ nodes, and an average route length of $m$ hops, then we could share each protection path by a maximum of $n/m$ protection routes. The following two experiments were designed to measure the effect of this relationship on the degree of sharing experienced, and on the final network cost.

In the first experiment the number of nodes was varied from 10 to 25, keeping the same number of edges per node. The traffic levels per node were adjusted to yield approximately the same average number of fibres per edge. For each scenario the lowest network cost using shared protection was compared with the lowest cost using dedicated protection, as a measure of the benefit gained by allowing shared protection. This is presented as a ratio in Figure 5.7. It clearly shows that as the number of nodes increases the benefits of allowing sharing increase, decreasing the network cost. This can also be seen in Figure 5.8, which shows the extent of sharing. Our unit of sharing bandwidth is a wavelength for the length of a waveband path. These lightpaths, if used for a protection route, are used by one or more protection routes. The proportions of how many are used by a given number of protection routes are shown in Figure 5.8. We can see that as we increase the number of nodes we typically have more sharing taking place, as a higher proportion of all bandwidth used for protection routeing is being shared by more protection routes.

The second experiment keeps the number of nodes constant at 15, and alters the connectedness of the topology. Similarly to the previous experiment, Figure 5.9 shows the

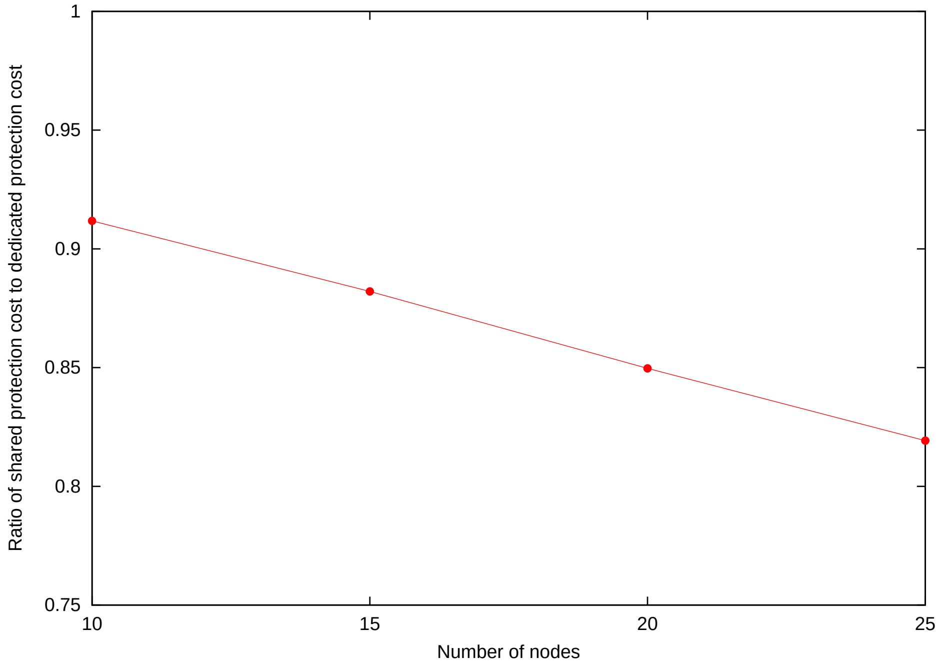

FIGURE 5.7: COST BENEFIT FROM SHARED PROTECTION OVER DEDICATED PROTECTION, VARYING WITH THE NUMBER OF NODES IN THE NETWORK

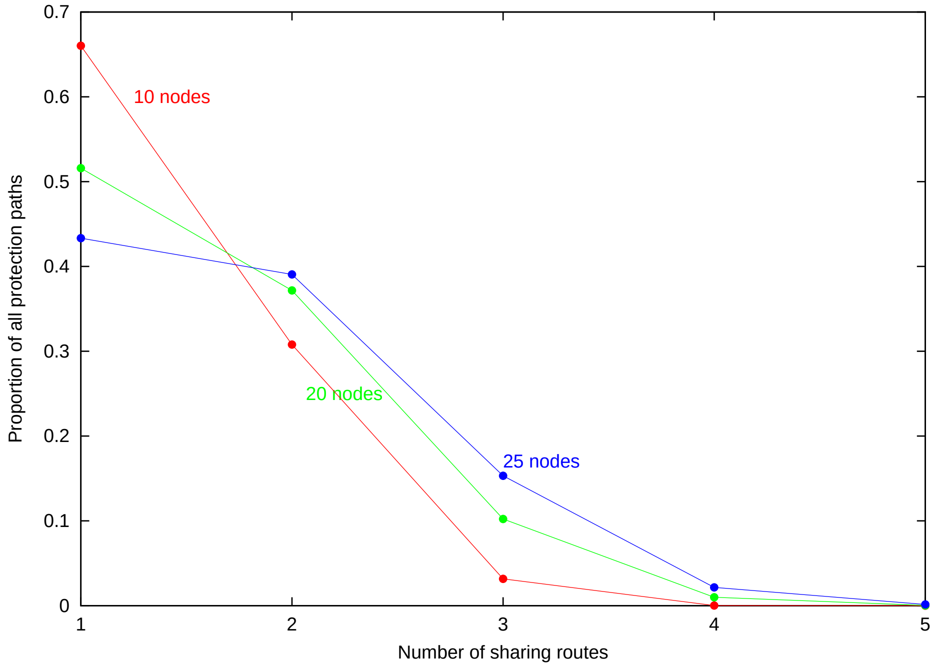

FIGURE 5.8: PROPORTION OF ALL SHARED PROTECTION BANDWIDTH SHARED BY A GIVEN NUMBER OF PROTECTION PATHS, VARYING WITH THE NUMBER OF NODES IN THE NETWORK

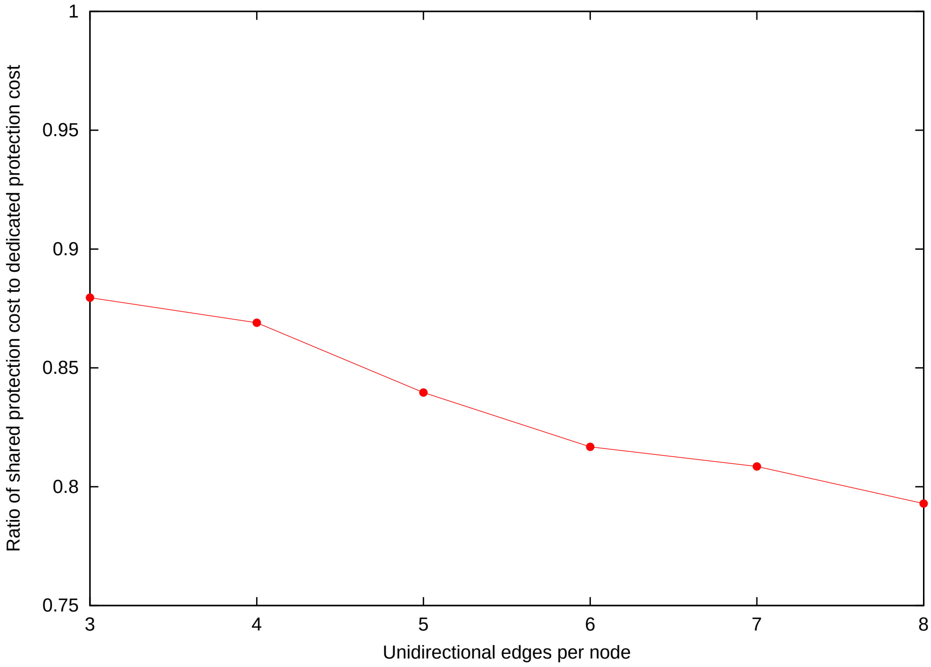

FIGURE 5.9: COST BENEFIT FROM SHARED PROTECTION OVER DEDICATED PROTECTION, VARYING WITH THE CONNECTEDNESS OF THE NETWORK

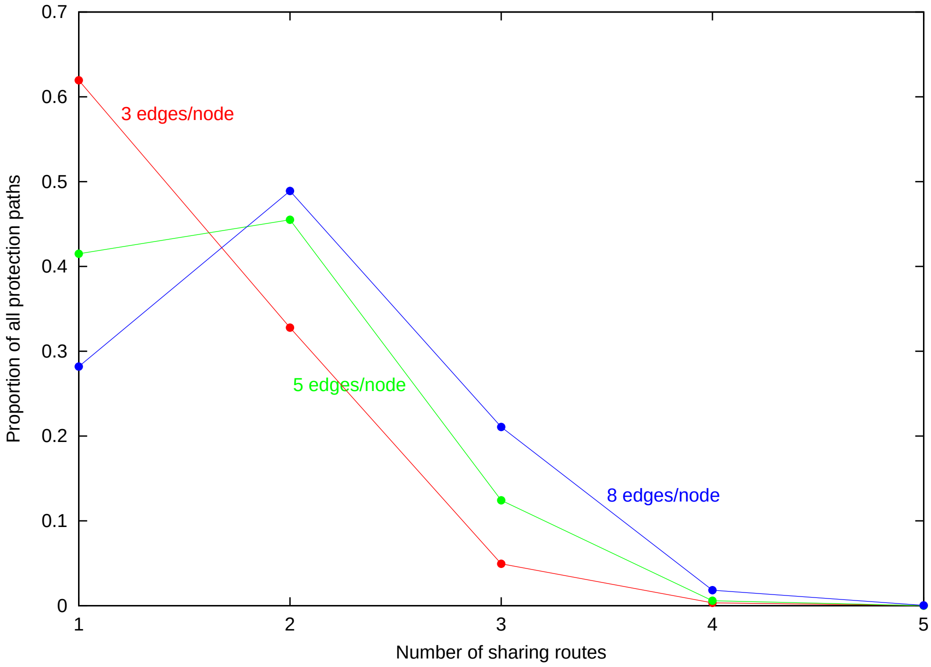

FIGURE 5.10: PROPORTION OF ALL SHARED PROTECTION BANDWIDTH SHARED BY A GIVEN NUMBER OF PROTECTION PATHS, VARYING WITH THE CONNECTEDNESS OF THE NETWORK

cost benefit from using shared protection over dedicated protection as the number of edges per node increases. Note that a figure of three edges per node means that each node is connected to three others on average. Figure 5.10 shows the degree of sharing in the network, which gets progressively greater as the connectedness of the network increases.

These experiments show that while shared protection does require less bandwidth than dedicated protection, for small and relatively sparsely connected networks the gain is limited by the requirement on disjoint primary routes which allows two protection routes to share bandwidth. The disadvantage of shared protection is higher latency when operating a protection switch, since intermediate nodes need to be signalled to put the shared path into operation rather than a single switch at the route's source. For relatively small and sparsely connected networks the gain from shared protection may be more than offset by this operational latency cost.

### 5.1.4 Route Ordering and Waveband Size

Although we are not seeking an optimal solution to the fixed demand problem, it is instructive to examine ways of finding better solutions, still using the routeing algorithm presented in Section 4.2.5. We can find the optimum waveband size for a set of scenarios, but we can also experiment with some of the assumptions used in previous experiments. This will enable us to see whether there is enough potential benefit to warrant trying to remove that assumption.

In the scenarios presented so far, routes have been placed in a random order. This simulates the arrival of traffic in a dynamic network, assuming that network reconfiguration is made incrementally to minimise latency between the stimuli for change and that change being made. We initially consider two alternatives to this: shortest first and longest first ordering. In these we take the set of demands to be routed and order them in terms of their shortest distance path from source to destination. Research has shown [Cinkler00] that shortest route first outperforms random ordering, however with waveband routeing we have the added complexity of forming new waveband paths. Compared to wavelength routeing, waveband routeing can save more resources the longer the waveband paths are, depending on the model of routeing flexibility chosen from Section 4.2.3. Routeing shortest paths first would initially create short waveband paths, so might decrease the average path length and therefore reduce the benefit from using waveband routeing.

The reduction in network cost due to waveband routeing also depends on the size of the wavebands. The scenarios presented so far use a waveband size of 4 wavelengths; fibres of 32 wavelengths were divided into 8 wavebands of 4 wavelengths each. In this experiment we use the notion of dynamic routeing flexibility covered in depth in Section 4.2.3. Essentially we assume an architecture that takes advantage of waveband routeing to decrease the cost of a provisioned network. This places fixed waveband paths in the network and allows routes to be added to the network that use these ex-

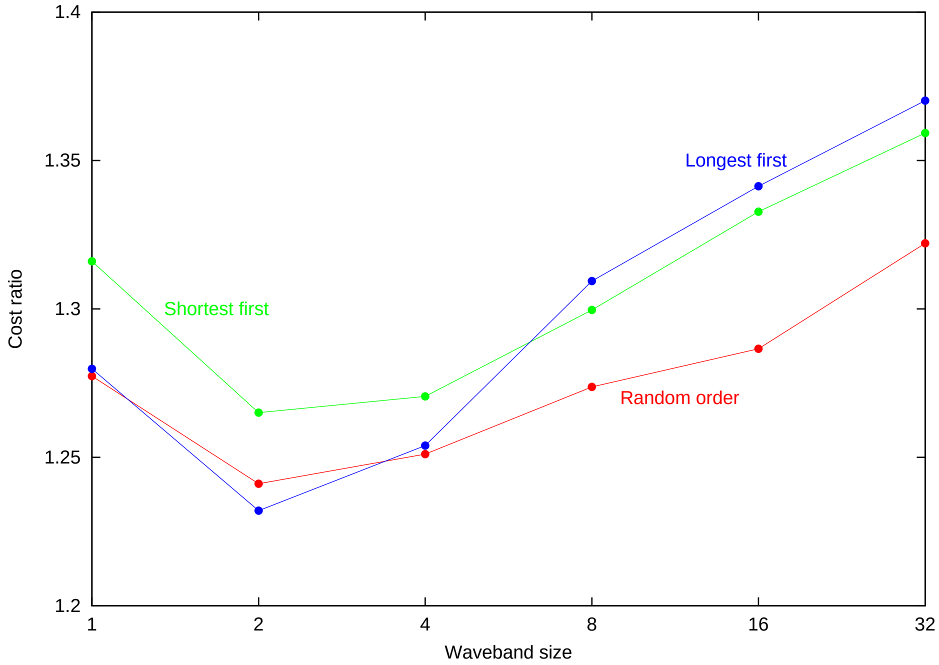

FIGURE 5.11: NO PROTECTION REQUIRED: NETWORK COST VARYING WITH THE WAVEBAND SIZE AND ROUTE ORDERING

isting paths. This reduces the network cost since we do not require lasers, receivers, or wavelength switches for those wavelengths that are part of a waveband path that flows through a node. This allows us to see how the possible cost benefits from waveband routeing are related to the size of the wavebands.

For each of 20 different network scenarios, we first find the *base line cost* to be able to give equal weight to each scenario. Then we select the waveband size, traffic demand ordering, and protection scheme to use. We try waveband sizes ranging from 1, just wavelength routeing, to 32, where we have one waveband in each fibre. We test every waveband size that leads to an integer number of wavebands in a single fibre. The protection schemes used are no protection, shared protection, and dedicated protection. Varying our routeing parameters $RF$, $FF$, and $SC$, we then find the lowest network cost which we divide by the base line cost for this network. These results can then be averaged from all scenarios to find the average cost for that protection scheme, waveband size, and demand ordering.

The results for no protection are shown in Figure 5.11, dedicated protection is shown in Figure 5.12, and shared protection in Figure 5.13. Note that these graphs do not contain standard error bars, where the confidence interval could be calculated by finding the standard deviation of the cost ratios for a given point across all the 20 scenarios tested. This is since the scenarios differ significantly enough in their results to yield large error bars that cover most of the vertical range plotted. A more insightful analysis takes each scenario in turn, and calculates the difference in cost ratio between two points plotted on these graphs. By calculating the mean and standard deviation of these differences

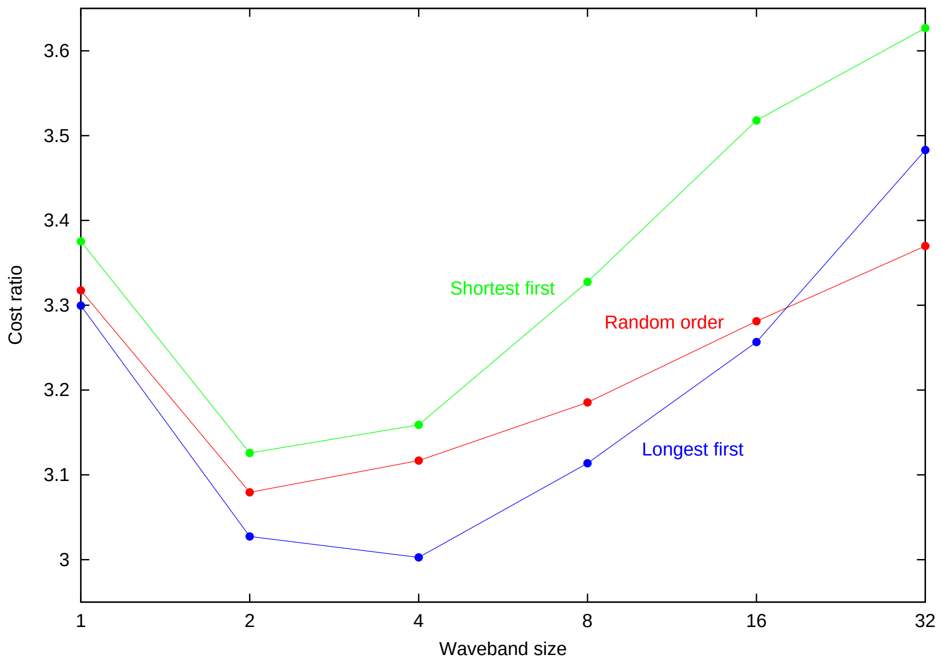

FIGURE 5.12: DEDICATED PROTECTION REQUIRED: NETWORK COST VARYING WITH THE WAVEBAND SIZE AND ROUTE ORDERING

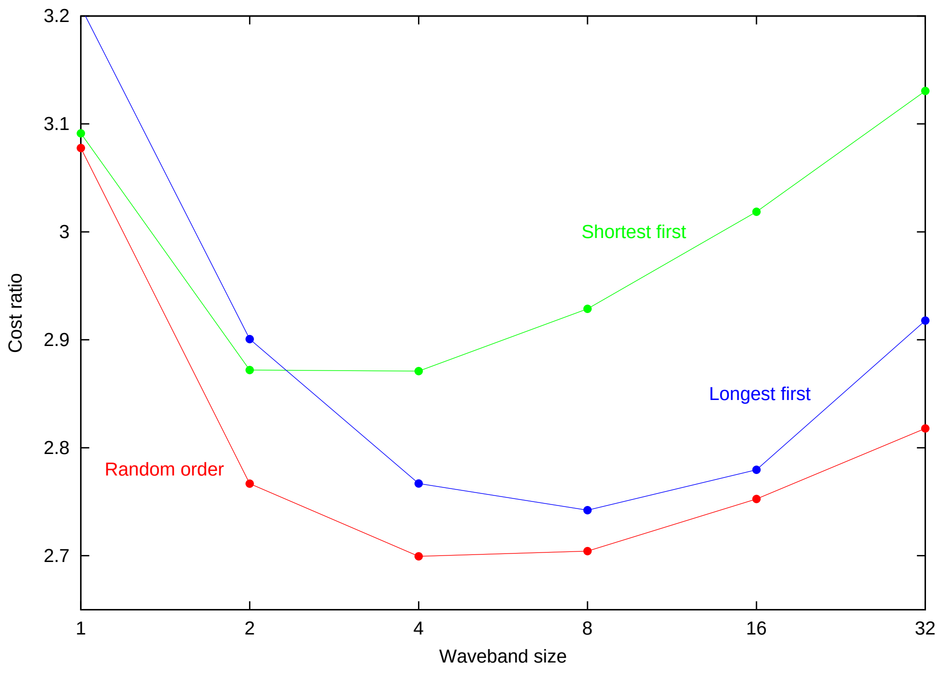

FIGURE 5.13: SHARED PROTECTION REQUIRED: NETWORK COST VARYING WITH THE WAVEBAND SIZE AND ROUTE ORDERING

across all scenarios, we can assess whether there is any significant difference in cost ratio between two points.

