2000

An analysis of label switching forwarding mechanisms in future IP over cell networks

Paul Andrew Boustead

University of Wollongong

Recommended Citation

NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
An Analysis of Label Switching Forwarding Mechanisms in Future IP over Cell Networks

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy

from

THE UNIVERSITY OF WOLLONGONG

by

Paul Andrew Boustead
Bachelor of Engineering (Honours Class I)

SCHOOL OF ELECTRICAL, COMPUTER AND TELECOMMUNICATIONS ENGINEERING
2000
Abstract

The label switching forwarding mechanism, which was derived from the connection oriented ATM protocol, is currently being applied to IP networks. Use of this mechanism brings significant advantages in terms of simplifying forwarding decisions as well as enabling implementation of comprehensive traffic engineering mechanisms. Label switching simplifies forwarding by replacing the longest-match prefix search algorithms with a simple lookup table.

This thesis concentrates on examining the use of label switching techniques for scalable best effort unicast forwarding. Most label switching protocols are capable of operating over cell based switched networks such as ATM, and we constrain this thesis to examining this case. In essence we examine the implementation of label switching techniques in an Internet wide environment. The aim of this work is to examine ways of maximizing the percentage of packets following label switched paths and therefore reducing the number of packets that require traditional IP forwarding. We concentrate on core Internet routers.

There is a significant body of literature proposing and examining implementations of several label switching approaches such as: Multi-Protocol Label Switching (MPLS), IP Switching, IP Navigator, Aggregate Route Based IP Switching, and Tag Switching. We present an extensive literature survey. In addition we present a comprehensive classification of label switching techniques to provide scalable best effort IP forwarding in the Internet. From this work we isolate several areas of work for further consideration. Specifically, these areas concern the use of label switching forwarding in future IP networks including: an IP version 6 environment with highly aggregated routing tables; use of label switching in conjunction with congestion sen-
sitive routing; and an optically switched network.

The examination of label switching in hierarchical networks examines the impact of a high level of IP routing table aggregation that will be associated with IP version 6. We show that the popular MPLS best effort forwarding mechanism will perform poorly in this environment and will require a high level of forwarding in core gateway routers. We propose and examine the performance of a hybrid label switching protocol, called Destination Site Label Switching (DSLS), which performs well in this environment. We show using traffic traces that the network layer forwarding requirement of DSLS can be reduced to below 0.2% of total packets regardless of routing table aggregation.

Some label switching protocols, such as IP switching, abstract forwarding from routing information. We examine the use of these protocols in conjunction with congestion sensitive routing protocols, where routing information will change frequently. Experiments, using traffic traces, were performed to examine the response of these protocols to changes in underlying routing information. We investigate several mechanisms to improve the sensitivity. It was found that altering flow detection parameters did not significantly improve the performance. However, the introduction of a maximum flow length in the order of 200 seconds improved the performance significantly even for highly aggregated DSLS flows with minimal (0.13% for DSLS) additional packet layer forwarding required.

The examination of label switching in optical cell switched networks highlights a problem with a mechanism called "VC Merge" which is used by many label switching protocols. VC Merge requires additional buffers for packet reassembly, and increases switch complexity. Previous studies have shown that VC Merge does not add significantly to total buffer requirements. However, these studies examined electronic switches which are capable of implementing large switch output buffers and complex forwarding mechanisms. We use analytical and simulation techniques to examine the buffer requirements of VC merge when traffic smoothing techniques are used to reduce buffer requirements. In this scenario we find the VC merge mechanism buffer size becomes a significant part of buffer requirements and actually exceeds the average output buffer size.
Statement of Originality

This is to certify that the work described in this thesis is entirely my own, except where due reference is made in the text.

No work in this thesis has been submitted for a degree to any other university or institution.