We illustrate this statistical test with a few examples on Figure 5.13, the results presented using shared protection. In all cases we calculate the ratio between the mean and standard deviation, and using the Gaussian distribution we find the probability level. This gives us the probability that we are measuring a significant difference, since at this level the mean is different from zero by the given number of standard deviations. When comparing random ordering and longest first ordering with a waveband size of 4, we have a mean difference in cost ratio across all scenarios of 0.067395 with a standard deviation of 0.0465577. Therefore the mean is 1.44756[10] standard deviations away from zero; the chance that this is a valid effect and not a result of sampling error is 85.2%. The difference between random ordering and shortest first ordering at a waveband size of 4 is significant to over 99.9%. Looking in more detail at the random ordering, the difference in cost ratio between a waveband size of 4 and a size of 2 is significant to 95.2%, and between a size of 4 and 8 it is significant to 12.7%. Between a waveband size of 4 and 16, the difference in cost ratio is significant to 70.9%. These results would lead us to believe that it is most likely that random ordering achieves the lowest cost ratios, with either a waveband size of 4 or 8.

From all three graphs is it clear that under this waveband architecture we can reduce the network cost by using waveband routeing; the wavelength routeing case is seen when the waveband size is one. For the no protection case the optimal waveband size is two, for the two protection cases the optimal waveband size is probably four. These experiments motivated the general choice a waveband size of four. This choice may also be motivated by considering the physical optical properties. It has been noted before that using multistage multiplexing brings possible benefits, leading to a waveband size of more than a single wavelength. It may also be true that some optical components have an upper limit on the power contained in a single waveband, leading to an upper bound on the number of wavelengths in a single waveband.

For all three graphs and most sizes of wavebands, routeing shorter paths first leads to the highest network cost. This supports the notion that the benefits from waveband routeing increase as the average length of waveband paths increases. Starting with small routes will tend to create short waveband paths initially, which are then re-used by longer routes — they will tend to be formed from many short paths. This is confirmed by Figures 5.14, 5.15, and 5.16. With a waveband size of 4 and routeing with dedicated protection, the number of waveband paths used for each route was calculated. These were sorted according to the minimum hop length for that route and those proportions were averaged over the 20 scenarios. Figure 5.14 corresponds to random route ordering, Figure 5.15 to shortest routes first and Figure 5.16 to longest first. Comparing the first two of these, for each hop length the shortest first graph has a more concentrated distribution: there are very few long routes taking a single waveband

[10] This figure of 1.44756 is called an *effect size*; a qualitative scale of effect sizes is 0.0: trivial, 0.2: small, 0.6: moderate, 1.2: large, 2.0: very large, 4.0: nearly perfect, $\infty$: perfect [Hopkins00]. In the results reported in this work we translate effect size into probability using the Gaussian distribution.

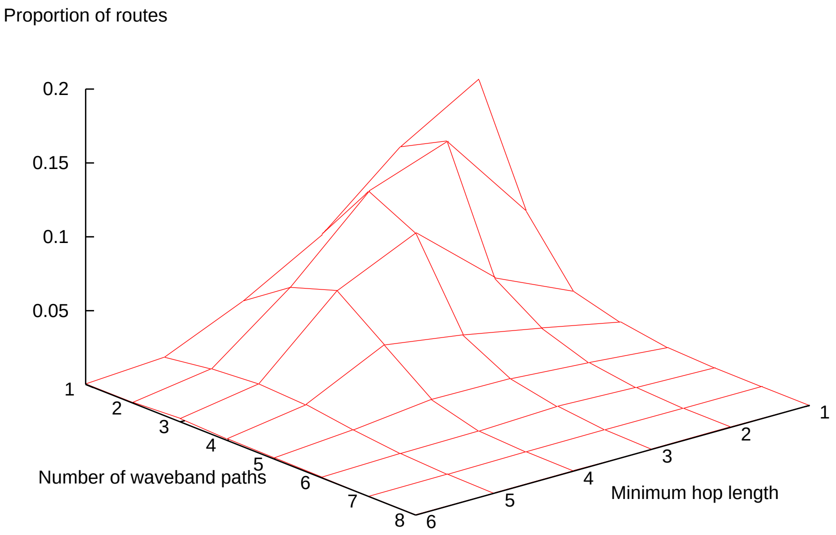

FIGURE 5.14: RANDOM ROUTE ORDERING: PROPORTIONS OF ROUTES HAVING A GIVEN MINIMUM HOP LENGTH AND CONSTRUCTED FROM A GIVEN NUMBER OF WAVEBAND PATHS

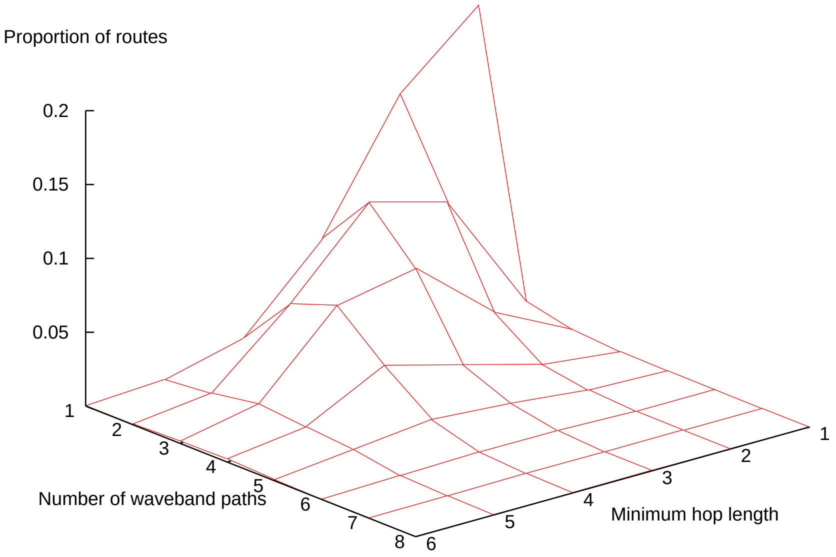

FIGURE 5.15: SHORTEST FIRST ROUTE ORDERING: PROPORTIONS OF ROUTES HAVING A GIVEN MINIMUM HOP LENGTH AND CONSTRUCTED FROM A GIVEN NUMBER OF WAVEBAND PATHS

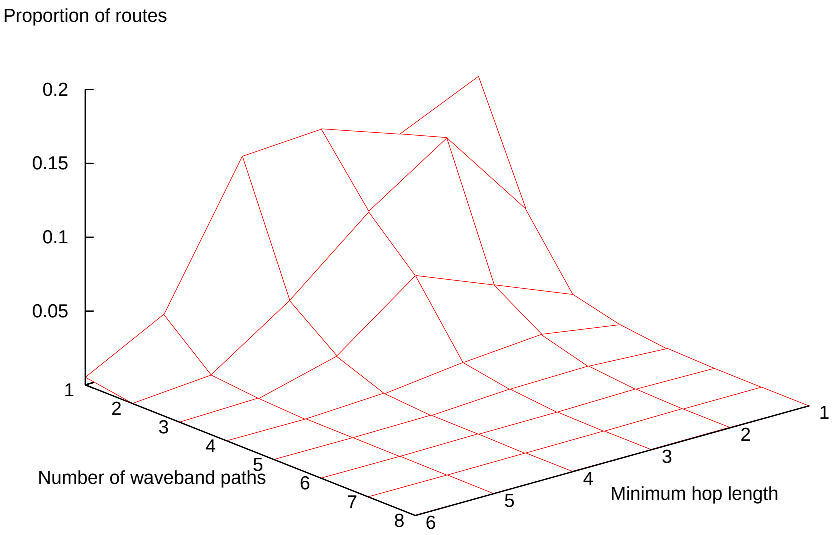

FIGURE 5.16: LONGEST FIRST ROUTE ORDERING: PROPORTIONS OF ROUTES HAVING A GIVEN MINIMUM HOP LENGTH AND CONSTRUCTED FROM A GIVEN NUMBER OF WAVEBAND PATHS

path, but conversely not many routes go through a large number of paths. There is a marked change for longest first ordering, since many of the long routes just use a single waveband path.

Taking advantage of waveband routeing by creating longer paths explains why, for no protection and dedicated protection routeing, longest routes first gives the optimal results. However, as the waveband size increases, this turns into a disadvantage, with random ordering giving better results than longest first. This is likely to be where the utilisation of these long waveband paths has an effect: long waveband paths are created with spare capacity which can only be used as part of short routes that must back-route at either end of the long waveband path.

As seen in Figure 5.13 the shared protection case also falls foul of the large number of longer waveband paths. Since the unit that we share is a wavelength for the duration of a waveband path we are limited to either sharing a long path or creating a new path that will follow a more direct route. For this reason random ordering, which seems to have a mixture of long and short paths gives the best results. This is useful since it matches the situation when demand is changing, since new routes to be added will arrive with no specific ordering.

## 5.2 Fixed Topology Problems

We now present a series of experiments considering the *fixed topology problem*, where a dynamic demand process is routed across a fixed network. Section 5.2.1 gives the methodology used in these types of experiments.

We compare the different models of network flexibility outlined in Section 4.2.3, namely the FULL, PATH, and FIXED architectures. The FULL architecture has all wavelengths being switched at each node, and each of these wavelengths may originate or terminate at this node. The PATH architecture has a fixed set of waveband paths across the network, created during the network provisioning system, and can create wavelength routes using a series of these paths. The FIXED architecture has a fixed number of wavelength routes available and relies on over-provisioning rather than reconfigurabilities for performance.

In Section 5.2.2 we compare the blocking probability for new routes in these different models, and their response to altering the dynamic traffic mix progressively away from that used to provision the network. In Section 5.2.3 we perform a cost sensitivity analysis, since the trade-off between different architectures depends on the relative network costs presented in Table 4.2.

Section 5.2.4 presents a complementary method of determining routeing flexibility based on using a limited number of tuneable lasers and receivers, rather than large banks of fixed frequency equipment.

### 5.2.1 Methodology

For a given network topology and traffic mix, we will mainly be comparing the impact of a change in the design of the network on the performance. This means that for each network design we need to provision a network with equal cost. We achieve this by using the traffic scaling parameter, $\beta$, introduced in Section 4.3.2. For one network design we set a value of $\beta$, and measure the network cost for the provisioned network. For other network designs we find a value of $\beta$ which yields a network with the same cost.

We use $\beta = 40$ for the FIXED architecture to set the network cost for a particular experiment. For the other more complex network architectures the value of $\beta$ will be lower in order to achieve the same cost.

For example, on one representative scenario, $\beta = 40$ yields 2610 pairs of primary and protection routes and the optimised cost for the FIXED architecture was calculated. For the PATH architecture we have to set $\beta = 18$, 1175 routes, achieves the same network cost. For the FULL architecture a further reduction of demand to $\beta = 14$, 914 routes, gets the same cost. Using this method we can accurately compare the relative performance of different network architectures.

Each provisioned network was then subjected to a dynamic demand stage, with a series of scaled demands. Here all architectures were given the same demand, the $\beta$ value of the subjected traffic shown in the x-axis of results presented. Routes were added and removed from the network continuously, and the proportion of routeing attempts which fail forms the blocking probability of the network. After an initial period of operation to remove start-up transient behaviour, the experiments carried on until either the 99% confidence interval on the blocking probability fell to under 0.2 relative error, or an upper limit on the length of the experiment was reached. This upper limit was set to achieve a 99% confidence interval with 0.2 relative error at a blocking probability of $10^{-4}$. Results from several scenarios were averaged to dampen effects of any one scenario and to reveal general trends.

### 5.2.2 Comparing Flexibility Architectures

The results presented here and in the next section are for four different network architectures. These are the FULL, PATH, and FIXED architectures, and a variant on PATH. PATH-1 is as previously described, where waveband paths are created while the network is provisioned. However this may leave some fibres with spare capacity. In PATH-2 after all routes have been placed, spare fibre capacity is filled by adding in extra waveband paths with the longest possible paths added first. This ensures that all fibres can be fully utilised. To achieve the same cost as PATH-1 a slightly smaller fraction of the original traffic demand is used when provisioning.

For all these experiments shared protection routes were created for all new routes. The waveband size was fixed at 4, which previously was found to be the best size when using shared protection.

The traffic mix of the dynamic process is the subject of this set of experiments. In the first, shown in Figure 5.17, the dynamic traffic mix is an exact proportion of the traffic mix used for provisioning. During the dynamic process each pair of nodes requires a mean quantity of bandwidth that is a linear scale of the number of paths that were provisioned. We can see that partial reconfigurability is better than full flexibility, since we obtain a lower blocking probability for a given traffic demand. However, the route-based architecture achieves even better results.

Since we would expect the traffic mix to change over time, either on a permanent or periodic basis, Figures 5.18, 5.19 and 5.20 show the results of applying a different dynamic traffic mix from that used to provision the network. In each case a new traffic mix, uncorrelated from the provisioning mix, is created and scaled to have the same total average demand. The mix used for the dynamic process is a proportion of the new mix, $N$, together with a proportion of the original, $(1 - N)$, keeping the average demand at the same level. We expect that the case of $N = 0.25$ is most realistic since research has shown that considering traffic demands by rank during a 24 hour period, rankings first spread then converge back [Feldmann00], but are always strongly correlated to the original ranking. Adding in a previously unseen traffic mix allows measurement of the response to unknown or new factors causing network load.

These results show that when confronted with an unexpected traffic mix, the FIXED architecture suffered greatly from its lack of flexibility. Although the gain over full flexibility decreases, the architecture proposed in Section 4.2.3 of partial flexibility is still measurably better over all proportions of traffic: by around two orders of magnitude at $N = 0.25$, an order of magnitude at $N = 0.5$, and a factor of two at $N = 1$.

Similarly to the results reported in Section 5.1.4 standard error bars are not shown on these graphs. For each scenario, each point was calculated to a relative error of 0.2 when the blocking probability was more than $10^{-4}$. However the differences between blocking probabilities across scenarios is large, giving rise to large error bars for the mean points plotted. We perform a similar statistical analysis of the results presented here as before. We present a representative analysis taken from Figure 5.18 where $\beta = 13$. The probability that the plotted difference between the PATH-1 and FULL architectures being significant is 88.3%, the difference between PATH-2 and FULL is 90.3% significant. The minor difference between PATH-1 and PATH-2 has only 9.8% probability of being significant; the difference between FULL and FIXED has only 24.7% probability of being significant.

### 5.2.3 Cost Sensitivity Analysis

A cost sensitivity analysis was performed by looking at the four types of cost involved in provisioning a network. The experiments shown in Figure 5.18 were repeated with each of the four costs first being increased by 50%, and then decreased by 50%, whilst the other three costs remained the same.

Since we are interested here in the change in relative performance of the different architectures we present the results in terms of the relative traffic load that can be supported at a given blocking probability. This is a relevant figure since network operators customarily design networks to operate at a given blocking probability. FULL is taken as a benchmark, so appears as a constant line at 1.

Figure 5.21 shows the effect of changing the cost of multiplexing. Since this cost is a small proportion of the total cost, the effects are small. The smooth lines show the relative performance change for no cost change, showing that here, for the same infrastructure cost, partial flexibility can take a 40%–50% increase in traffic load for the same blocking probability from full flexibility. The up and down arrows show the change after the cost of multiplexing increases and decreases respectively. The results for the partial flexibility architectures, PATH-1 and PATH-2 are shown offset from each other, with PATH-1 on the left and PATH-2 on the right of each pair. All pairs of these results are for the same blocking probability, and merely offset for clarity.

The effect of changing the cost of transmitters and receivers is shown in Figure 5.22. We indicate on the figure an example point to describe. The point is at a blocking prob-

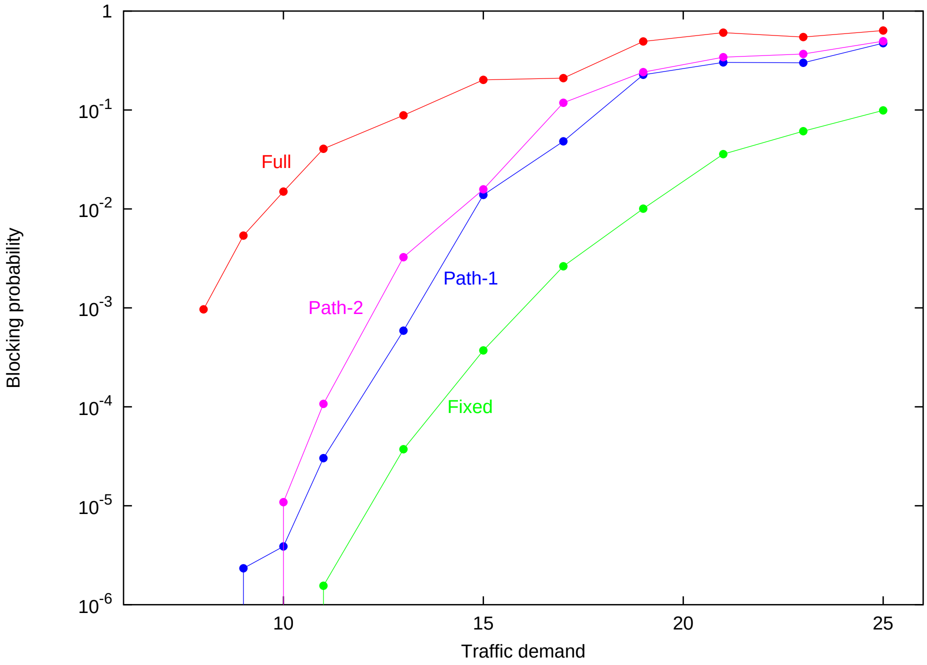

FIGURE 5.17: BLOCKING PROBABILITIES FOR PROVISIONED TRAFFIC MIX

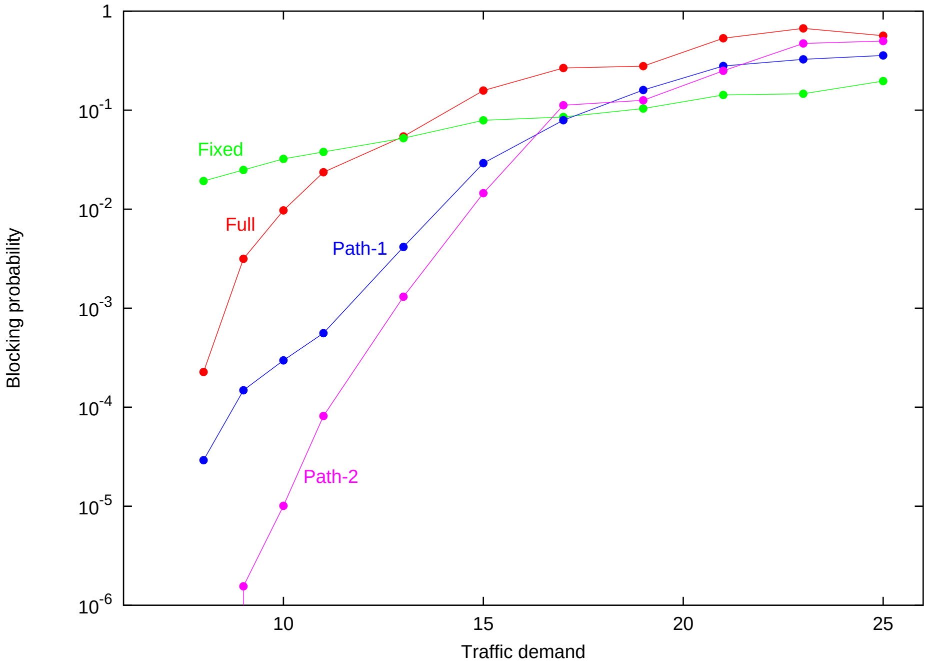

FIGURE 5.18: BLOCKING PROBABILITIES FOR 25% NEW TRAFFIC MIX

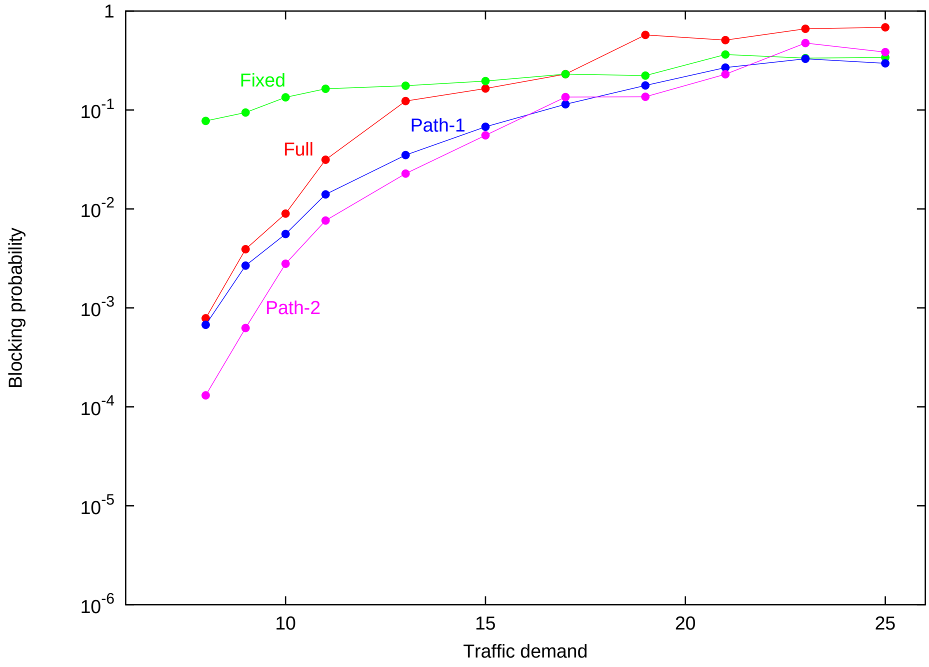

FIGURE 5.19: BLOCKING PROBABILITIES FOR 50% NEW TRAFFIC MIX

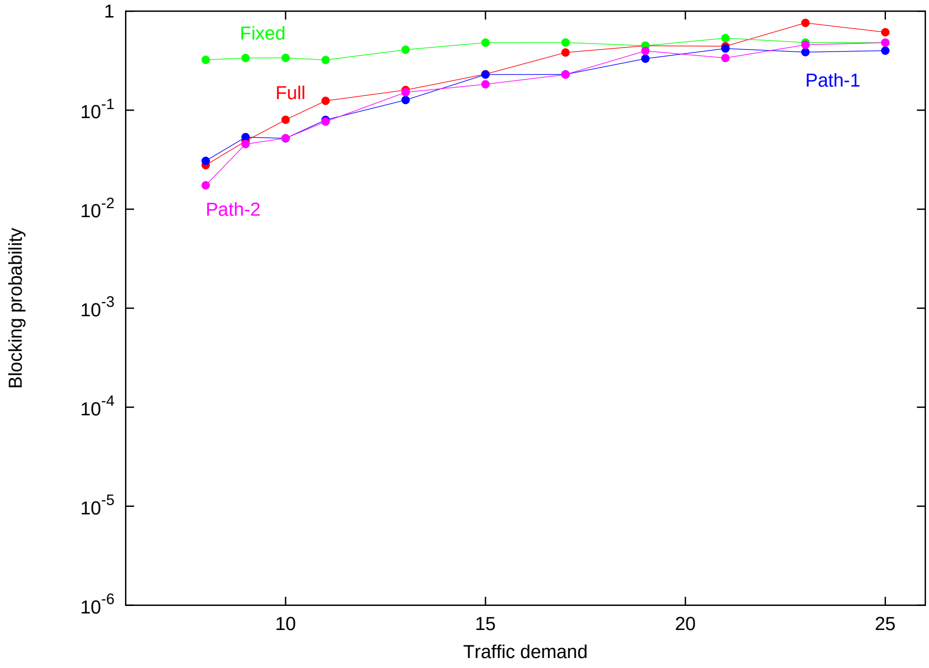

FIGURE 5.20: BLOCKING PROBABILITIES FOR 100% NEW TRAFFIC MIX

ability of 0.001, is a downward arrow symbol, is a result from the PATH-1 architecture, and is around 1.07 on the y-axis. This point indicates that if the cost of transmitters and receivers falls by 50%, the other equipment costs remaining the same, that when achieving a blocking probability of 0.001 the PATH-1 architecture can handle 7% more traffic than the FULL architecture. This is a fall from around 45% more traffic, with the original equipment costs, as shown by the curved line. So as this cost falls, the PATH-1 architecture has less of an advantage over the FULL architecture.

In general in Figure 5.22 the relative inefficiency of using FULL increases as the cost of opto-electrical conversion increases and vice versa. This is to be expected since the gains from using PATH are due to reducing the number of these sorts of components relative to the quantity of fibre used. Figure 5.23, where we change the optical switching cost, shows a similar trend for the same reason. It is interesting to note that in the two cases a different PATH architecture out-performs the other, showing that there is no single optimum architecture for all relative costs. In both cases the worst-case performance of the best algorithm results in at least a 25% increase in traffic capacity over FULL.

The effects of changing fibre and amplifier cost are shown in Figure 5.24. Here the benefits of using PATH or FIXED increase as this cost decreases, and as the fibre and amplifier cost increases, the FULL architecture performs relatively better. This is expected since the FULL architecture is likely to have a higher utilisation of fibre, so requires less of it to satisfy the same demand. Here the worst-case performance of the PATH-1 architecture is within 5% of the performance of the FULL architecture.

The original cost estimates used are the current best guess. It is difficult to say in which direction relative prices will move; however, it has been shown that for a wide range of relative costs, the PATH architecture through waveband routeing results in significant increases in traffic capacity over FULL or FIXED. We would estimate that the benefits of a partial flexibility architecture would be maintained until a decrease of around 75% in the relative cost of either optical switches or opto-electrical conversion, or to a 50% increase in the relative cost of fibres and amplifiers.

### 5.2.4 Tuneable Laser Flexibility

An alternative way of decreasing possible equipment cost for a waveband routed network would be to reduce the number of laser transmitters at a given node, replacing them with tuneable lasers which are able to tune to a range of different frequencies. This would be of benefit if in general the quantity of traffic originating from any given node were smaller than the fibre capacity present at this node. This has partly been addressed by the partial flexibility architectures proposed in Section 4.2.3 where we have transmitters only at the head of waveband paths. Using tuneable lasers would further reduce the quantity of transmitters, allowing the equivalent budget to increase capacity of other areas of the network.

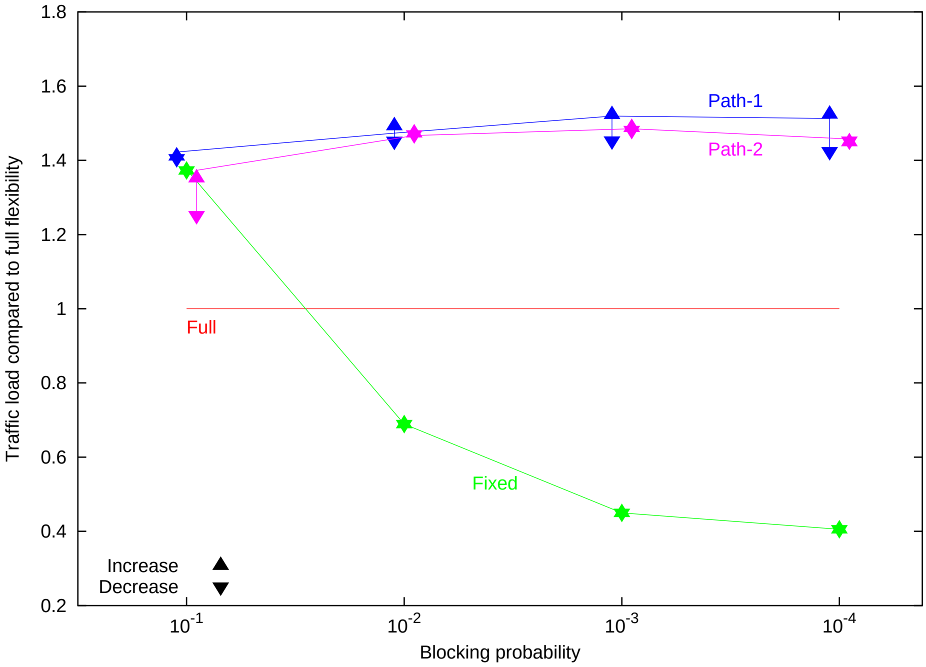

FIGURE 5.21: RESPONSE TO A CHANGE IN MULTIPLEXING COST. VERTICAL BARS INDICATE THE LOAD RATIO AFTER A COST CHANGE

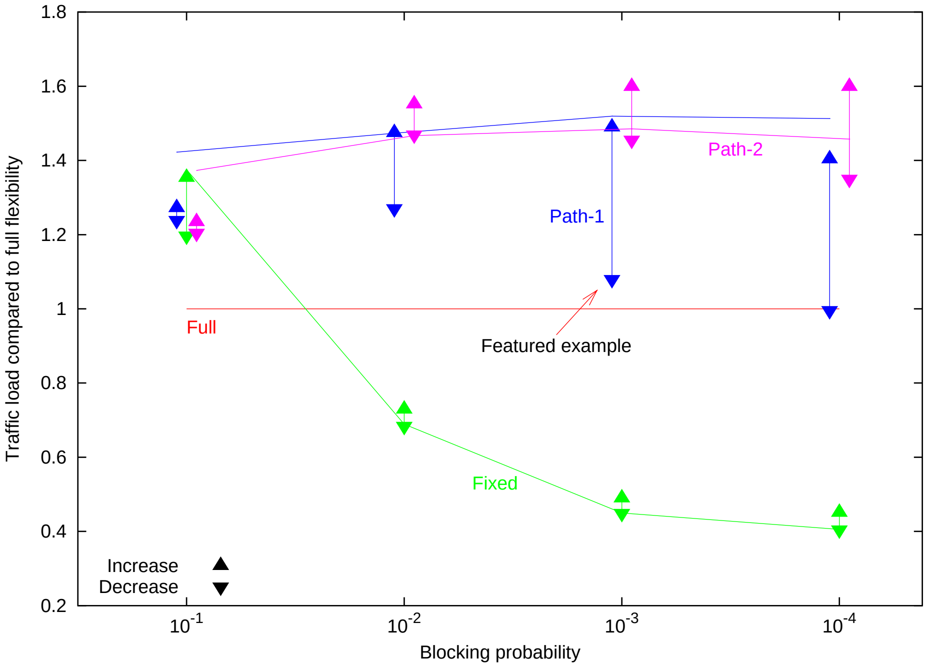

FIGURE 5.22: RESPONSE TO A CHANGE IN TRANSMITTER/RECEIVER COST. VERTICAL BARS INDICATE THE LOAD RATIO AFTER A COST CHANGE

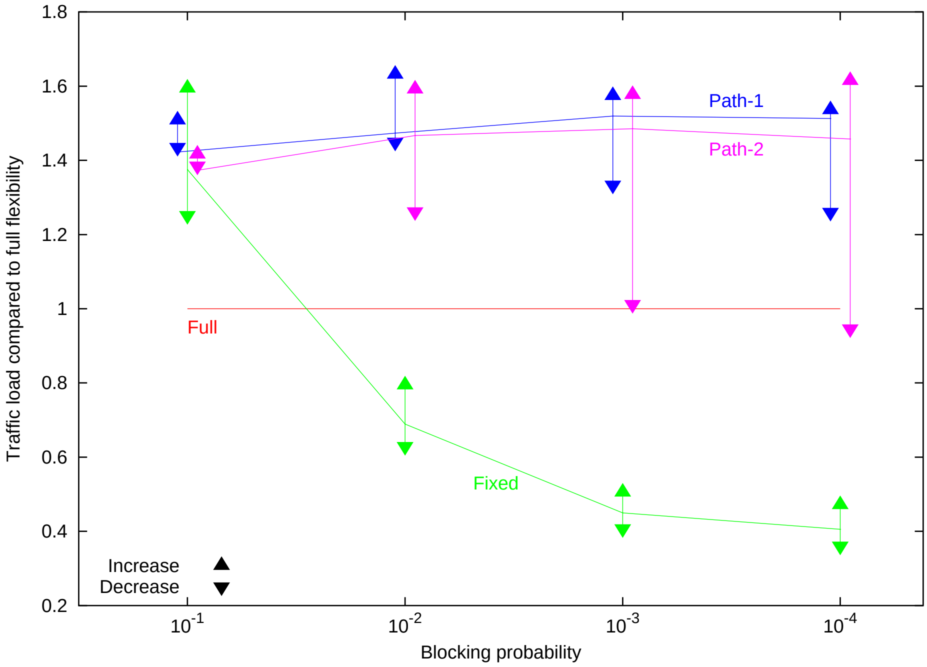

FIGURE 5.23: RESPONSE TO A CHANGE IN OPTICAL SWITCHING COST. VERTICAL BARS INDICATE THE LOAD RATIO AFTER A COST CHANGE

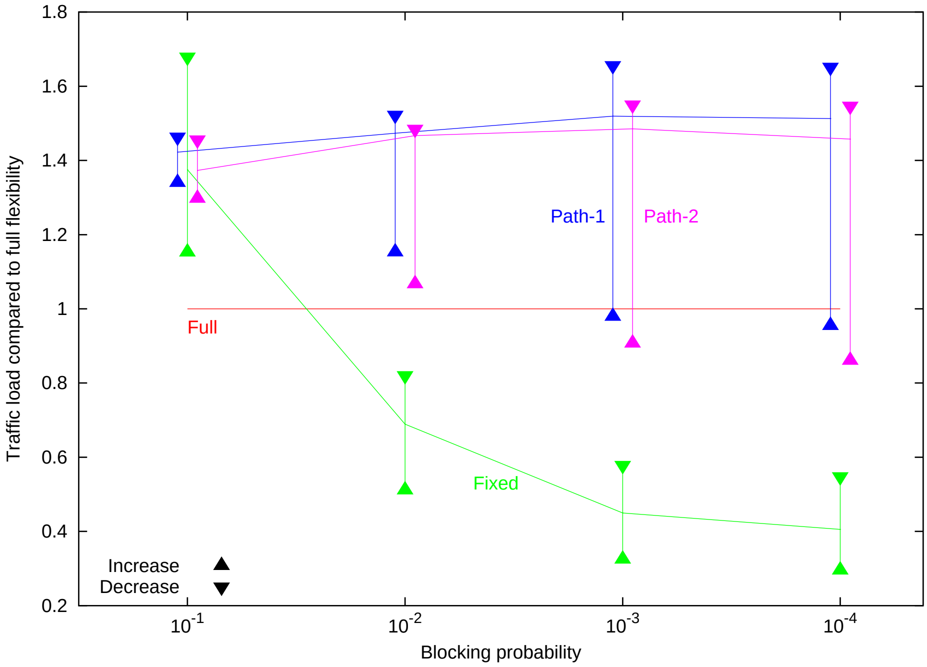

FIGURE 5.24: RESPONSE TO A CHANGE IN FIBRE AND AMPLIFIER COST. VERTICAL BARS INDICATE THE LOAD RATIO AFTER A COST CHANGE

Whilst some types of tuneable laser can span a whole window of transmission frequencies, it is much more simple to restrict tuneable lasers to a small set of wavelengths; we consider the case where transmitters are able to tune to any wavelength in a given waveband. This assumes that wavelengths in a single waveband are contiguous, which is most likely to be the case. For a node in the PATH architecture where we have $n$ waveband paths originating here for a given waveband we would ordinarily have $n$ transmitters for each of the wavelengths in the waveband. If wavebands contain 4 wavelengths each, we have $4n$ single frequency transmitters. In our new model we replace these by $\lfloor 4n\gamma \rfloor$[11] tuneable transmitters, $0 < \gamma < 1$. Similarly at the termination of $m$ waveband paths we assume $\lfloor 4m\gamma \rfloor$ tuneable receivers.

Although it would depend on the exact hardware used to create a tuneable transmitter or receiver, it is very likely that these components will cost more than their fixed counterparts. This is partly for the facility to be able to tune the laser, but also for the extra switching arrangements needed to connect the transmitter bank to the correct set of switches. For example if we have a single switch per wavelength, the model assumed here, then a single tuneable laser would in general need to be connected to an input port on a wavelength switch for each wavelength in the waveband. If we assume a monolithic switch architecture, then this extra cost is reduced; the optimum approach may be for a single MEMS type switch handling the wavelength switching for each wavelength in a given waveband.

We term the ratio of cost increase from a fixed component to a tuneable component $\delta$. Clearly if $\delta\gamma > 1$ then our bank of tuneable components costs more than the previous bank of fixed components, so we will have to provision our network with a smaller traffic load to attain the same final cost. This, together with the possible increase in blocking due to a smaller pool of transmitters at each node, is certain to increase the final blocking ratio.

A set of experiments was carried out using our model of tuneable transmitter and receivers. Several values of $\gamma$ and $\delta$ were used, with the results presented averaged from four different scenarios. Within each scenario all networks were provisioned to the same cost, the traffic load used for provisioning was the maximum possible to support given the value of $\gamma$ and $\delta$.

Figure 5.25 shows the blocking probability achieved by using tuneable components, compared with that achieved by the default algorithm, shown in Figure 5.18. These results are with the traffic scaling parameter set at $\beta = 13$: a range of values of $\beta$ were investigated, and $\beta = 13$ gives representative results. Four different values for $\delta$ were investigated, which are shown as four curves, each labelled with the percentage cost increase. The x-axis shows $\delta\gamma$, indicating the relative cost of a bank of tuneable components versus the equivalent fixed components' cost. A value of 1 indicates that $\delta\gamma = 1$: since the costs are the same, networks are provisioned with the same traffic capacity. Results to the left of this have a higher provisioned network traffic capacity, due to the decreased total cost of tuneable components, and conversely for points to

[11] $\lfloor x \rfloor$ returns the largest integer value smaller or equal to $x$

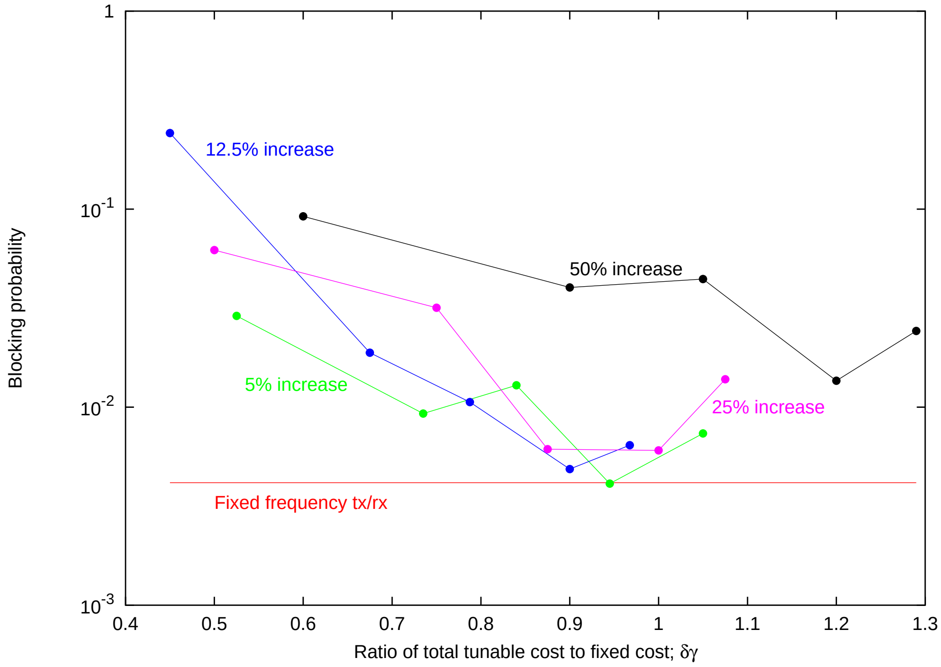

FIGURE 5.25: BLOCKING PROBABILITY ACHIEVED BY USING TUNEABLE TRANSMITTERS AND RECEIVERS, VARYING THE DENSITY AND COST COMPARED TO FIXED FREQUENCY EQUIPMENT

the right.

It is clear that as the cost increase factor $\delta$ increases, our best possible solution becomes worse — we have a higher minimum blocking probability. For smaller values of $\delta$, the lowest blocking probability is very close to that obtained by using fixed frequency components. For other values of $\beta$ tested, similar results were obtained; the minimum blocking probability for tuneable frequency components were achieved where $\delta$ is small and that minimum was very close to that obtained by fixed frequency components. With current technology it is estimated that the value of $\delta$ is perhaps as low as 1.10 [Rigby03]. This leads us to conclude that the use of tuneable frequency components is not of any benefit when used in the manner tested in this experiment.