Signed

Paul Andrew Boustead
11 August, 2000
Acknowledgments

I would like to thank my supervisor Professor Joe Chicharo for his academic advice and friendship throughout the PhD. I am particularly grateful to Joe for his guidance through difficult stages of the project. I would also like to acknowledge assistance from Professor Gary Anido in the early stages of this work.

I wish to convey my appreciation of all the other members of the SNRC lab. Their friendship made the lab a pleasant place to work. The many discussions about problems encountered were also invaluable.

I am grateful to Telstra Research Laboratories who funded this work. In particular I appreciate the many hours of assistance from Brian Keck.

Finally, I would like to thank Renee O’Leary for her support and understanding over the past few years. Her late night visits to the lab with hot dinners, as well as emotional support were very important to me.
Contents

1 Introduction ................................................. 1
  1.1 Background ........................................... 1
  1.2 Overview ............................................. 2
  1.3 Contributions ........................................ 5
  1.4 Publications ........................................ 7

2 Literature Review ......................................... 9
  2.1 Introduction ......................................... 9
  2.2 IP Label Switching Concepts ............................ 9
    2.2.1 Traditional IP Forwarding ......................... 10
    2.2.2 Label Swapping Forwarding ....................... 11
    2.2.3 Cell and Packet Label Switching .................. 12
  2.3 IP Label Switching Protocols ......................... 15
    2.3.1 Classical IP over ATM (CLIP), LAN Emulation (LANE), Multi-Protocol Over ATM (MPOA) and the Next Hop Resolution Protocol (NHRP) ......................... 16
    2.3.2 IP on ATM ...................................... 21
    2.3.3 IP Switching .................................... 24
    2.3.4 Cell Switch Router ............................... 28
    2.3.5 Threaded Indexes ................................ 28
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.6</td>
<td>Tag Switching .............................. 31</td>
</tr>
<tr>
<td>2.3.7</td>
<td>IP Navigator .............................. 34</td>
</tr>
<tr>
<td>2.3.8</td>
<td>Aggregate Route Based IP Switching ......... 35</td>
</tr>
<tr>
<td>2.3.9</td>
<td>Multi-Protocol Label Switching ............. 36</td>
</tr>
<tr>
<td>2.3.10</td>
<td>Hybrid Label Switching Proposals .......... 41</td>
</tr>
<tr>
<td>2.4</td>
<td>General Label Switching Studies .......... 43</td>
</tr>
<tr>
<td>2.4.1</td>
<td>VC Merge Evaluation ..................... 43</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Flow Analysis ............................ 44</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary .................................. 48</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Deficiencies in Existing Literature ....... 48</td>
</tr>
</tbody>
</table>

3 Classification of Scalable IP Over Cell Packet Forwarding Mechanisms  50

3.1 Introduction .................................. 50
3.2 Classification ............................... 51
3.3 Cut-Through Mechanism ....................... 54
  3.3.1 Packet Threshold Cut-through Creation .... 54
  3.3.2 Packet Type Cut-Through Creation .......... 56
3.4 Forwarding Classification .................... 57
  3.4.1 Forwarding Granularity .................... 58
  3.4.2 Forwarding Sensitivity .................... 59
  3.4.3 Inter-AS Forwarding ...................... 61
3.5 Path Determination ........................... 64
  3.5.1 Scalability .............................. 65
  3.5.2 IP Routing Extensions .................... 65
3.6 Hybrid Label Switching: Destination Site Label Switching ....... 67
3.7 Discussion .................................................. 68
3.8 Conclusions ................................................ 69