With technological advances it may be possible for $\delta$ to decrease, rejuvenating interest in the approach outlined in this section. An alternative possibility is that tuneable frequency components are used as an alternative to waveband routeing, since both could decrease the equipment cost at a network node by reducing the number of transmitters and receivers. This would require some scheme for coping with the small range of obtainable frequencies from any single component compared to the number of wavelengths available, or further technological advances in extending this tuning range. Other benefits not measured here such as the reduction in space taken by physical components may also impact on the benefits from using tuneable components.

## 5.3 Network Optimisation Techniques

We now present two extensions to the results presented in this chapter. Firstly in Section 5.3.1 we change the ordering in which we provision a network, in a more radical fashion than that explored in Section 5.1.4, by initially routeing full wavebands. Secondly an option to improve the capacity of a dynamically reconfiguring network, especially where routes are added incrementally with only local optimisation, is to reorganise existing routes periodically to a more optimal configuration. Sections 5.3.2 and 5.3.3 outline algorithms for doing this, and assess the benefits to be gained from this approach.

### 5.3.1 Two-stage Waveband Routeing

A more radical method of route ordering can take better advantage of the idea of limited dynamic routeing flexibility, used in Section 5.1.4, where we used a fixed set of waveband paths to reduce the cost of the provisioned network.

Given that a longer average waveband path length can give a larger reduction in cost, the aim of *two-stage* waveband routeing is to maximise the length of paths. Considering we require $d$ wavelengths between a given source and destination pair where our wavebands are $w$ wavelengths each, we first route $\lfloor d/w \rfloor$ wavebands from source to destination. We perform this waveband routeing stage for all source destination pairs. After this stage we have $(d \bmod w)$ wavelengths left between each source and destination pair, which we route individually as before.

For the second stage of this algorithm, we multiply the new fibre factor, $FF$, by a constant. This means that the second stage of our routeing process can try to avoid adding more fibres to our network, preferring to increase the utilisation of existing resources, whereas the first waveband routeing stage should prefer short efficient routes. Several values of this constant were tried, with an $FF$ increase by a factor of 4 being the most common optimal value.

Since the benefits from this approach are highly dependent on the relationship between $d$ and $w$ we experimented with a range of values for each. The results are shown in Figure 5.26, averaged from the 20 scenarios used previously. The traffic scaling parameter $\beta$ is given, which was defined in Section 4.3.2; the average number of wavelengths per source and destination pair will be $\bar{d} = \beta/4$. In this experiment routes were placed without associated protection routes. The relative cost reported in the results is the cost of the two-stage provisioned network divided by the cost of a normally provisioned network, and this ratio is averaged over the different scenarios tested.

Firstly when the waveband size is 1, there is no difference between the normal and two-stage algorithms, given a relative cost of 1. For each level of demand, the advantage from the two-stage algorithm increases as the waveband size increases, before de-

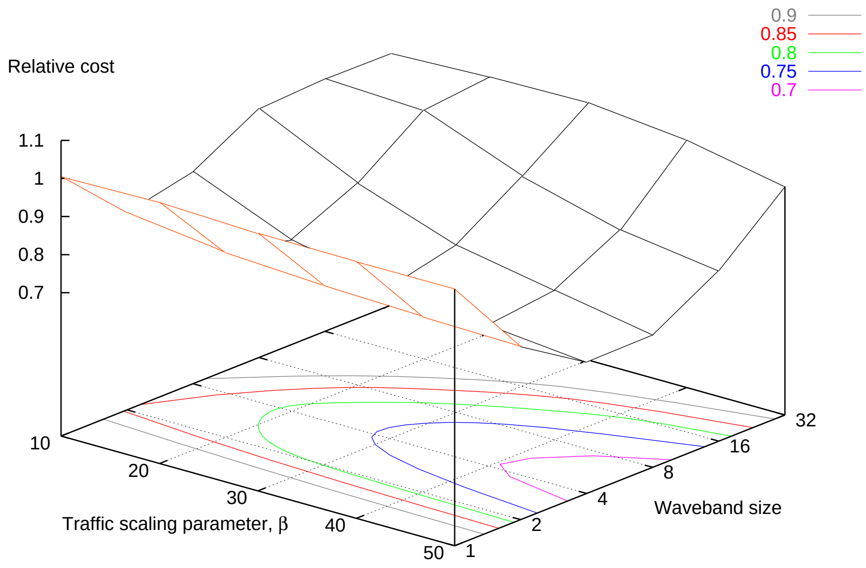

FIGURE 5.26: COST BENEFIT FROM TWO-STAGE WAVEBAND ROUTEING OVER WAVELENGTH ROUTEING, VARYING WITH THE QUANTITY OF TRAFFIC AND THE WAVEBAND SIZE

creasing. The optimal waveband size increases slightly as the demand increases, from around 2 with low traffic load which is comparable to results presented in Figure 5.11, to just over 4 with a high traffic load. Whilst a considerable reduction in network cost can be achieved by this method, the resulting network is likely to be less reconfigurable than when most routes are composed from multiple waveband paths.

To assess this loss of flexibility we re-run the set of experiments in Section 5.2.2 that assessed the benefits of different architectures under a range of traffic mixes. We now compare these with a network provisioned using the two-stage waveband routeing process outlined above. Since these networks are cheaper, we equalise the final network cost by increasing the traffic scaling parameter $\beta$. We recall that $\beta$ is the maximum number of wavelength routes required between two nodes, where the average number of routes required will be $\beta/4$. When used in provisioning traffic, increasing $\beta$ increases the capacity of the network. We are using the same methodology for this experiment as explained in Section 5.2.1.

To achieve the same network cost as the standard network using $\beta = 40$, a two-stage waveband routed network required around $\beta = 70$. These two networks were then subjected to the same dynamic traffic demand to measure the resulting blocking probability. The traffic mix used was a proportion, $(N - 1)$ of the traffic mix used for provisioning, added to a proportion $(N)$ of a new random traffic mix. We experiment with $N = 0.25$ and $N = 0.50$; since two-stage routeing decreases the flexibility of our network by decreasing the number of waveband paths, it may be suited better to situations where the dynamic traffic mix is closer to that used for provisioning.

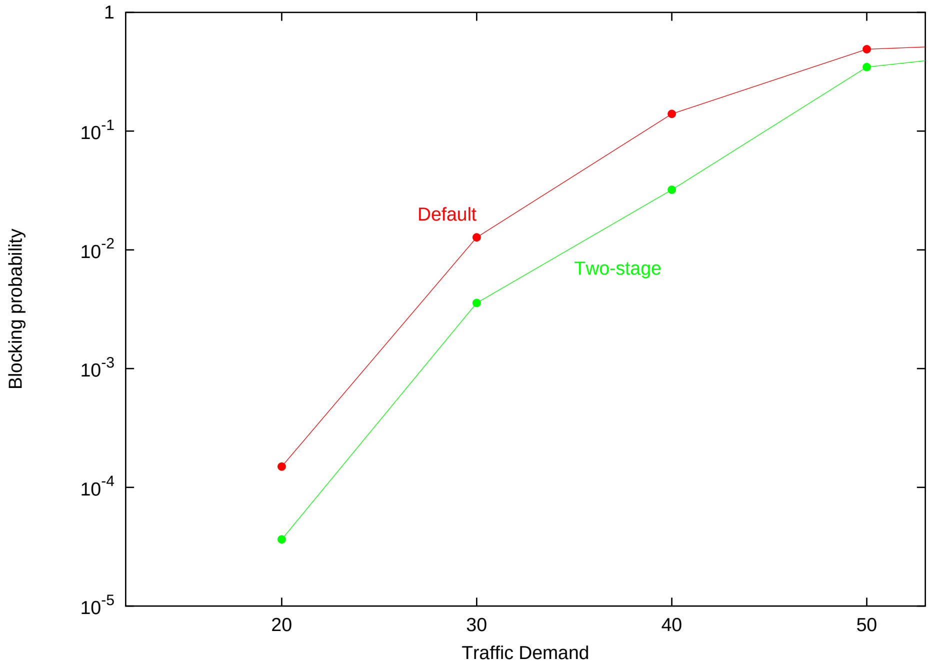

FIGURE 5.27: COMPARING THE BLOCKING PROBABILITY ACHIEVED WITH NORMAL AND TWO-STAGE PROVISIONING, USING A 25% PROPORTION OF A NEW TRAFFIC MIX

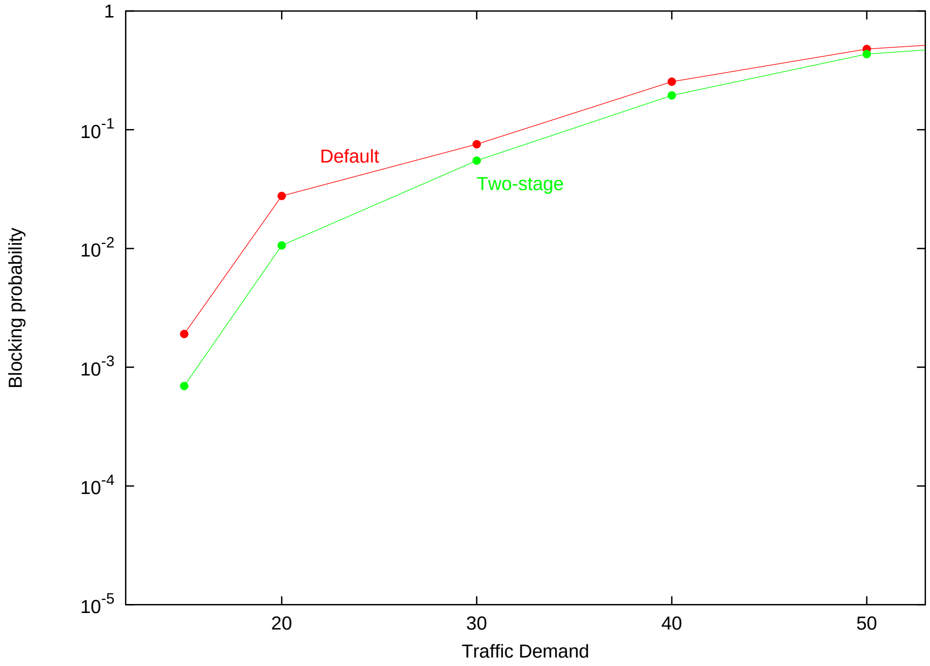

FIGURE 5.28: COMPARING THE BLOCKING PROBABILITY ACHIEVED WITH NORMAL AND TWO-STAGE PROVISIONING, USING A 50% PROPORTION OF A NEW TRAFFIC MIX

Figure 5.27 shows the blocking probabilities achieved by the default and two-stage algorithms at a range of different traffic loads. The traffic mix used here was 75% of the traffic mix used for provisioning, and 25% of a new mix; Figure 5.28 shows the same results for a 50/50 mix. In both, the large cost decrease brought by two-stage provisioning does outweigh the decrease in routeing flexibility, giving a lower blocking probability for an equal cost network. The advantage in two-stage provisioning is slightly smaller for the experiment using a larger unknown traffic mix; in a two-stage provisioned network waveband paths are generally longer but orientated to the provisioned traffic mix so it will suit this traffic mix better.

### 5.3.2 Rearranging During Provisioning

In a dynamic system where routes arrive and depart over time, new routes being placed with only local optimisation, it is likely that at any one point in time the current set of routes will be placed sub-optimally. For example a route could be routed around some congested part of the network that later becomes less congested; this route could be made shorter. This sub-optimality would then lead to a higher blocking percentage than necessary since the existing routes are not utilising resources efficiently. A method of addressing this could be to perform an online optimisation of rearranging existing paths, undertaken during slack periods of control traffic.

In the following work we assume that if we are to remove a route and replace it with another alternative, then the new route may use resources currently in use by the existing route. Because of the possible limitations in setting up new wavelength route this is may not be possible since it would result in a loss of service whilst the new route is being configured. We make this assumption since we will get an upper-bound on the possible gain from rearranging routes. Care is also needed if components such as optical switches are only rearrangeably non-blocking, since when reconfiguring a switch existing paths may have to be disrupted for a brief period to achieve the desired switch configuration.

To develop an understanding of how much rearranging routes may help, we first perform this rearranging algorithm on a static network, measuring the benefits by the reduction in network cost during the provisioning process. In Section 5.3.3 we will then move to performing a similar algorithm on a dynamic network.

Our algorithm is based on the understanding that we want to maximise our gain by changing the fewest routes. A way to achieve that is to find fibres which are under-utilised and attempt to remove those fibres from our provisioned network; if a fibre contains just a single route, then we could remove that fibre and have to only rearrange a single route.

The algorithm contains the following steps. At each stage we order the fibres based on their utilisation: how many routes go through this fibre. We take the fibres in turn, starting with the least utilised. As a tie-break on utilisation we consider the longest

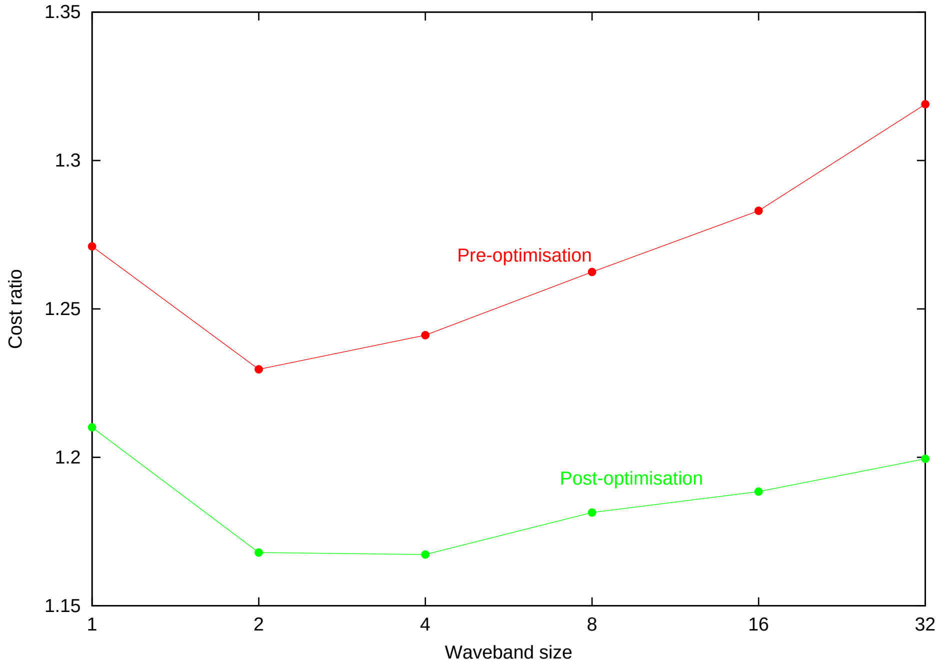

FIGURE 5.29: THE COST RATIO BEFORE AND AFTER REARRANGING ROUTES, BY WAVEBAND SIZE

fibre first. We remove the fibre and all routes using this fibre from our network. If any of these routes is part of a protected route, then its pair is also removed. We then attempt to re-route all these paths, firstly not allowing any new fibres to be added to the network; then, if some routes cannot be added, allowing fibres to be added. We shall either succeed in re-routeing all removed demands at a lower network cost than previously, or we fail and our new network is at a higher cost. In the failure case we revert to the previous network and proceed with the next fibre on our utilisation-ordered list. If we succeed then we commit the changes and recalculate our ordered list of fibres. We can stop either after a certain number of changes have taken place, or until we have exhausted our list of fibres; we attempted to remove each fibre in our network but every attempt failed. For the experiments below we used the exhaustive version of the algorithm.

We take the 20 scenarios used previously, and for each waveband size we place routes with a variety of routeing metric parameters. We then rearrange routes using the algorithm above, noting the fall in cost ratio as we successively rearrange more routes. All routes are placed without using any protection routeing for this experiment.

Figure 5.29 shows the best average cost ratio achieved before optimisation, the same as shown in Figure 5.11, and the best average cost ratio achieved post optimisation. Both curves are similar, but with a greater cost decrease for larger waveband sizes. This indicates that at the levels of traffic used our initial algorithm gives relatively higher costs at larger waveband sizes, and with optimisation a larger choice of waveband size could be warranted. This is a related result to that shown in Figure 5.26 where the

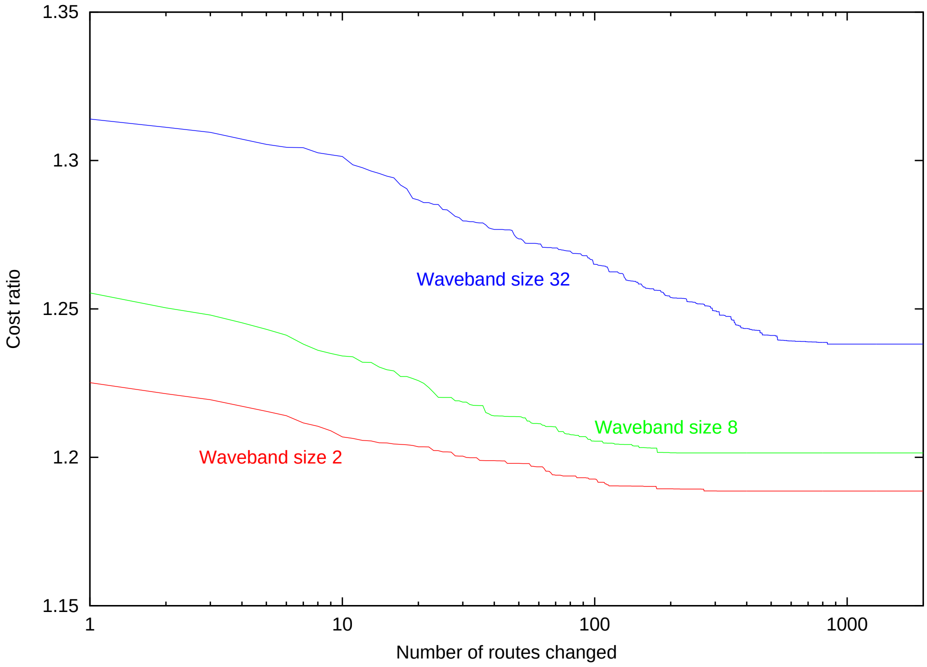

FIGURE 5.30: THE DECREASE IN COST RATIO GIVEN HOW MANY ROUTES ARE REARRANGED, FOR THREE WAVEBAND SIZES

optimal waveband size increases with traffic levels.

The choice of routeing metric parameters that give the lowest cost after optimisation are different from those previously identified in Section 5.1.1. The previous values were $FF = 1, RF = 0.8$, but to achieve the best optimisation results we need to initially perform almost shortest path routeing: $FF = 0.1, RF = 0.99$. Since this keeps the length of each route short, at the expense of fibre utilisation, it allows the rearranging algorithm the best chance to finish with untouched routes being near minimum length and only rearranged routes to be longer than necessary.

Figure 5.30 shows the gradual decrease in cost ratio as a function of the number of routes rearranged, for three different waveband sizes. After the first 10 routes rearranged, the cost decrease is approximately linear in the log of the number of routes rearranged. This is due to the algorithm design: it attempts to rearrange the fewest routes to gain the greatest cost decrease.

### 5.3.3 Dynamic Route Rearranging

From the results presented in Section 5.3.2 it is likely that there will be benefits from performing some form of route rearranging whilst routes are dynamically changing. We alter our rearranging algorithm from that previously presented, to now try to minimise the length of existing routes. We order all routes in the network in terms of the ratio between their current length and the minimum possible length. For routes with

| Traffic load, $\beta$ | Default | Remove longest (% of Default) | Periodic rearrange (% of Default) | Rearrange and remove longest (% of Default) |
|---|---|---|---|---|
| 17 | $9.46 \cdot 10^{-2}$ | 102.9% | 43.1% | 50.4% |
| 15 | $1.17 \cdot 10^{-2}$ | 78.2% | 70.9% | 70.2% |
| 13 | $1.41 \cdot 10^{-3}$ | 77.3% | 69.3% | 66.3% |
| 11 | $7.56 \cdot 10^{-5}$ | 75.9% | 65.3% | 71.6% |

TABLE 5.1: BLOCKING PROBABILITY FOR DIFFERENT DYNAMIC OPTIMISATION ALGORITHMS. FOR THE THREE ALGORITHMS ON TEST, BLOCKING PROBABILITY IS REPRESENTED AS A PERCENTAGE OF THE DEFAULT RESULT

protection routes we compare the total length of both primary and protection routes against their minimum disjoint routes, similar to the examples featured in Figures 4.10 and 5.2.