4 Label Switching in Hierarchical Networks ................. 71
  4.1 Introduction .............................................. 71
  4.2 Label Switching Approaches ................................ 74
    4.2.1 Table Linked Label Forwarding ..................... 74
    4.2.2 Other Label Switching Protocols ................... 76
  4.3 Hierarchical Networks ................................... 77
    4.3.1 IP Version 6 ....................................... 78
  4.4 Impact of Hierarchical Routing on Label Switching ........ 79
  4.5 Destination Site Label Switching ...................... 80
  4.6 Preliminary Examination of DSLS ........................ 82
  4.7 Framework for Performance Evaluation .................. 83
    4.7.1 Hierarchical Trace Collection ..................... 84
    4.7.2 Determining Site Address ........................... 87
    4.7.3 Simulation ......................................... 87
  4.8 Results ................................................. 88
    4.8.1 Performance of Destination-Site Label Switching .......... 88
    4.8.2 Higher Levels of Aggregation ....................... 91
    4.8.3 Parameter Adjustment ............................... 91
  4.9 Discussion on the Choice Packet Threshold Parameters ...... 94
  4.10 Conclusions ............................................. 95

5 Flow Based Forwarding and Frequently Changing Routing Tables .... 97
  5.1 Introduction ............................................. 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>Cut-through Forwarding Classification</td>
<td>99</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Flow Based Cut-through Forwarding</td>
<td>99</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Table-linked Cut-through Forwarding</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Dynamic Routing</td>
<td>101</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Label Switching and Dynamic Routing</td>
<td>105</td>
</tr>
<tr>
<td>5.4</td>
<td>Previous Studies of Aggregated Flows</td>
<td>106</td>
</tr>
<tr>
<td>5.5</td>
<td>Methodology</td>
<td>108</td>
</tr>
<tr>
<td>5.6</td>
<td>Traffic Analysis</td>
<td>110</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Case Study 1: Historical Examination of Fix West 1995-1997</td>
<td>110</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Case Study 2: Hierarchical Aggregation</td>
<td>116</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusions</td>
<td>123</td>
</tr>
<tr>
<td>6</td>
<td>Packet Forwarding in Optically Switched Networks</td>
<td>125</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>125</td>
</tr>
<tr>
<td>6.2</td>
<td>Optical Switching</td>
<td>127</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Wavelength Routing</td>
<td>127</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Slotted and Unslotted Optical Switches</td>
<td>130</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Optical Cell Switch and Buffer Designs</td>
<td>131</td>
</tr>
<tr>
<td>6.3</td>
<td>Packet Forwarding Techniques</td>
<td>133</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Non-Aggregated Forwarding</td>
<td>133</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Aggregated Forwarding</td>
<td>135</td>
</tr>
<tr>
<td>6.4</td>
<td>Factors Effecting Buffer Usage</td>
<td>136</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Optical VC Merge</td>
<td>136</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Cell Inter-arrival Time</td>
<td>136</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Traffic Burstiness</td>
<td>137</td>
</tr>
<tr>
<td>6.5</td>
<td>Methodology</td>
<td>138</td>
</tr>
</tbody>
</table>
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5.1</td>
<td>Analytical Model</td>
<td>138</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Simulation</td>
<td>144</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Verification of Simulation and Analysis</td>
<td>146</td>
</tr>
<tr>
<td>6.6</td>
<td>Comparison of Buffer Requirements</td>
<td>148</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Effect of Low Switch Utilisation</td>
<td>149</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Effect of Cell Gap on Buffer Requirements</td>
<td>150</td>
</tr>
<tr>
<td>6.7</td>
<td>Network Layer Packet Forwarding</td>
<td>155</td>
</tr>
<tr>
<td>6.7.1</td>
<td>Adjunct Router</td>
<td>156</td>
</tr>
<tr>
<td>6.7.2</td>
<td>Optical Reassembly</td>
<td>157</td>
</tr>
<tr>
<td>6.7.3</td>
<td>Simulation of Network Layer Forwarding Options</td>
<td>157</td>
</tr>
<tr>
<td>6.7.4</td>
<td>Results</td>
<td>158</td>
</tr>
<tr>
<td>6.8</td>
<td>Conclusions</td>
<td>162</td>
</tr>
<tr>
<td>7</td>
<td>Conclusions</td>
<td>165</td>
</tr>
<tr>
<td>7.1</td>
<td>Overview</td>
<td>165</td>
</tr>
<tr>
<td>7.2</td>
<td>Classification</td>
<td>165</td>
</tr>
<tr>
<td>7.3</td>
<td>Hierarchical IP Version 6 Networks</td>
<td>166</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Summary of Major Findings</td>
<td>167</td>
</tr>
<tr>
<td>7.4</td>
<td>Congestion Sensitive Routing Protocols</td>
<td>167</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Summary of Major Findings</td>
<td>168</td>
</tr>
<tr>
<td>7.5</td>
<td>Optical Cell Switches</td>
<td>169</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Summary of Major Findings</td>
<td>171</td>
</tr>
<tr>
<td>7.6</td>
<td>Future Work</td>
<td>171</td>
</tr>
</tbody>
</table>