Considering each route in this order, we try to find an alternative route that is shorter. Since shorter routes are more efficient, using fewer network resources, this should help to decrease future blocking rates for new routes. Our current implementation allows the alternative route to use resources being used by the current route. This leads to optimistic results since, as discussed previously, this is unlikely to be the case for realistic network.

We have two parameters to control the operation of this algorithm. The first is a frequency parameter: how often do we run our rearranging algorithm. This is presented in terms of a number of new routes added to our network between each run, the *period cycle*; this was either 10 or 100 in the results presented below. The second parameter is how many routes to rearrange successfully during each algorithm run, and was set at 5, 20, or 50. If we exhausted our ordered list of routes before reaching this limit, we restarted the process and continued searching. Since some routes have changed, others may now be able to be made shorter. We terminate when we either reach our target number to move, or we have been through the whole list of routes without being able to rearrange any.

An alternative method of rearranging is also explored in this section, which attempts to gain the same benefits as the above algorithm but for no extra routeing cost. Since each pair of nodes has multiple routes between them, when we delete one of these we are free to delete any we choose. This does assume that the characteristics of each route are identical: one of our original assumptions from Section 4.2. In particular it assumes that if we wish to keep a route that has just fallen idle, we can switch traffic currently using another less efficient route onto this route before deleting the now idle route. In the following set of experiments we either have the default behaviour, which corresponds to a *first in, last out* (FILO) queueing system, or we remove the longest of all existing routes. Since the experiments presented here had shared protection paths, for this modification we removed the pair of routes with the longest combined length.

Table 5.1 shows the lowest blocking probability achieved by the four different algo-

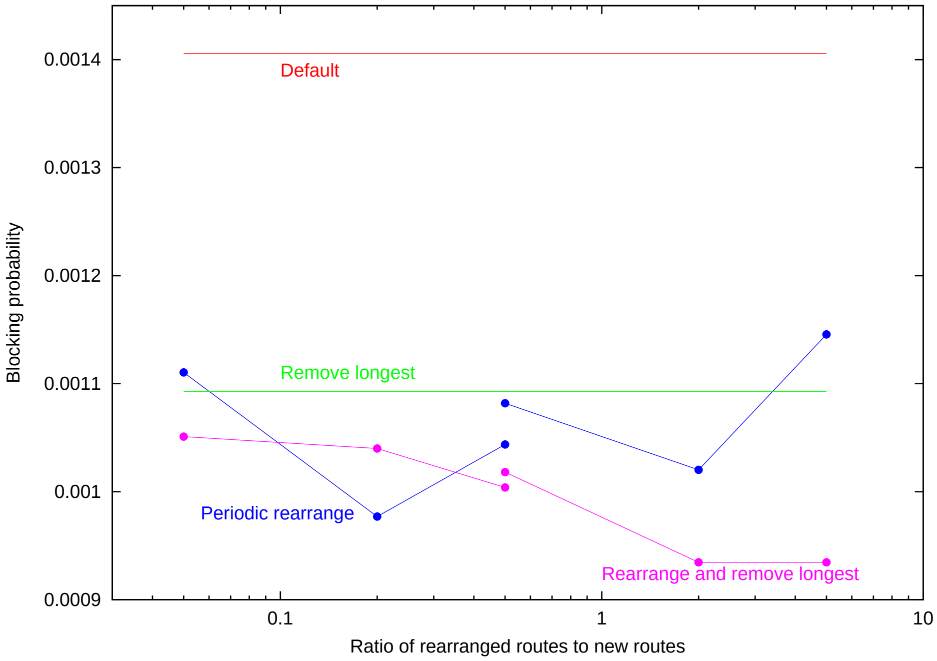

FIGURE 5.31: BLOCKING PROBABILITY FOR $\beta = 13$, BY HOW MANY ROUTES ARE REARRANGED

rithms. We compare our original algorithm with that using the 'remove longest' option to delete the longest path instead of the default FILO queueing system, and then both of these alternatives using a periodic rearranging of existing paths. For this table the best results were chosen from the various combinations of how often to perform this and how many routes to alter. We can see that in low loss situations all three algorithms improve on the default algorithm. If we are performing periodic rearranging then removing the longest seems to give only a very small improvement; however, most of the gain from performing periodic rearranging can be achieved just by removing the longest existing route. This gives around a 20%–25% decrease in blocking probability, compared with a decrease of 30%–35% for the best periodic rearranging algorithm results.[12]

Figure 5.31 shows the performance of the different algorithms in detail for $\beta = 13$. Firstly the default algorithm is shown, as well as the algorithm that just removes the longest existing path. For the two algorithms which perform periodic rearranging, the x-axis shows the ratio between the number of routes rearranged against the number of new routes added. The break in both curves indicates where the *period cycle*, the interval between each rearranging, changes from every 10 new routes added to 100.

For the periodic rearranging algorithm, a smaller period cycle performs less well, even though more routes are being rearranged. This is probably due to the algorithm always starting at the longest route first and working downward; it is possible that some

[12] Note that experiments are run to achieve at best 0.2 relative error on a blocking probability of $10^{-4}$. Reported blocking probabilities smaller than $10^{-4}$ will have a larger error factor.

routes are never checked for a more efficient solution. Looking at the periodic rearranging with removing longest algorithm, as we rearrange more routes, our solution improves as expected. To achieve the best result we are rearranging up to five times the number of new routes added to the network. A more conservative approach, rearranging less than a tenth of the number of new routes added, only slightly improves on the algorithm that does not perform any rearranging.

The cost of algorithms that perform periodic rearranging, apart from the potential switch-over loss mentioned above, is the extra control traffic needed to find candidates for rearranging and to implement the changeover. To be able to perform periodic rearranging would depend on having sufficient capacity in the control fabric, over and above that needed for reacting to traffic shifts and other external stimuli that lead to routeing changes. In addition to the capacity of the control fabric, in terms of an average number of route changes per time, the latency of any one route change would need to be considered. We would have to balance the possible delay to a 'real' routeing request caused by a network optimisation algorithm against the slight decrease in blocking probability for new routeing requests. Alternatively we can decrease the blocking probability for new routes by provisioning a larger network, so these algorithms could trade off the cost of the network with the observed latency for new routes.

## 5.4 Conclusion

In this chapter we reported on results gained from the network simulator that we have developed. Firstly we reported on results gained from experiments on the fixed demand problem from Section 5.1.

- We assessed the routeing algorithm proposed in Section 4.2.5, and found that we can find parameters that lead to good results. These values seem to indicate that the method of using weighted lengths as a routeing metric is better than the extremes of either using shortest path or maximum capacity efficiency routeing.
- We tested our routeing algorithm using protection routeing, where we have to find both primary and protection routes simultaneously. We see that the benefits of shared protection over dedicated protection depend to a great extent on the size and intra-connectivity of the network, and that for small sparsely connected networks the cost reduction from using shared protection is relatively small.
- We experimented with the ordering in which we place routes whilst provisioning a network, and see that in general placing the longest routes first method is best. The exception is when using shared protection where the existence of long waveband paths, our granularity for sharing, inhibits gains from sharing leading to a random ordering outperforming longest-first.
- We also tested the size of a waveband, seeing that for the network topologies and traffic levels tested, a choice of 4 wavelengths per waveband is generally a good

choice.

A methodology for conducting experiments using a changing demand routed across a fixed network architecture was explained, and the different architectures outlined in Section 4.2.3 were tested fairly. We now report on the results found when investigating the fixed topology problem from Section 5.2.

- When the network is exposed to the same demand as that used for provisioning best performance is achieved when the network is exposed to the same demand as that used for provisioning a minimal infrastructure using no switches at all.
- With dynamic demand shifting away from that used to provision the network, a partial flexibility architecture exploiting waveband routeing is best.
- This advantage is maintained over a range of relative equipment costs, as shown by the cost sensitivity analysis performed. The advantage of the waveband routed architectures are maintained until a relative decrease of around 75% in the cost of optical switches or transmitters and receivers, or a 50% relative increase in the cost of fibre and amplifiers.
- We also looked at using tuneable components to further reduce equipment costs, but find that it requires a cost increase per component smaller than that currently available to match performance using fixed frequency components.

We then looked at two methods to improve performance in Section 5.3.

- We used a two-stage provisioning approach, where full wavebands are routed first. This reduces costs by a large factor, but leads to a network with lower flexibility. However, the trade-off does reduce blocking probability over the previously used method for provisioning networks.
- Secondly we looked at ways to ensure that routes in our network remain efficient, firstly by an intelligent removal strategy, and secondly by periodically rearranging existing routes. Both approaches reduce the blocking probability, but a large proportion of the possible gains are made by the intelligent removal strategy alone, thus obviating the need for a potentially costly periodic rearranging algorithm.

The results presented in this chapter indicate that the architecture and routeing algorithms presented in Chapter 4 can be used to operate a dynamically reconfiguring optical network, which creates and removes wavelength routes between edge routers. The work presented in Chapter 3 shows that for a single pair of edge routers this ability to change the total available bandwidth can lead to an increase in utilisation by tracking large scale traffic movements. In the next chapter we look at how these ideas can be combined to form an optical network, and how this layer will interact with higher network layers. We look at areas such as how to achieve fairness in bandwidth allocation, how the operation of the optical network can affect the observed network performance, and the scaling parameters of this network.

# Chapter 6

# Implementation Issues

In the previous chapter we looked at the performance of an optical layer which reconfigures the set of virtual connections in response to traffic-led changes. We have seen that with the correct architecture and routeing protocols an optical network can reconfigure, adding and removing routes over time, to best match the current traffic matrix. This chapter looks in more detail at the implementation of the optical network, especially in terms of how it relates to, and interacts with, the higher network layers.

## 6.1 Optical Layer Operation

We reproduce the Figure 1.2 here for convenience, as Figure 6.1.

At each IPON we need to perform several tasks. Incoming packets need to be routed through the network; the OXC-IPON egress node needs to be identified for each packet. This is a standard IP routeing table look-up. The OXC will present a set of virtual interfaces to the IPON, one for each possible network egress. These interfaces will accept packets, and be responsible for choosing the correct wavelength route for transmission across the network; each virtual interface will have a set of wavelength routes which go directly to the corresponding egress point.

### 6.1.1 Failure Recovery

To maintain full-mesh virtual connectivity, the optical network needs actively to respond to a failure. If the failed wavelength route has an associated protection route then traffic can be automatically switched over, waiting for the necessary signalling time if the protection route is shared.

The route may not have a protection route, either through deliberate choice or since the

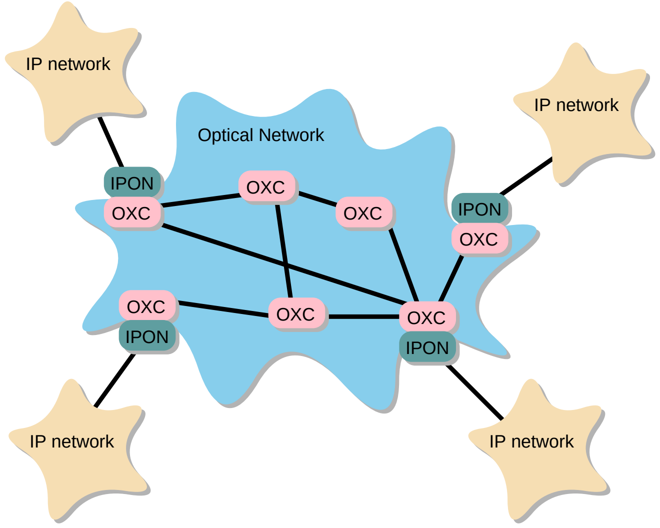

FIGURE 6.1: OPTICAL SUBNET ARCHITECTURE

network has undergone multiple failures affecting both primary and protection routes. In this case, if there are multiple routes between the affected pair of edge routers, then some other primary paths may be unaffected by the failure. In this case then all traffic can be multiplexed onto the remaining routes. This may be a deliberate provisioning strategy, where we create multiple routes which are either disjoint or mostly disjoint to maximise resilience. In normal operation traffic can be split between these routes, and a low utilisation will be maintained. In the event of failure of one or more routes, traffic will be multiplexed onto the remaining routes.

If all routes between a pair of edge routers are affected, then if we are using some form of Time Division Multiplexing, covered in more detail in Section 6.3.2, we may be able to use a time-slot on existing routes which will go via a different edge node. This could be a prearranged strategy, although we would have to choose two routes to use which are disjoint to the route to be protected, where either of those routes may be deleted over time. The alternative would be for this to happen at the time of failure, which would lead to a longer delay before connectivity is restored.

If this is not feasible we could add a new route across the network, where the delay would now include the time to find the route and the possible latency for bringing up a new wavelength route as previously discussed in Section 2.3.2. If these are not possible, or the delay in adding a new route is too long, then the higher-level recovery mechanisms would be required, such as either using intra-AS routeing to select an intermediate edge router at the IP level, or inter-AS routeing to select an alternative AS path for the affected Internet routes.

The trade-off between reserved bandwidth and recovery time is likely to depend on the use of the route. If the route is part of a high-cost high-reliability VPN then the service-level agreement may require the use of dedicated protection. If the route is used to support peering arrangements with other ASs for a best-effort IP network, then we may decide to rely on multiple disjoint primary routes between each IPON where, in the event of failure of one route, traffic is multiplexed onto any remaining routes. The degree of protection may also depend on the utilisation of the optical network. Since networks are typically re-provisioned on a cycle of 6 months or longer, immediately after a re-provisioning, network capacity will be larger than traffic demand. Thus we may be able to support dedicated protection for all routes while maintaining a low blocking probability for new routes. In the period before a re-provisioning we may have to operate with less protection overhead to ensure normal operating activity. One advantage of using route-based protection is that the degree of protection can easily be varied between different routes and over time.

There are reasons why we may choose not to maintain a full mesh optical topology. Firstly failures may occur in the IP level equipment which are not visible in the optical layer: failure of routers for example. These failures are perhaps more naturally dealt with at the IP level. If we are promoting an integrated layering strategy, motivated by the discussion in Section 6.3.1, then the optical protection strategy could include failure of any optical or IP equipment. The signal to activate the protection mechanism would therefore come either from optical signal monitoring equipment or IP connectivity monitoring depending on the cause of the failure.

Secondly the cost of any protection or restoration scheme has to be considered. Over-provisioning the available bandwidth of an IP network and relying on a fast IP re-route operation in the event of failure may be cheaper than over-provisioning reserved optical bandwidth. As stated above the method of generating revenue would also have to be taken into account, if the difference in recovery speed is critical.

### 6.1.2 Effects of Local Optimisation

In this dissertation we have taken the decision to add new routes without altering existing routes; we perform a local optimisation to place each route. Over time this may lead to a situation where the optical network is in a state which is globally sub-optimal, which can only be substantially improved by a reorganisation procedure which would be disruptive to existing routes.

Following the analysis in Section 5.3.3 our procedure for avoiding this state is that when a route can be removed due to decreases in traffic levels the most inefficient route is removed. This was shown to give most of the benefits associated with a larger scale periodic rearranging function.

This procedure is least effective for avoiding a globally sub-optimal configuration when a low number of routes being added and deleted, coupled with high demand. In this situation it may be necessary to have the option to suspend the automated reconfiguration and to force the global rearrangement of existing routes. In the event that there are multiple waveband routes connecting each pair of edge routers, it may be possible to achieve this global reconfiguration without losing the connectivity of the network by temporarily moving all flows onto a single route.

### 6.1.3 Route Creation Competition

Since we have many pairs of IPONs, there is competition for resources in the optical network. In the experiments reported on in Chapter 5, requests for new routes were dealt with sequentially and in a first-come-first-served manner. The outcome of this type of operation is that pairs of IPONs that are further apart have a higher blocking probability: since they require more resources to create a route there is more chance that any one required resource will not be available. This is a similar result to that found in measurement-based admission control algorithms [Jamin97].

We would typically want the blocking probability to be independent of the distance between the two nodes. We may want the blocking probability to depend on the type of network the routes are supporting — for example a high cost VPN or a general IP network. For routes supporting an IP network we may want to bias the choice of routes to support a fair allocation for that network, especially in the face of elastic traffic as discussed in Section 6.2.5.

A solution to this priority problem would to have a reservation scheme, where if a routeing attempt fails we can reserve a route instead. Reservation stickers would have a priority ordering, created using the criteria outlined above. Free resources on the desired route would be reserved automatically, and resources currently in use marked with a reservation sticker. If this scheme should be pre-emptive then when the sticker is placed, the current user of those resources is notified and they may delete that route. The priority ordering could determine the result of contention where multiple sources wish to reserve a resource.[13]

When it comes to delete a route due to falling traffic, an IPON pair would delete the route with the highest priority reservation stickers attached. This is similar to the scheme tested in Section 5.3.3 where the longest route is deleted; longer routes use more resources and are more likely to be wanted for new routes. When resources with a sticker on are freed, ownership transfers to the owner of that sticker, and they are notified that they own that resource.

A common problem for multiple algorithms running in a distributed system is where they interact to synchronise or beat against each other. With the scheme above there is obvious scope for some degree of auto-synchronisation, since the removal of a route may cause notifications to be sent resulting in several new routes being created. The only way more routeing changes would be triggered by these new routes would be if some lower priority routes needed to be deleted to create these new routes. Clearly the

[13] Using strict priority as a scheduling mechanism can lead to problems such as priority inversion and resource starvation. Well-known solutions to this problem exist such as dynamic priority, however it might be that strict priority between classes of stickers is needed to enforce management led decisions

priority scheme would need some thought to avoid these problems, either in designing the ordering properties of the scheme or removing the possibility for pre-emptive resource requests.

## 6.2 Network Traffic Interaction

For each pair of IPONs, the optical layer will operate some number of wavelength routes across the optical network. It will seek to have these routes use the minimum necessary resources, subject to the management-level traffic agreements. The virtual interface will monitor the quantity of traffic passing between these two nodes.

This operation of the optical layer will affect higher network layers, since the properties of the optical network may change over time. In this section we look at these interactions and ways to minimise these effects.

### 6.2.1 Flow Identification

Considering the behaviour of the optical network from a higher network layer perspective, we would wish the observable performance of the optical network to be consistent over time. This allows elastic traffic protocols such as TCP to correctly identify the correct window size, which is based on available bandwidth and Round Trip Time (RTT), and inelastic traffic such as real time media streaming to correctly set the connection rate and application level performance. Therefore the traffic originating from a source address and travelling to a destination address should have a similar performance. If the virtual interface used has multiple routes with different lengths and therefore differing propagation times, then we would wish all associated packets to use the same wavelength route to give a stable RTT. When this is not achieved, and there is a significant different between propagation delays, then packet reordering may occur, which increases the burden on receiving end systems and may lead to needless packet retransmission. This is discussed further in Section 6.2.3.

At what level packets are associated together is a critical decision, since we have to store state to match packet flows with wavelength routes inside the virtual interface and perform a form of pattern matching on each data packet. The finer grained the association, TCP flows identified by source and destination IP and port numbers for example, the more state we store and the longer the look-up procedure. However, this allows us to perform better load balancing between different wavelength routes; this is covered in Section 6.2.2. Other options for possible matches would include matching just on IP source and destination, matching on IP destination or source only, or matching using network prefix or AS information. We refer to the data packets matching in our chosen stream as a *flow*. All the information needed to perform this matching in the schemes listed above is already available through the standard IP routeing look-up that has to be done to determine the correct virtual interface to use, so it could be passed to the virtual interface with the packet to assist in flow identification.

The aggregation level in the network would determine the practical minimum size of a flow, depending on the capability of the hardware which performs matching and accounting. It is likely that for core networks performing flow matching on destination AS or both destination and source AS would be realistic.

### 6.2.2 Flow Balancing

Whilst the total traffic moving from one IPON to another may be stable over some time period, the distribution of traffic across the series of flows may change. The definition of a flow is flexible, but should allow us the capability of load balancing across different wavelength routes travelling between the same pair of IPONs. This allows us to allocate the total bandwidth available across this set of wavelength routes in some fair manner.

The coarser grained our flow definition the less state and processing time required to load balance, but since we move whole flows across is becomes harder to achieve the optimal balance. Given that a flow is an aggregation of network traffic, it will be subject to similar statistical properties as seen in Chapter 3, so appropriate analysis would have to be performed to identify when to move a single flow from one wavelength route to another.

On a longer timescale the virtual interface may observe a total traffic change large enough to warrant adding or removing a wavelength route, using the algorithm developed in Section 3.6.2. In the case of adding a new route, the existing flows would have to be spread out, with some flows from each existing route being moved to the new route. When deleting a route, the reverse procedure would be used, with the flows on the selected route being moved to other routes. Therefore, for the higher network layers, all actions by the optical layer can be considered as load balancing between different wavelength routes. Note that following the results in Section 5.3.3 routes to be deleted are likely to be those using the most optical resources, rather than those requiring the least traffic to be moved onto other routes.

### 6.2.3 Consequence of Flow Balancing

The consequence of moving a flow is that for the application level traffic comprising that flow, the network would appear to change RTT and loss behaviour. If the flow is moving to a shorter route with a corresponding drop in RTT some packets may be reordered, as the packets already in flight on the old route will be overtaken by those on the new route. The quantity of reordering can be calculated by looking at the maximum number of packets in flight at any one time and the different in RTT caused by changing routes.

For example a North American continental network has a direct coast-to-coast distance of 4000 km, whereas perhaps the longest route might take twice this distance, 8000 km. Since the speed of light in a fibre is approximately 200,000 km/s the difference in RTT is 20ms. With a single TCP connection transmitting 1 Mbit/s would have around 90 1500 byte packets transmitted per second, resulting in a packet every 10ms. Thus an instantaneous change in RTT of 20ms would result in each TCP connection on the link having one or two packets reordered. If this happens relatively infrequently it is unlikely to be a problem. The current way that TCP transmits actually sends bursts of packets when an acknowledgement is sent which increases the available window, with a relatively longer period between bursts. Therefore connections would only be affected by a 20ms change in RTT if this occurred while a collection of packets inside a single burst was being transmitted.

It is envisaged that the IP layer routeing protocols are not aware of the change in RTT caused by the optical layer reconfiguring, since the same connectivity and cost information will be propagated into the EGP used to route between neighbouring ASs.

### 6.2.4 Fairness of Flow Balancing

The benefit of flow balancing between wavelength routes is that all flows would receive fair treatment from the network. The major two network attributes affected are the propagation delay and the loss characteristics.

The difference in propagation delay could be an order of two; for example moving from 20 ms to 40 ms delay. This would depend on the maximum path difference between the available wavelength routes; this may be constrained by only using a fixed set of alternative routes rather than full unconstrained routeing, or by the use of Service Level Agreements which constrain the permissible cross-network latency.

The loss characteristics will depend on the possibility of buffer overflow at the virtual interface into the optical layer. We are likely to have separate packet buffers for each wavelength route, since they would be associated with each optical transmitter, so loss will depend on the effective bandwidth of the flows selected for a given route. Flows are likely to exhibit self-similarity, and as seen in Table 3.1 are likely to be statistically different from the aggregate total traffic. Some flows may be very bursty, while others have a more Gaussian-like marginal distribution, probably depending on the bandwidth limits elsewhere in the network and the degree of aggregation present in each flow.

If all flows have similar characteristics and are just different in magnitude, then load balancing to achieve similar loss characteristics in all routes is simple since average bandwidth allocation will be fair. Packet loss is caused by the high peaks of a bursty traffic flow: if there are differences then we might wish to consider two schemes.

Firstly we allocate flows in matched sets to routes; for example with $N$ routes we pick the $N$ burstiest flows and allocate one to each route. We then pick the next $N$ bursty flows, and so on, taking care that the sum of the average bandwidth of the set of flows in each route is equal.

Secondly we may decide that since the bursty flows cause the majority of the packet loss that we should attempt to use statistical multiplexing properties and place the burstiest flows together in the same route. The most stable flows would also share the same route; it is likely that the sum of the average bandwidth for these stable flows would be higher to achieve a similar loss ratio, so this may lead to a more efficient packing of flows.

### 6.2.5 Elastic Traffic Adaptation

If the traffic levels across the network are high, then the elastic proportion of the traffic will adapt to fill the available bandwidth; this section of the network may be the bottleneck link. If this prompts the optical network to increase the available bandwidth, then the elastic traffic may just fill the extra. This leads to two issues on different scales.

Firstly when considering the fairness of flow balancing, as discussed in Section 6.2.4, if all wavelength routes linking a pair of edge connected routers have high utilisation then the affects of elastic traffic need to be considered. This is since the measured aggregate statistics may not be enough to share the available bandwidth fairly. The concept of fairness among many elastic traffic sources is contentious. Max-min fairness says that no connection should increase its share if that leads to another decreasing its share [Bertsekas87], whereas 'a system is proportionally fair if any change in the distribution of the rates would result in the sum of the proportional changes being negative' [Crowcroft98]. Proportional fairness has been investigated as an outcome of a possible pricing strategy where people pay some amount per unit time for their network bandwidth [Kelly97]. Current TCP implementations share bandwidth such that the average bandwidth is proportional to $1/T^a$ where $1 \leq a < 2$, and $T$ is approximately the RTT [Lakshman97].

Without some form of pricing system which could then be used to correctly weight connections, it is clear that achieving any kind of fairness is hard. Since we are operating with large aggregations of connections some simplified scheme may be appropriate; maybe keeping the number of TCP connections on each route equal. To save administrative overheads it may be feasible to have a mechanism that just spots TCP SYN or FIN packets to approximate the number of connections, rather than some form of accurate TCP analysis model. The experience gained in Chapter 3 would suggest that FIN packets would be more accurate, since SYN packets may be retransmitted when connections are refused due to an end-system admission control policy. The fairness that this scheme would achieve depends on the degree to which the TCP connections in different flows share similar sets of properties; the distribution of 'mice' and 'elephants' would need to be similar for all flows.

Secondly when the traffic load across the whole network is large, there will be competition between different pairs of IPONs for the available optical bandwidth in the network. To keep the allocation fair we need some way of comparing the relative benefits from the requested wavelength route against the existing routes in the network. However this would break one of the assumptions behind the routeing algorithm, that new routes are added without changing existing routes. One solution to this was discussed in Section 6.1.3.

Another consideration is that the action of elastic transport protocols such as TCP should not interact badly with the altering of total bandwidth available. It has been observed in the past that loosely-coupled systems can synchronise to affect the large scale behaviour, both in routeing protocols [Floyd94] and TCP aggregate behaviour through a bottleneck link [Zhang90]. Since the likely timescale of bandwidth change is over a much longer timescale than the adaptation mechanisms for short-lived flows, this is not expected to be a problem.

The problems mentioned in this section are mainly prevalent under high load. Current backbone networks are typically provisioned so that high load is rare, and generally only occurs when some part of the network fails, leaving the rest to carry the offered load.

## 6.3 Network Scale

When designing a network architecture, some notion of how that architecture scales with size is essential. Scaling with traffic levels, the total number of nodes, and the interconnectedness of nodes are all important.

### 6.3.1 Routeing Adjacencies

The ability to scale easily with network size is a persuasive argument for packet switched networks since each node only has routeing adjacencies with directly connected nodes. As the size of the routeing domain increases, each node still has a similar number of physical adjacencies, and each packet travels through more hops.

With a circuit switched approach, the work is performed at the edge of the network. If the optical network achieves stable full-mesh connectivity then whilst each edge-router has a large number of routeing adjacencies, each of these virtual connections apparently never changes so no internal routeing updates are necessary. The length of the routeing table is presumed not to change between the packet switched and circuit switched networks since this reflects the size of the Internet, but this node has more outgoing connections, or virtual interfaces to the optical network, from which to choose.

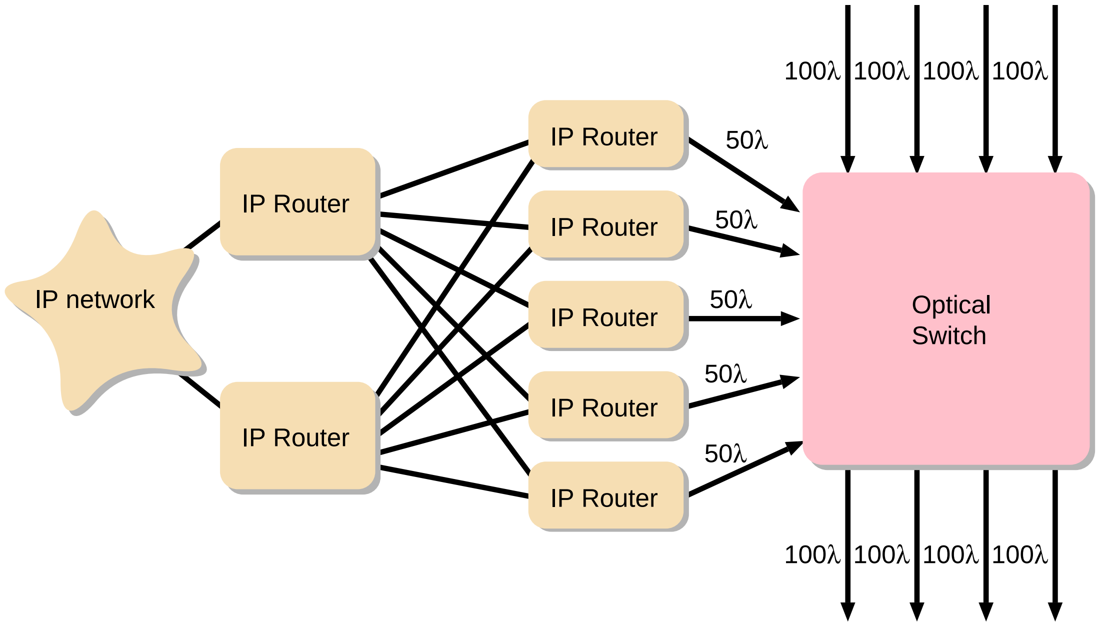

FIGURE 6.2: MULTI-STAGE ROUTER

Informal understanding of both currently deployed routeing protocols, such as OSPF and ISIS, and implementations of IP routers suggest that high degrees of connectivity cause problems [McAuley03]. Since major network nodes typically are connected to between 2 and 10 adjacent nodes [Iannaccone03], there has been no need to design protocols or engineer routers for higher than a few tens of connections. However, since the operating assumptions for an internal routeing protocol based on a full mesh topology are so different than a typical partial mesh IP network, there is no reason why a new or improved routeing protocols should not cope. New router implementations may be more challenging than existing routers, causing the cost benefit of this approach to decrease. If fundamental limits on the connectivity of an IP router do exist, then one solution would be to a have a more complex hierarchy of routers at a network node, where traffic is spread out towards the correct fibres. By classifying traffic as flows as discussed previously, traffic destined for a given fibre could be routed through this set of routers by computing the relevant routeing table entries, enabling several routers to emulate a single router with a high degree of connectedness.

An example of this can be seen in Figure 6.2. Here an access network is connected to two first-stage routers, used for redundancy. These in turn connect to multiple second-stage routers, which each handle the proportion of the traffic bound for a set of wavelengths. These all enter a wavelength switch, or the correct combination of wavelength and waveband switches depending on the architecture. This removes the need for a single router to physically cope with a large number of output linecards; the operation of the second-stage routers is also comparatively simpler than the first-stage routers. This could be part of the process to deployment, since all individual units are similar to those current available. Over time it is expected that these would partly or entirely coalesce.

AS-level routeing updates are still processed, changing the path across this network for an external destination. To respond to this change there will be a single update at each edge router to divert packets with the updated destination.

In the event that a network failure causes an unrecoverable failure in the full mesh connectivity as discussed in Section 6.1.1, external routeing protocols will be compelled to find an alternative AS path across the network.

### 6.3.2 Allocation Granularity

If we wanted to create a full-mesh virtual topology with disjoint routes between each IPON pair to create resilience in the event of a single fibre, then if we have 50 nodes then we would need at least 100 wavelengths leaving each node. This is perhaps equivalent to 2 fibres, with a peak bandwidth capacity of 1 Tbit/s at each node, assuming 10 Gbit/s per wavelength.

If the traffic levels in the network do not scale with the size of the network, we may not have sufficient wavelengths available to create our resilient full-mesh connectivity. In this case then we might require Time Division Multiplexing (TDM). The simplest integration with the network architecture described in this dissertation would be treat each (wavelength,time slot) pair at the same level as a wavelength. Thus we create a wavelength route by allocating a (fibre,wavelength,time slot) tuple on each link.

Figures 3.17 and 3.18 show that a smaller granularity of allocation, relative to traffic levels, can lead to higher utilisation with more frequent changes of allocation. With the addition of TDM, careful engineering is required to ensure that the bandwidth in a single (wavelength,time slot) balances off the average utilisation and frequency of change. It may be that dividing the wavelength into only a few timeslots would be sufficient to allow resilient full-mesh connectivity.

### 6.3.3 Concurrent Reconfigurations

As the number of IPON-pairs increases, if each wants to add or delete routes with the same frequency, the total frequency of network reconfigurations will increase. If the system is distributed through the network, then we may have the scope for some of these reconfigurations to occur concurrently. This would probably only be possible when the two new routes use different sets of optical resources, so that they do not interact when being added to the network.

As an upper bound on the limit to the scale of the network we first consider the centralised case, where all routeing changes are handled by a single network controller. In the worst case a single controller would serialise routeing changes, although a simple pipe-lined approach would be possible since as we configure the switches to add a new route we are searching our topology database for the paths taken by the next route to add.

With $N$ nodes in our network, we have $N \cdot (N-1)$ pairs of IPONs, if we consider unidirectional demands, or $N \cdot (N-1)/2$ pairs using bidirectional demands. We use $s$ as the allocation delay, the time taken to add a new route to the network. We assume that this is the limiting factor to the speed of adding new routes, as opposed to the initial route determination. Thus on average we can change the allocation for any IPON pair every $t = s \cdot N \cdot (N-1)$ seconds; for $N = 50$ and $s = 2$, $t$ is 1.4 hours. This is the operating region identified in Section 3.6.2; a higher allocation delay would result in too few changes, and would need a better predictive algorithm than that included in this work. Considering bidirectional demands and an allocation delay of one second then on average we are able to reconfigure every 20 minutes. As $N$ increases, the pressure to reduce $s$ further will increase.

Considering traffic streams between pairs of IPONs, it is likely that they will be different sizes in relation to the allocation granularity of a wavelength, and that some streams will change bandwidth levels more frequently than others. Thus, with an example network average of a reconfiguration every hour, some traffic streams will require reconfiguring more than one an hour, whilst some only a few times a day. It may be that sufficiently few streams require actively reconfiguring to reduce the burden on the optical control-plane.

This analysis considers a centralised approach. For a distributed approach the degree of simultaneous reconfiguration is important. As the diameter of the network increases the number of pairs of edge routers increases. However, especially if the physical interconnectedness also increases, there is more scope of concurrent reconfigurations since routes overlap less; this is similar reasoning to that behind the cost advantage increase in shared protection over dedicated protection as shown in Figure 5.9.

## 6.4 Conclusion

In this chapter we looked at some of the aspects of optical network design that affect the higher network layers. In Section 6.1 we looked how the optical network may decide to deal with failure, and how these decisions may depend on the traffic being carried over the network. We also look at how the network may escape from a local optimum configuration and how fairness can be introduced when there is competition for optical resources.

In Section 6.2 we looked at the explicit interaction between the optical layer and the IP layer. We examined the need for flows, to keep related packets using the same wavelength route, and discussed the granularity of these flows, the consequences of moving flows between routes, how to achieve fairness between flow allocations, and the problems caused by elastic traffic adaptation.

Section 6.3 contains a qualitative analysis of how the operation of an optical network scales with size. We looked at the IP-level routeing adjacencies caused by a full virtual mesh topology, whether we will require TDM in addition to WDM, and the scope and impact of concurrent reconfigurations.

While these issues need to be considered carefully, and some may have a large effect on the design of the optical network, none seem insurmountable.

# Chapter 7

# Conclusion

In this chapter we summarise the work presented in this dissertation, and list the contribution. We then outline future work that would extend or compliment this dissertation, before finally concluding.

## 7.1 Summary

We first introduced our work, and gave an overview of the essential background material to this work. The introduction focused on the current networking architecture, how to make simulating networks practical and relevant, the properties of optical components, and reviewed past work on creating virtual connection schemes.

Chapter 3 discussed a 48 hour period of measurement data at the connection point of a large university to the Internet. Separating the traffic into groups based on the measured AS topology, we analysed the traffic for a Gaussian Marginal Distribution (GMD). We found that while the separated traffic distribution varied greatly in approximating a GMD, most likely caused by bandwidth limits elsewhere in the network, the aggregated traffic was very close to having a GMD. This analysis has impact on the core of the network, where traffic is aggregated from many heterogeneous access networks. Using the assumption of a GMD we measured the self-similarity of the traffic and, in common with other studies, found that the measured traffic is approximately self-similar. These findings give us more understanding of the likely behaviour of large traffic aggregates in future networks.

We next introduced two algorithms to perform bandwidth allocation based on a set of traffic measurements. Analysis of the off-line algorithm shows the trade-off between the frequency of bandwidth change, and the average allocated demand: to decrease the necessary bandwidth — which increases the average utilisation — the bandwidth needs to change more frequently. The impact of changing the granularity of allocation was also measured: as the granularity decreases we can achieve better utilisation at the cost of changing more frequently. The limitations of the assumption about applying this work to elastic traffic were noted, and a possible scheme with a target utilisation was introduced. The on-line algorithm takes into account the lack of future knowledge and the latency of allocation caused by optical device limitations. Where latency is high, we identified a high utilisation operating region where we change allocation approximately every hour. Finally we discussed the impact of multiplexing many traffic streams together and the impact this would have on the analysis performed.

This chapter concluded that it is likely that, for a single aggregate stream of traffic, the utilisation can be increased markedly by changing the bandwidth allocation. This has little impact on the observed data loss, and can be achieved by a simple on-line algorithm.

Chapter 4 introduced the concept of waveband routeing. Here we are able to treat a waveband — a set of wavelengths — as a single routeable object. We outlined possible optical switch architectures suitable for dynamically rearranging the set of wavelength routes across the network. These take advantage of waveband routeing to reduce the equipment cost for the same throughput. These architectures trade off the equipment cost reduction against the decrease in routeing flexibility. We considered three main architectures. The first, the FULL architecture, allows full flexibility with high cost. The PATH architecture uses a fixed set of waveband paths and has optical transmitter and receivers only at the end of each path. Finally the FIXED architecture has a fixed set of wavelength routes; there is no flexibility to reconfigure, and we rely on over-provisioning for performance.

We described a routeing algorithm which is able to add routes to a network serially by using Dijkstra's shortest path algorithm with modified weights. It can route protection routes, including shared protection paths, and be used to provision networks as well as add routes to a dynamically reconfiguring network. We also described the configuration of the routeing simulator used in this dissertation. The generation of the topology model, traffic matrix, and traffic data are described, all of which are plausible models for future single AS networks. We gave a simple cost model, which separates the cost of an optical network into four categories. The ratio of costs for each category is designed to model the cost of a long-haul backbone network.

This chapter illustrated that it is possible to perform simulation studies on circuit switched optical networks, by giving the node architectures, routeing algorithm, topology of the network, traffic demands, and a method for costing networks to ensure a fair comparison.

Chapter 5 reported on the results from our network routeing simulator. Firstly we analysed the routeing algorithm previously presented, and found that for a set of scenarios we can find good values for our internal routeing parameters. We presented experiments to find how the benefits from sharing protection change with the scale and connectedness of the network, and to find the best size of a waveband.

We compared the different network architectures given in the previous chapter with a range of traffic matrices, ranging from the matrix used to provision the network to an uncorrelated traffic matrix. We took care to equalise the cost of each network, independent of the architecture used, to ensure a fair test of performance. We found that with the matrix used to provision the network the FIXED architecture outperformed the others, but when the proportion of the uncorrelated traffic matrix increased, the PATH architecture gave the lowest blocking probability for the same traffic demand. We then tested how these results are affected by changing the cost model, and find that if we considered two variants of the PATH architecture, for a change of 50% in any single cost component, either the PATH-1 or PATH-2 architecture matched or outperformed the FULL architecture.

We finally looked at network optimisations, and found that a two-stage provisioning process may improve performance. Benefits were also gained from a scheme which removes the route using the most resources, rather than using a strict call model to add and remove routes.

This chapter concluded that the routeing algorithm is feasible, and can route changing traffic loads over the node architectures proposed previously. In most realistic traffic conditions a waveband-based architecture is best, and this holds for a range of relative equipment costs.