# A Hierarchical Trace Details

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>Subnet Size Distributions</td>
<td>183</td>
</tr>
</tbody>
</table>
B Examination of Optical VC Merge Penalty Using Australian Traces 188

B.1 Introduction .................................................. 188

B.2 Results ......................................................... 188
## List of Figures

2.1 AAL5 encapsulation ........................................ 13
2.2 ATM UNI cell header ........................................ 14
2.3 Classical IP Over ATM (CLIP) network .................... 17
2.4 NHRP shortcut creation .................................... 20
2.5 MPOA architecture ......................................... 21
2.6 Architecture of an IP on ATM switch/router .......... 22
2.7 Block diagram of an IP on ATM routing card .......... 23
2.8 Architecture of an IP Switch ............................... 25
2.9 Hop by hop forwarding .................................... 26
2.10 Cut-through creation (a) stage 1 (b) stage 2 .......... 26
2.11 Operation of Threaded Indexes ............................ 30
2.12 Propagating indexes with routing protocol information . 30
2.13 ARIS label switch path setup ............................ 36
2.14 MPLS tunnel ................................................. 37
2.15 The overlay solution: (a) physical topology; (b) logical topology (Awduche, 1999) .................. 40
2.16 VC merge module architecture (Widjaja and Elwalid, 1999) .... 44
2.17 Flow definition ............................................. 45
3.1 Non-aggregated packet forwarding ....................... 58
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Aggregated packet forwarding</td>
<td>59</td>
</tr>
<tr>
<td>3.3</td>
<td>IP route aggregation</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>Hierarchical label stack</td>
<td>63</td>
</tr>
<tr>
<td>3.5</td>
<td>Hybrid label switching</td>
<td>68</td>
</tr>
<tr>
<td>4.1</td>
<td>Autonomous system path</td>
<td>76</td>
</tr>
<tr>
<td>4.2</td>
<td>Aggregatable global unicast address</td>
<td>78</td>
</tr>
<tr>
<td>4.3</td>
<td>Destination-site label switching</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary of preliminary flow results</td>
<td>83</td>
</tr>
<tr>
<td>4.5</td>
<td>Example address hierarchy</td>
<td>87</td>
</tr>
<tr>
<td>4.6</td>
<td>Subnet sizes</td>
<td>88</td>
</tr>
<tr>
<td>4.7</td>
<td>Percentage of packets routed</td>
<td>89</td>
</tr>
<tr>
<td>4.8</td>
<td>Average number of VCs required</td>
<td>90</td>
</tr>
<tr>
<td>4.9</td>
<td>Connection setups required per second</td>
<td>90</td>
</tr>
<tr>
<td>4.10</td>
<td>Percentage of packets routed versus aggregation</td>
<td>92</td>
</tr>
<tr>
<td>4.11</td>
<td>Effect of aggregation on (a) the average VC requirement and (b) the network layer forwarding requirement</td>
<td>93</td>
</tr>
<tr>
<td>4.12</td>
<td>Subnet aggregation with reduced packet threshold</td>
<td>95</td>
</tr>
<tr>
<td>5.1</td>
<td>MPLS Adaptive Traffic Engineering (MATE) example</td>
<td>104</td>
</tr>
<tr>
<td>5.2</td>
<td>Multipath network example</td>
<td>106</td>
</tr>
<tr>
<td>5.3</td>
<td>Traffic trace analysis</td>
<td>109</td>
</tr>
<tr>
<td>5.4</td>
<td>Time response to a single route change (a) over whole trace duration and (b) for the first 60 seconds after the route change</td>
<td>111</td>
</tr>
<tr>
<td>5.5</td>
<td>Flow characteristics</td>