Chapter 6 considered the implementation of a reconfiguring optical layer, presenting a qualitative analysis of the interaction between a reconfiguring optical layer and higher network layers. We first considered the operation of the optical layer, with possible schemes for recovery in the event of failures and the possibility of a dynamic network being stuck in a local optimisation. We also considered how to correctly resolve competition for resources.

We looked at how the network can appear both consistent and fair to network traffic, especially where traffic is elastic and adapts to the different bandwidth allocations a stream may be given. Finally problems of network scale were addressed. Where a network is fully connected the edge routers have large adjacencies, which is a problem for current routers and protocols. Where traffic is small compared to the size of the network we may use Time Division Multiplexing, in addition to Wavelength Division Multiplexing. Lastly as the number of nodes increases we may be faced with a large number of reconfigurations.

This chapter contained some of the issues faced in the interaction between the optical and IP layers. While some are serious and would have a large effect on the design of the optical network, none seem insurmountable.

## 7.2 Contribution

The concepts of virtual connections and how they might ease the burden on IP networks are not new. Such schemes usually require a fully-functional network layer to cope with the rich, but not fully meshed, virtual topology. This work has started to explore some of the benefits and drawbacks of maintaining a full mesh topology at the optical layer, which may allow a large simplification of the routeing protocol at the network layer.

Measurements from many traffic sites have been analysed for self-similarity in a similar way to the techniques used in this dissertation. That more aggregation, both on the time and traffic scales, leads to a more Gaussian Marginal Distribution (GMD) is known. Analysis of traffic by splitting traffic up into disjoint streams based on external routeing knowledge has not been carried out before: this gives insight into how streams with different statistical properties can multiplex together to form an aggregate stream with a GMD.

The only published bandwidth allocation algorithm which is comparable to that given in this work uses only history based on the mean traffic over the last fixed length period, and is based on high and low watermarks, rather than any target for user perceived performance such as loss. That approach also discards the detailed burst timescale information critical for determining performance. The work presented in this dissertation uses a moving average which targets real network performance, and contains a full analysis of the effects of changing both the averaging parameter and the latency of allocation. We also find that with the measured data there is no obvious timescale at which networks should reconfigure, and that reconfiguring over shorter timescales can increase bandwidth utilisation. However we also find that when the latency of setting up optical routes is high, the most efficient operating region changes allocation on average once each hour.

Wavebands have been explored in previous work, but no algorithms have been developed to use them in reconfiguring optical networks. This work presents such an algorithm, together with the analysis of its performance. The algorithm presented is fully capable of being run in a distributed and incremental fashion, and to be tuned for different network scenarios. This work also contributes algorithms for keeping the network in a more optimal state, by removing routes taking up the most resources.

The concept of trading off the equipment cost and routeing flexibility has not been carried out before. This work has shown that with good traffic predictions a fixed architecture with no routeing flexibility gives the lowest blocking probability. When traffic deviates away from that used to provision the network routeing flexibility is of advantage, and a form of limited flexibility outperforms full flexibility.

## 7.3 Future Work

The model used in this thesis used a simple node structure with protection paths using node-disjoint routes. Since networks cope with multiple types of failure, alternative node architectures should be modelled to assess the impact of more complex failure models. The WAVEBAND architecture, using dynamically changing wavebands, could perhaps be exploited by a more advanced routeing algorithm to achieve better performance than static wavebands, or be confirmed as being of no benefit over the PATH architecture.

Both the routeing algorithm and the bandwidth allocation algorithms use internal parameters. The routeing algorithm uses the weight factors to alter the trade-off between keep routes short and increasing utilisation of equipment. Results presented in this work show that the optimal values for these parameters do change with the network scenario, but this only leads to an emperical understanding of what values these parameters should be in the general case. Whether these parameters could be determined by an efficient algorithm, or should change dynamically depending on the current state of the network, are both interesting problems.

The bandwidth allocation algorithms are simple and may not be optimal. Better algorithms could be developed, especially in the on-line case where future traffic is not known, as currently knowledge of past traffic is not fully exploited. With the current algorithms it is clear than some form of parameter adaptation would be required in a deployed network, with user controls to set targets on the frequency of allocation change.

While the traffic measurements taken show that for a single stream dynamic allocation may give benefits, it is not clear how streams may interact in a real network. For example, if all traffic in the network increases and decreases in phase, then we can only perform peak rate allocation on all traffic. Reconfiguring networks relies on traffic shifts, where one stream decreasing in required bandwidth coincides with another stream increasing. Further measurements would be needed to assess the occurrence of traffic shifts large enough to be of benefit.

## 7.4 Conclusion

The original thesis of this work was that:

> *"It is possible and desirable for an optical physical layer based on wavelength switching using DWDM to create and maintain a virtual topology distinct from the physical interconnection and to present this virtual topology to higher network layers such as IP."*

With the possible architectures and routeing algorithms given, it is certainly possible to create virtual routes across the optical network which can change over time. With the bandwidth allocation algorithms presented, it is possible to determine when allocation shifts should occur. It is an open question how good these algorithms are, since each requires internal parameters which need to be statically or dynamically tuned.

It is not clear whether traffic shifts occur in real networks. If changes in traffic levels are primarily related to a daily cycle, then a network would need a large geographic spread so that enough data from different time zones was carried to give rise to traffic shifts. Currently traffic is biased towards more developed countries, and networks tend to be centred in a single geographical area. Both traffic and the scope of networks may need to spread wider around the globe before traffic shifts become a reality.

There is currently a large deployment of IP networks, and large investment into existing protocols, equipment, and training of staff to use particular equipment. Whether some or all of the functionality of an IP network should be implemented in the optical layer would require one of two scenarios.

- Firstly, that IP networks fail to cope with rising traffic levels and competition delivering services that people are willing to pay for. With a large research investment into this area this is not an easy question to answer: estimates as to the end of Moore's Law or the exhaustion of the IPv4 address space are consistently shown to be in error.
- Secondly, the demonstration that there are clear price benefits to adopting an adapting optical layer. If the IP layer *can* cope with traffic levels with sufficient performance, then it would mean that the combined infrastructure of optical and IP layers must be cheaper than the IP layer alone. Thus there must be some simplification of the IP layer, otherwise two active layers cannot be cheaper or easier to manage than one. This in turn would require a major change in current IP deployed technology.

An astute network provider, predicting that either of these may be true at some point in the future, could then start a slow migration process. Existing technology could add static optical paths to enrich the virtual topology of an IP network. This would aid the deployment of IP equipment capable of dealing with a higher degree of connectivity. Selected nodes could then use switches to start changing the most critical routes on a smaller timescale, before making this behaviour more widespread througout their network.

The conclusion of this dissertation is that it *is* possible and *may be* desirable to create an active optical layer to enable simplification of the IP layer.

# Bibliography

[Aksyuk02] V.A. Aksyuk, S. Arney, N.R. Basavanhally, D.J. Bishop, C.A. Bolle, C.C. Chang, R. Frahm, A. Gasparyan, J.V. Gates, R. George, C.R. Giles, J. Kim, P.R. Kolodner, T.M. Lee, D.T. Neilson, C. Nijander, C.J. Nuzman, M. Paczkowski, A.R. Papazian, R. Ryf, H. Shea, and M.E. Simon. *238x238 Surface Micromachined Optical Crossconnect With 2dB Maximum Loss*. In Optical Fibre Communication, March 2002. Post Deadline Paper. (p 47)

[Austin01] Gary P. Austin, Bharat T. Doshi, Christopher J. Hunt, Ramesh Nagarajan, and M. Akber Qureshi. *Fast, Scable, and Distributed Restoration in General Mesh Optical Network*. Bell Labs Technical Journal, 6(1):67–81, January–June 2001. (pp 45, 52)

[Bala95a] Krishna Bala, Thomas E. Stern, and Kavita Bala. *Algorithms For Routing in a Linear Lightwave Network*. In Proceedings of IEEE INFOCOM, 1–9, 1995. (p 41)

[Bala95b] Krishna Bala, Thomas E. Stern, David Simchi-Levi, and Kavita Bala. *Routing in a Linear Lightwave Network*. IEEE/ACM Transactions on Networking, 3(4):459–469, August 1995. (p 45)

[Banerjee01] Ayan Banerjee, John Drake, Jonathan P. Lang, Brad Turner, Kireeti Kompella, and Yakov Rekhter. *Generalized Multiprotocolo Label Switching: An Overview of Routing and Management Enhancements*. IEEE Communications Magazine, 39(1):144–150, January 2001. (p 55)

[Barabási99] Albert-László Barabási and Réka Albert. *Emergence of Scaling in Random Networks*. Science, 286:509–512, October 1999. (p 36)

[Barford01] Paul Barford and David Plonka. *Characteristics of Network Traffic Flow Anomalies*. In AGM SIGCOMM Internet Measurement Workshop, 2001. (p 39)

[Barford02] Paul Barford, Jeffery Kline, David Plonka, and Amos Ron. *A Signal Analysis of Network Traffic Anomalies*. In AGM SIGCOMM Internet Measurement Workshop, 2002. (p 39)

[Baroni00] Stefano Baroni, John O. Eaves, Manoj Kumar, M. Akber Qureshi, Antonio Rodriguez-Moral, and David Sugerman. *Analysis and Design of Backbone Architecture Alternatives for IP Optical Networking*. IEEE Journal on Selected Areas in Communications, 18(10):1980–1994, October 2000. (pp 50, 111)

[Beneš65] V. E. Beneš. *Mathematical Theory of Connecting Networks and Telephone Traffic*. Academic Press, New York, 1965. (p 47)

[Beran94] Jan Beran. *Statistics for Long-Memory Processes*, volume 61 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1994. (pp 38, 72)

[Bertsekas87] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, 1987. (p 158)

[Binetti00] Stefano Binetti, Attilo Bragheri, Eugenio Iannore, and Filippo Bentivoglio. *Mesh and Multi-Ring Optical Networks for Long-Haul Applications*. IEEE/OSA Journal of Lightwave Technology, 18(12):1677–1684, December 2000. (pp 111, 112)

[Boll00] D.W. Boll, J. Donovan, R.L. Graham, and B.D. Lubachevsky. *Improving Dense Packings of Equal Disks in a Square*. Electronic Journal of Combinatorics, 7(R46), 2000. (p 108)

[Borella97] M. S. Borella, J. P. Jue, D. Banerjee, B. Ramamurthy, and B. Mukherjee. *Optical Components for WDM Lightwave Networks*. In Proceedings of the IEEE, 85:1274–1307, August 1997. (p 40)

[Box76] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ: Prentice Hall, 1976. (p 37)

[Calient02] *Calient Networks DiamondWave Ships for Revenue, Passes NEBS and ETSI Testing*. `http://www.calient.net/press.html`, 2002. (p 47)

[Calvert97] Ken Calvert, Matt Doar, and Ellen W. Zegura. *Modeling Internet Topology*. IEEE Communications Magazine, June 1997. (p 36)

[Cao01a] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun. *The Effect of Statistical Multiplexing on the Long-Range Dependence of Internet Packet Traffic*. Bell Labs Technical Journal, 2001. (p 86)

[Cao01b] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun. *On the Nonstationarity of Internet Traffic*. In SIGMETRICS/Performance, 102–112, 2001. (p 39)

[Cao02] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun. *Internet Traffic: Statistical Multiplexing Gains*. DIMACS Workshop on Internet and WWW Measurement, Mapping and Modeling, 2002. (p 86)

[Chen02] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger. *The Origin of Power Laws in Internet Topologies Revisited*. In Proceedings of IEEE INFOCOM, 2002. (p 37)

[Chiaroni97] D. Chiaroni, B. Lavigne, A. Jourdan, L. Hamon, C. Janz, and M. Renauld. *New 10 Gbit/s 3R NRZ Optical Regenerative Interface based on Semiconductor Optical Amplifiers for All-Optical Networks*. In Proceedings of IOOC/ECOC, 5:41–44, September 1997. (p 48)

[Chiba01] T. Chiba, H. Arai, K. Ohira, H. Nonen, H. Okano, and H. Uetsuka. *Wavelength Splitters for DWDM Systems*. IEEE LEOS Newsletter, 15(5), October 2001. (p 46)

[Chlamtac92] I. Chlamtac, A. Ganz, and G. Karmi. *Lightpath Communications: An Approach to High-Bandwidth Optical WAN's*. IEEE Transactions on Communications, 40(7):1171–1182, July 1992. (p 41)

[Cinkler00] Tibor Cinkler, Daniel Marx, Claus Popp Larsen, and Daniel Fogaras. *Heuristic Algorithms for Joint Configuration of the Optical and Electrical Layer in Multi-Hop Wavelength Routing Networks*. In Proceedings of IEEE INFOCOM, 1000-1009, 2000. (pp 41, 124)

[Coffman01] K. G. Coffman and A. M. Odlyzko. *Is there a "Moore's Law" for Data Traffic?* In J. Abello, P. Pardalos, and M. Resende, editors, *Handbook of Massive Data Sets*. Kluwer Academic Publishers, 2001. (p 19)

[Colle02] Didier Colle, Sophie De Maesschalck, Chris Develder, Pim Van Heuven, Adelbert Groebbens, Jan Cheyns, Ilse Lievens, Mario Pickavt, Paul Lagasse, and Piet Demeester. *Data-Centric Optical Networks and Their Survivability*. IEEE Journal on Selected Areas in Communications, 20(1):6–20, January 2002. (p 50)

[Crochat00] Olivier Crochat, Jean-Yves Le Boudec, and Ornan Gerstel. *Protection Interoperability for WDM Optical Networks*. IEEE/ACM Transactions on Networking, 8(3):384–395, June 2000. (pp 27, 50)

[Crowcroft98] Jon Crowcroft and Philippe Oechslin. *Differentiated End-to-End Internet Services using a Weighted Proportional Fair Sharing TCP*. Computer Communication Review, 28(3):53–67, 1998. (p 158)

[Czumaj99] A. Czumaj and A. Lingas. *On Approximability of the Minimum-Cost k-Connected Spanning Subgraph Problem.* In 10th Annual ACM-SIAM Symposium on Discrete Algorithms, 281–290, 1999. (p 108)

[D'Agostino86] R. B. D'Agostino and M. A. Stephens, editors. *Goodness of Fit Techniques.* Marcel Dekker, 1986. (pp 68, 69)

[Davis01] R. Drew Davis, Krishnan Kumaran, and Iraj Saniee. *SPIDER: A Simple and Flexible Tool for Design and Provisioning of Protection Lightpaths in Optical Networks.* Bell Labs Technical Journal, 6(1):82–97, January–June 2001. (p 36)

[Dijkstra59] E. W. Dijkstra. *A Note on Two Problems in Connexion with Graphs.* Numerische Mathematik, 269–271, 1959. (p 104)

[Doar96] M.B. Doar. *A Better Model for Generating Test Networks.* In Proceedings of IEEE Globecom, November 1996. (pp 36, 107)

[Doshi99] Bharat T. Doshi, Subrahmanyam Dravida, P. Harshavardhana, Oded Hauser, and Yufei Wang. *Optical Network Design and Restoration.* Bell Labs Technical Journal, 4(1):58–84, January–March 1999. (p 52)

[Doshi01] Bharat T. Doshi, Ramesh Nagarajan, G. N. Srinivasa Prasanna, and M. Akber Qureshi. *Future WAN Architecture Driven by Services, Traffic Volume, and Technology Trends.* Bell Labs Technical Journal, 6(1):13–32, January–June 2001. (p 22)

[Elmirghani00] J. M. H. Elmirghani and H. T. Mouftah. *All-Optical Wavelength Conversion Technologies and Applications in DWDM Networks.* IEEE Communications Magazine, 38(3):86–92, March 2000. (p 48)

[Erlang09] A. K. Erlang. *The Theory of Probabilities and Telephone Conversation.* Nyt Tidsskrift for Matematisk, 1909. (p 37)

[Even76] S. Even, A. Itai, and A. Shamir. *On the Complexity of Timetable and Multicommodity Flow Problems.* SIAM Journal on Computing, 5(4):691–703, December 1976. (p 41)

[Feldmann00] Anja Feldmann, Albert Greenburg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True. *Deriving Traffic Demands for Operation IP Networks: Methodology and Experience.* In Proceedings of ACM SIGCOMM, 257–270, 2000. (p 131)

[Floyd94] Sally Floyd and Van Jacobson. *The Synchronization of Periodic Routing Messages.* IEEE/ACM Transactions on Networking, 2(2):122–136, April 1994. (p 159)

[Fraleigh03] C. Fraleigh, S. Moon, C. Diot, B. Lyles, and F. Tobagi. *Packet-Level Traffic Measurements from the Sprint IP Backbone*. IEEE Network, 2003. (pp 19, 79, 93)

[Frost94] Victor S. Frost and Benjamin Melamed. *Traffic Modeling For Telecommunications Networks*. IEEE Communications Magazine, March 1994. (p 37)

[Gençata03] Ayşegül Gençata and Biswanath Mukherjee. *Virtual-Topology Adaption for WDM Mesh Networks Under Dynamic Traffic*. IEEE/ACM Transactions on Networking, 11(2):236–247, April 2003. (p 56)

[Gerstel99] Oli Gerstel, Galen Sasaki, Shay Kutten, and Rajiv Ramaswami. *Worst-Case Analysis of Dynamic Wavelength Allocation in Optical Networks*. IEEE/ACM Transactions on Networking, 7(6):833–845, December 1999. (p 48)

[Gerstel00a] Ornan Gerstel and Rajiv Ramaswami. *Optical Layer Survivability: A Services Perspective*. IEEE Communications Magazine, 38(3):104–113, March 2000. (p 50)

[Gerstel00b] Ornan Gerstel and Rajiv Ramaswami. *Optical Layer Survivability: An Implementation Perspective*. IEEE Journal on Selected Areas in Communications, 18(10):1885–1899, October 2000. (p 50)

[Gerstel00c] Ornan Gerstel, Rajiv Ramaswami, and Weyl-Kuo Wang. *Making Use of a Two Stage Multiplexing Scheme in a WDM Network*. OFC 2000 Technical Digest, ThD1-1–ThD1-3, 2000. (p 45)

[Giles90] C. R. Giles and D. J. Giovanni. *Dynamic Gain Equalization in a Twostage Fiber Amplifier*. IEEE Photonic Technology Letters, 2:866–868, December 1990. (p 43)

[Grossglauser99] Matthias Grossglauser and Jean-Chrysostome Bolot. *On the Relevance of Long-Range Dependence in Network Traffic*. IEEE/ACM Transactions on Networking, 7(5):629–640, October 1999. (p 40)

[Grover98] Wayne D. Grover and Demetrios Stamatelakis. *Cycle-Orientated Distributed Preconfiguration: Ring-like Speed with Mesh-like Capacity for Self-planning Network Restoration*. In IEEE International Conference on Communications, 537–543, June 1998. (p 50)

[Grover02] Wayne D. Grover and John Doucette. *Design of a Meta-Mesh for Chain Subnetworks: Enhancing the Attractiveness of Mesh-Restorable WDM Networking on Low Connectivity Graphs*.

IEEE Journal on Selected Areas in Communications, 20(1):47–61, January 2002. (p 52)

[Hall01] James Hall, Ian Pratt, and Ian Leslie. *Non-Intrusive Estimation of Web Server Delays*. In Proceedings of IEEE LCN2001, 215–224, November 2001. (p 60)

[Hauser02] Oded Hauser, Murali Kodialam, and T. V. Lakshman. *Capacity Design of Fast Path Restorable Optical Networks*. In Proceedings of IEEE INFOCOM, 2002. (pp 41, 52)

[Heffes86] J. R. M. Heffes and D. M. Lucantoni. *A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance*. IEEE Journal on Selected Areas in Communications, 4(6):856–868, September 1986. (p 37)

[Hopkins00] W. G. Hopkins. *A New View Of Statistics*, chapter A Scale of Magnitudes for Effect Statistics. Internet Society for Sport Science: `http://www.sportsci.org/resource/stats/effectmag.html`, 2000. (p 127)

[Hosking81] J. R. M. Hosking. *Fractional Differencing*. Biometrika, 68:165–176, 1981. (p 39)

[Hurst51] H. E. Hurst. *Long-Term Storage Capacity of Reservoirs*. Transactions of the American Society of Civil Engineers, 116:770–779, 1951. (p 38)

[Iannaccone03] Gianluca Iannaccone, Chen-Nee Chuah, Supratik Bhattacharyya, and Christophe Diot. *Feasability of IP Restoration in a Tier-1 Backbone*. Technical Report 030666, Sprint ATL, March 2003. (pp 49, 109, 160)

[Ihaka96] Ross Ihaka and Robert Gentleman. *R: A Language for Data Analysis and Graphics*. Journal of Computational and Graphical Statistics, 5(3):299–314, 1996. (p 72)

[Jamin97] Sugih Jamin and Scott Shenker. *Measurement-based Admission Control Algorithms for Controlled-load Service: A Structural Examination*. Technical Report CSE-TR-333-97, University of Michigan, April 1997. (p 154)

[JAN] *JANET*. `http://www.ja.net/`. (p 60)

[Jeong96] G. Jeong and E. Ayanoglu. *Comparison of Wavelength Interchanging and Wavelength Selective Cross Connects in Multiwavelength All-Optical Networks*. In Proceedings of IEEE INFOCOM, 156-163, 1996. (p 49)

[JRoute] *jroute: internet route measurement software*. `http://www.cl.cam.ac.uk/Research/SRG/netos/netx/`. (p 64)

[Jung02] Jaeyeon Jung, Balachander Krishnamurthy, and Michael Rabinovich. *Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites*. In 11th World Wide Web Conference, May 2002. (p 39)

[Kalmanek02] Charles Kalmanek. *A Retrospective View of ATM*. Computer Communication Review, 32(5):13–21, November 2002. (p 94)

[Karasan98] Ezhan Karasan and Ender Ayanoglu. *Effects of Wavelength Routing and Selection Algorithms on Wavelength Conversion Gain in WDM Optical Networks*. IEEE/ACM Transactions on Networking, 6(2):186–196, April 1998. (p 49)

[Katsube97] Y. Katsube, K. Nagami, and H. Esaki. *RFC2098: Toshiba's Router Architecture Extensions for ATM : Overview*, February 1997. (p 54)

[Kelly97] Frank Kelly. *Charging and rate control for elastic traffic*. European Transactions on Telecommunications, 8:33–37, 1997. (p 158)

[Kilpi02] Jorma Kilpi and Ilkka Norros. *Testing the Gaussian approximation of aggregate traffic*. In AGM SIGCOMM Internet Measurement Workshop, 2002. (p 38)

[Krishmaswamy01] Rajesh M. Krishmaswamy and Kumar N. Sivarajan. *Design of Logical Topologies: A Linear Formulation for Wavelength-Routed Optical Networks with No Wavelength Changers*. IEEE/ACM Transactions on Networking, 9(2):186–198, April 2001. (p 36)

[Kruskal56] J. B. Kruskal. *On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem*. Proceedings of the American Mathematical Society, 7(1):48–50, 1956. (p 108)

[Labourdette97] Jean-François P. Labourdette. *Performance Impact of Partial Reconfiguration on Multihop Lightwave Networks*. IEEE/ACM Transactions on Networking, 5(3):351–358, June 1997. (p 45)

[Labovitz99] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. *Origins of Internet Routing Instability*. In Proceedings of IEEE INFOCOM, 218–226, 1999. (p 87)

[Lakshman97] T. V. Lakshman and Upamanyu Madhow. *The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss*. IEEE/ACM Transactions on Networking, 5(3):336–350, June 1997. (p 158)

[Lau95] W.-C. Lau, A. Erramilli, J. L. Wang, and W. Willinger. *Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and its Properties*. In Proceedings of the IEEE ICC, 466–472, 1995. (p 39)

[Laude84] J. P. Laude and C-N. Zah. *Wavelength Division Multiplexing/Demultiplexing (WDM) using Diffraction Gratings*. SPIE-Application, Theory and Fabrication of Periodic Structures, 503:22–28, 1984. (p 44)

[Lee02] Myungmoon Lee, Jintae Yu, Yongbum Kim, Chul-Hee Kang, and Jinwoo Park. *Design of Hierarchical Crossconnect WDM Networks Employing a Two-Stage Multiplexing Scheme of Waveband and Wavelength*. IEEE Journal on Selected Areas in Communications, 20(1):166–171, January 2002. (pp 45, 94)

[Lee03] Kayi Lee and Kai-Yeung Siu. *On the Reconfigurability of Single-Hub WDM Ring Networks*. IEEE/ACM Transactions on Networking, 11(2):273–284, April 2003. (p 56)

[Leland94] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. *On the Self-Similar Nature of Ethernet Traffic (Extended Version)*. IEEE/ACM Transactions on Networking, 2(1):1–15, February 1994. (pp 37, 38)

[Li94] Chung-Sheng Li, Franklin Fuk-Kay Tong, Christos J. Georgiou, and Monsong Chen. *Gain Equalization in Metropolitan and Wide Area Optical Networks using Optical Amplifiers*. In Proceedings of IEEE INFOCOM, 130–137, June 1994. (p 44)

[Madamopoulos02] Nicholas Madamopoulos, D. Clint Friedman, Ioannis Tomkos, and Aleksandra Boskovic. *Study of the Performance of a Transparent and Reconfigurable Metropolitan Area Network*. IEEE/OSA Journal of Lightwave Technology, 20(6):937–945, June 2002. (p 44)

[McAuley03] In conversation with Derek McAuley, Director of Intel Research, Cambridge UK, June 2003. (pp 44, 160)

[McKeown03] Nick McKeown. *Circuit Switching in the Core*. Talk given at Open Architectures and Network Programming Conference, April 2003. `http://tiny-tera.stanford.edu/~nickm/talks/`. (p 22)

[Médard02] Muriel Médard, Richard A. Barry, Steven G. Finn, Wenbo He, and Steven S. Lumetta. *Generalized Loop-Back Recovery in Optical Mesh Networks*. IEEE/ACM Transactions on Networking, 10(1):153–164, February 2002. (p 51)

[Medina01] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. *BRITE: An Approach to Universal Topology Generation*. In Proceedings of MASCOTS, August 2001. (p 36)

[Medina02] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. *Traffic Matrix Estimation: Existing Techniques and New Directions*. In Proceedings of ACM SIGCOMM, 2002. (p 85)

[Mitra01] Partha P. Mitra and Jason B. Stark. *Nonlinear limits to the information capacity of optical fibre communications*. Nature, 411:1027–1030, 2001. (p 22)

[Mohan01] G. Mohan, C. Siva Ram Murthy, and Aran K. Somani. *Efficient Algorithms for Routing Dependable Connections in WDM Optical Networks*. IEEE/ACM Transactions on Networking, 9(5):553–566, October 2001. (p 52)

[Mokhtar98] Admed Mokhtar and Murat Azizoglu. *Adaptive Wavelength Routing in All-Optical Networks*. IEEE/ACM Transactions on Networking, 6(2):197–206, April 1998. (p 41)

[Molinero-Fernández02] Pablo Molinero-Fernández and Nick McKeown. *TCP Switching: Exposing Circuits to IP*. IEEE Micro Magazine, 22(1):82–89, January/February 2002. (p 54)

[Molinero-Fernández03] Pablo Molinero-Fernández and Nick McKeown. *The Performance of circuit switching in the Internet*. OSA Journal of Optical Networking, 2(4), March 2003. (p 54)

[Neukermans01] Armand Neukermans and Rajiv Ramaswami. *MEMS Technology for Optical Networking Applications*. IEEE Communications Magazine, 39(1):62–69, January 2001. (p 47)

[Newman98] Peter Newman, Greg Minshall, and Thomas L. Lyon. *IP Switching – ATM Under IP*. IEEE/ACM Transactions on Networking, 6(2):117–129, April 1998. (p 54)

[Noel00] Eric Noel and K. Wendy Tang. *Performance Modeling of Multihop Network Subject to Uniform and Nonuniform Geometric Traffic*. IEEE/ACM Transactions on Networking, 8(6):763–774, December 2000. (p 36)

[Palm38] C. Palm. *Analysis of the Erlang Traffic Formulae for Busy-Signal Arrangements*. Technical Report, Ericsson, 1938. (p 37)

[Papagiannaki02] Konstantina Papagiannaki, Nina Taft, Supratik Bhattacharyya, Patrick Thiran, Kave Salamatian, and Christophe Diot. *A Pragmatic Definition of Elephants in Internet Backbone Traffic*. In AGM SIGCOMM Internet Measurement Workshop, 2002. (p 38)

[Papagiannaki03] Konstantina Papagiannaki, Nina Taft, Zhi-Li Zhang, and Christophe Diot. *Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models*. In Proceedings of IEEE INFOCOM, 2003. (pp 74, 85, 92)

[Paxson95] Vern Paxson and Sally Floyd. *Wide-Area Traffic: The Failure of Poisson Modelling*. IEEE/ACM Transactions on Networking, 3(3):226–244, June 1995. (p 37)

[Paxson97a] Vern Paxson. *End-to-End Routing Behaviour in the Internet*. IEEE/ACM Transactions on Networking, 5(5):601–615, October 1997. (pp 20, 87)

[Paxson97b] Vern Paxson. *Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic*. Computer Communications Review, 27(5):5–18, October 1997. (pp 38, 39)

[Procket03] *PRO/8812 High-Availability Router Datasheet*. `http://www.procket.com/`, 2003. (p 85)

[Ramamurthy98a] Byrav Ramamurthy, Jason Iness, and Biswanath Mukherjee. *Optimizing Amplifier Placements in a Multiwavelength Optical LAN/MAN: The Unequally Powered Wavelengths Case*. IEEE/OSA Journal of Lightwave Technology, 16(9), September 1998. (p 44)

[Ramamurthy98b] Byrav Ramamurthy, Jason Iness, and Biswanath Mukherjee. *Optimizing Amplifier Placements in a Multiwavelength Optical LAN/MAN: The Unequally Powered Wavelengths Case*. IEEE/ACM Transactions on Networking, 6(6):755–767, December 1998. (p 44)

[Ramamurthy03] Ramu Ramamurthy and Biswanath Mukherjee. *Fixed-Alternate Routing and Wavelength Conversion in Wavelength-Routed Optical Networks*. IEEE/ACM Transactions on Networking, 11(3):351–367, June 2003. (p 41)

[Ramaswami95] Rajiv Ramaswami and Kumar N. Sivarajan. *Routing and Wavelength Assignment in All-Optical Networks*. IEEE/ACM Transactions on Networking, 3(5):489–500, October 1995. (pp 41, 48)

[Ramaswami98] Rajiv Ramaswami and Galen Sasaki. *Multiwavelength Optical Networks with Limited Wavelength Conversion*. IEEE/ACM Transactions on Networking, 6(6):744–754, December 1998. (pp 36, 48)

[Ramaswami02] Rajiv Ramaswami and Kumar N. Sivarajan. *Optical Networks: A Practical Perspective*. Morgan Kaufmann, 2002. (pp 40, 43, 49)

[Rekhter97] Y. Rekhter, B. Davie, D. Katz, E. Rosen, and G. Swallow. *RFC2105: Cisco Systems' Tag Switching Architecture Overview*, February 1997. (p 55)

[Rigby03] Pauline Rigby. *Tunable Lasers Revisited*. LightReading, January 2003. `http://www.lightreading.com/`. (p 139)

[Rose96] O. Rose. *Estimation of the Hurst Parameter of Long-Range Dependent Time Series*. Technical Report 137, University of Würzburg, February 1996. (p 38)

[Rosen01] E. Rosen, A. Viswanathan, and R. Callon. *RFC3031: Multiprotocol Label Switching Architecture*, January 2001. (p 55)

[Roughan99] Matthew Roughan and Darryl Veitch. *Measuring Long-Range Dependence under Changing Traffic Conditions*. In Proceedings of IEEE INFOCOM, 1513–1521, 1999. (p 39)

[RViews] *Route Views, BGP routing table archive*. `http://www.routeviews.org/`. (p 64)

[Sabella98] Roberto Sabella, Eugenio Iannone, Marco Listanti, Massimo Berdusco, and Stefano Binetti. *Impact of Transmission Performance on Path Routing in All-Optical Transport Networks*. IEEE/OSA Journal of Lightwave Technology, 16(11):1965–1972, November 1998. (pp 48, 111)

[Sahu99] Sambit Sahu, Philippe Nain, Don Towsley, Christophe Diot, and Victor Firoiu. *On Achievable Service Differentiation with Token Bucket Marking for TCP*. Technical Report 99-72, UMASS CMPSCI, 1999. (p 76)

[Sarvotham01a] Shriram Sarvotham, Rudolf Riedi, and Richard Baraniuk. *Connection-level Analysis and Modeling of Network Traffic*. In AGM SIGCOMM Internet Measurement Workshop, 2001. (p 38)

[Sarvotham01b] Shriram Sarvotham, Rudolf Riedi, and Richard Baraniuk. *Connection-level Analysis and Modeling of Network Traffic*. Technical Report, ECE Department, Rice University, July 2001. (p 70)

[Shapiro68] S. S. Shapiro, M. B. Wilk, and H. J. Chen. *A Comparative Study of Various Tests for Normality*. Journal of American Statistical Association, 63:1343–1372, 1968. (p 68)

[Sharma00] Vishal Sharma and Emmanouel A. Varvarigos. *An Analysis of Limited Wavelength Translation in Regular All-Optical WDM Networks*. IEEE/OSA Journal of Lightwave Technology, 18(12):1606–1619, December 2000. (p 48)

[Smith90] David A. Smith, Jane E. Baran, John J. Johnson, and Kwok-Wai Cheung. *Integrated-Optic Acoustically-Tunable Filters for WDM Networks*. IEEE Journal on Selected Areas in Communications, 8(6):1151–1159, August 1990. (p 44)

[Sridharan02] Murari Sridharan, Murti V. Salapaka, and Arun K. Somani. *A Practical Approach to Operating Survivable WDM Networks*. IEEE Journal on Selected Areas in Communications, 20(1):34–46, January 2002. (p 51)

[Stamatelakis00] D. Stamatelakis and W. D. Grover. *Theoretical Underpinnings for the Efficiency of Restorable Networks Using Preconfigured Cycles ("p-cycles")*. IEEE Transactions on Communications, 48(8):1262–1265, August 2000. (p 51)

[Strand01] John Strand, Robert Doverspike, and Guangzhi Li. *Importance Of Wavelength Conversion In An Optical Network*. Optical Networks Magazine, 2(3):33–44, May/June 2001. (pp 48, 91)

[Subramaniam96] Suresh Subramaniam, Murat Azizoglu, and Arun K. Somani. *All-Optical Networks with Sparse Wavelength Conversion*. IEEE/ACM Transactions on Networking, 4(4):544–557, August 1996. (p 48)

[Subramaniam99] Saresh Subramaniam, Marat Azizoglu, and Arun K. Somani. *On Optimal Converter Placement in Wavelength-Routed Networks*. IEEE/ACM Transactions on Networking, 7(5):754–766, October 1999. (pp 36, 48)

[Tanmunarunkit02] Hongsuda Tanmunarunkit, Ramesh Govindan, Sugih Jamin, Scott Skenker, and Walter Willinger. *Network Topology Generators: Degree-Based vs. Structural*. In Proceedings of ACM SIGCOMM, 2002. (p 37)

[Taqqu85] M. S. Taqqu. *A bibliographical guide to self-similar processes and long-range dependence*. In E. Eberlain and M. S. Taqqu, editors, *Dependence in Probability and Statistics*, 137–165. Basel:Birkhauser, 1985. (p 38)

[Teverovsky97] V. Teverovsky and M. S. Taqqu. *Testing for Long Range Dependence in the Presence of Shifting Means or a Slowly Decaying Trend using a Variance Type Estimator*. Journal of Time Series Analysis, 18:279–304, 1997. (p 39)

[Thompson97] Kevin Thompson, Gregory J. Miller, and Rick Wilder. *Wide-Area Internet Traffic Patterns and Characteristics*. IEEE Network, 11(6):10–23, November/December 1997. (pp 39, 54, 74)

[Toba93] H. Toba, K Takemoto, T Nakanishi, and J Nakano. *A 100-Channel Optical FDM Six-Stage In-line Amplifer System Emplying Tunable Gain Equalization*. IEEE Photonic Technology Letters, 5:248–250, February 1993. (p 43)

[Tornatore02] Massimo Tornatore, Guido Maier, and Achille Pattavina. *WDM Network Optimization by ILP Based on Source Formulation*. In Proceedings of IEEE INFOCOM, 2002. (p 41)

[Véhel97] Jacques Lévy Véhel and Rudolf Riedi. *Fractional Brownian motion and data traffic modeling: The other end of the spectrum*. Fractals in Engineering, 1997. (p 39)

[Veitch99] Darryl Veitch and Patrice Abry. *A Wavelet Based Joint Estimator of the Parameters of Long-Range Dependence*. IEEE Trans. Inform. Theory, 45(3):878–897, April 1999. (p 38)

[Vellekoop91] Arjen R. Vellekoop and Meint K. Smit. *Four-channel integrated-optic wavelength demultiplexer with weak polarization dependence*. IEEE/OSA Journal of Lightwave Technology, 9(3):310–314, March 1991. (p 44)

[Vutukury01] Srinivas Vutukury and J. J. Garcia-Luna-Aceves. *MDVA: A Distance-Vector Multipath Routing Protocol*. In Proceedings of IEEE INFOCOM, 557–564, 2001. (p 95)

[Waxman88] B. M. Waxman. *Routing of Multipoint Connections*. IEEE Journal on Selected Areas in Communications, 6(9):1617–1622, 1988. (p 36)

[Whittle51] P. Whittle. *Hypothesis testing in time series analysis*. Hafner, New York, 1951. (p 72)

[Willinger95] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson. *Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*. In Proceedings of ACM SIGCOMM, 100–113, 1995. (p 39)

[Willinger96] Walter Willinger, Murad S. Taqqu, and Ashok Erramilli. *A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High-Speed Networks*. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*, 339–366. Clarendon Press, Oxford, 1996. (p 37)

[Xiao99] Gaoxi Xiao and Yiu-Wing Leung. *Algorithms for Allocating Wavelength Converters in All-Optical Networks*. IEEE/ACM Transactions on Networking, 7(4):545–557, August 1999. (p 48)

[Xiong99] Yijun Xiong and Lorne G. Mason. *Restoration Strategies and Spare Capacity Requirements in Self-Healing ATM Networks*. IEEE/ACM Transactions on Networking, 7(1):98–110, February 1999. (p 50)

[Zang03] Hui Zang, Canhui Ou, and Biswanath Mukherjee. *Path-Protection Routing and Wavelength Assignment (RWA) in WDM Mesh Networks Under Duct-Layer Constraints*. IEEE/ACM Transactions on Networking, 11(2):248–258, April 2003. (p 52)

[Zaumen91] William T. Zaumen and J. J. Garcia-Luna Aceves. *Dynamics of Distributed Shortest-Path Routing Algorithms*. In Proceedings of ACM SIGCOMM, 31–42, 1991. (p 36)

[Zaumen98] William T. Zaumen and J. J. Garcia-Luna-Aceves. *Loop-Free Multipath Routing Using Generalized Diffusing Computations*. In Proceedings of IEEE INFOCOM, 1408–1417, 1998. (p 95)

[Zegura97] Ellen W. Zegura, Kenneth L. Calvert, and Michael J. Donahoo. *A Quantitative Comparison of Graph-based Models for Internet Topology*. IEEE/ACM Transactions on Networking, 5(6):770–783, December 1997. (p 36)

[Zhang90] L. Zhang and D Clark. *Oscillating Behavior of Network Traffic: A Case Study Simulation*. Internetworking: Research and Experience, 1(2):101–112, December 1990. (p 159)

[Zhang95] Zhensheng Zhang and Anthony S. Acampora. *A Heuristic Wavelength Assignment Algorithm for Multihop WDM Networks with Wavelength Routing and Wavelength Re-Use*. IEEE/ACM Transactions on Networking, 3(3):281–288, June 1995. (p 41)

[Zhu03] B. Zhu, L. Nelson, S. Stulz, A. Gnauck, C. Doerr, J. Leuthold, L. Gruner-Nielsen, M. Pedersen, J. Kim, R. Lingle, Y. Emori, Y. Ohki N. Tsukiji, A. Oguri, and S. Namiki. *6.4-Tb/s (160 x 42.7 Gb/s) transmission with 0.8 bit/s/Hz spectral efficiency over 32 x 100 km of fiber using CSRZ-DPSK format*. In Optical Fibre Communication, March 2003. Post Deadline Paper 19. (p 22)