<td>113</td>
</tr>
<tr>
<td>5.6</td>
<td>Vary flow timeout</td>
<td>115</td>
</tr>
<tr>
<td>5.7</td>
<td>Vary maximum flow length</td>
<td>117</td>
</tr>
</tbody>
</table>
5.8 MFL penalty ................................................................. 118
5.9 Time response to route changes for each Australian trace with no aggregation ................................................................. 119
5.10 Time response to route changes for different levels of aggregation (a) over whole trace and (b) after 60 seconds ................................................................. 120
5.11 Maximum Flow Length (a) MFL=100 (b) MFL=200 ................................................................. 121
5.12 Percent packets routed versus packet threshold for MFL=200 ................................................................. 122
5.13 Percent packets routed versus packet threshold with no MFL (from Chapter 4) ................................................................. 123
6.1 Wavelength routing network (a) physical topology (b) logical topology ................................................................. 128
6.2 Wavelength switched path (a) without wavelength converter (b) with wavelength converter ................................................................. 129
6.3 Fiber delay line switch (Masetti et al., 1996) ................................................................. 131
6.4 Fiber loop switch (Hunter et al., 1998b) ................................................................. 132
6.5 Non-aggregated packet forwarding ................................................................. 134
6.6 Aggregated packet forwarding ................................................................. 135
6.7 Markov chain for reassembly buffer with cell interval of g x dt ................................................................. 139
6.8 Markov chain output process of reassembly buffer ................................................................. 143
6.9 Simulation block diagram ................................................................. 145
6.10 Validation of case where G=0 ................................................................. 147
6.11 Reassembly buffer validation for G=0,5,10,15,20 ................................................................. 147
6.12 Reassembly + output buffer validation for G=0,5,10,15,20 ................................................................. 148
6.13 Comparison of buffer requirement for VC merge and non VC merge versus utilisation ................................................................. 149
6.14 Additional buffer required for VC merge (%) versus utilisation ................................................................. 150
6.15 Total buffer requirement for VC merge versus utilisation ................................................................. 151
<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.16</td>
<td>Reassembly buffer requirement for VC merge versus utilisation</td>
<td>152</td>
</tr>
<tr>
<td>6.17</td>
<td>Output buffer requirement for VC merge versus utilisation</td>
<td>153</td>
</tr>
<tr>
<td>6.18</td>
<td>Total buffer requirement for non-VC merge versus utilisation</td>
<td>153</td>
</tr>
<tr>
<td>6.19</td>
<td>Additional buffer requirement for VC merge</td>
<td>154</td>
</tr>
<tr>
<td>6.20</td>
<td>Trace: additional buffer requirement for VC merge</td>
<td>155</td>
</tr>
<tr>
<td>6.21</td>
<td>Adjunct router</td>
<td>156</td>
</tr>
<tr>
<td>6.22</td>
<td>Percentage of cells and packets routed for non-aggregated forwarding</td>
<td>158</td>
</tr>
<tr>
<td>6.23</td>
<td>VC usage versus packet threshold</td>
<td>159</td>
</tr>
<tr>
<td>6.24</td>
<td>Buffer size versus packet threshold for $10^{-5}$ packet loss probability</td>
<td>160</td>
</tr>
<tr>
<td>6.25</td>
<td>Average packet size versus packet threshold</td>
<td>161</td>
</tr>
<tr>
<td>6.26</td>
<td>Buffer size versus switch utilisation for $10^{-5}$ overflow probability</td>
<td>162</td>
</tr>
<tr>
<td>A.1</td>
<td>Example address hierarchy</td>
<td>184</td>
</tr>
<tr>
<td>A.2</td>
<td>Length distribution</td>
<td>184</td>
</tr>
<tr>
<td>A.3</td>
<td>Subnet size distribution: level 1</td>
<td>185</td>
</tr>
<tr>
<td>A.4</td>
<td>Subnet size distribution: level 2</td>
<td>186</td>
</tr>
<tr>
<td>A.5</td>
<td>Subnet size distribution: level 3</td>
<td>186</td>
</tr>
<tr>
<td>A.6</td>
<td>Subnet size distribution: level 4</td>
<td>187</td>
</tr>
<tr>
<td>B.1</td>
<td>Total buffer requirement for VC merge versus utilisation</td>
<td>189</td>
</tr>
<tr>
<td>B.2</td>
<td>Re-assembly buffer requirement for VC merge versus utilisation</td>
<td>189</td>
</tr>
<tr>
<td>B.3</td>
<td>Output buffer requirement for VC merge versus utilisation</td>
<td>190</td>
</tr>
<tr>
<td>B.4</td>
<td>Total buffer requirement for non-VC merge versus utilisation</td>
<td>190</td>
</tr>
<tr>
<td>B.5</td>
<td>Trace: additional buffer requirement for VC merge</td>
<td>191</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Complete classification ........................................ 52
3.2 Cut-through classification ......................................... 54
3.3 Forwarding mechanism classification ............................ 57
3.4 Path determination .................................................. 64
4.1 Hierarchy information .............................................. 86
5.1 Traffic traces used .................................................. 109
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAL</td>
<td>ATM Adaptation Layer</td>
</tr>
<tr>
<td>APIC</td>
<td>ATM Port Interconnect Chip</td>
</tr>
<tr>
<td>ARIS</td>
<td>Aggregate Route based IP Switching</td>
</tr>
<tr>
<td>AS</td>
<td>Autonomous System</td>
</tr>
<tr>
<td>ATM</td>
<td>Asynchronous Transfer Mode</td>
</tr>
<tr>
<td>ATMARP</td>
<td>ATM Address Resolution Protocol</td>
</tr>
<tr>
<td>BGP</td>
<td>Border Gateway Protocol</td>
</tr>
<tr>
<td>BUS</td>
<td>Broadcast Unknown Server</td>
</tr>
<tr>
<td>CIDR</td>
<td>Classless Interdomain Routing</td>
</tr>
<tr>
<td>CLIP</td>
<td>Classical IP Over ATM</td>
</tr>
<tr>
<td>CRC</td>
<td>Cyclic Redundancy Check</td>
</tr>
<tr>
<td>CSR</td>
<td>Cell Switch Relay</td>
</tr>
<tr>
<td>DBMAP</td>
<td>Discrete Batch Markov Arrival Process</td>
</tr>
<tr>
<td>DSLS</td>
<td>Destination Site Label Switching</td>
</tr>
<tr>
<td>ELAN</td>
<td>Emulated LAN</td>
</tr>
<tr>
<td>EOP</td>
<td>End of Packet</td>
</tr>
<tr>
<td>FATDLM</td>
<td>Flow Aggregated Traffic Driven Label Mapping</td>
</tr>
<tr>
<td>FEC</td>
<td>Forwarding Equivalence Class</td>
</tr>
<tr>
<td>FIB</td>
<td>Forwarding Information Base</td>
</tr>
<tr>
<td>GFC</td>
<td>Generic Flow Control</td>
</tr>
<tr>
<td>GSMP</td>
<td>General Switch Management Protocol</td>
</tr>
<tr>
<td>IETF</td>
<td>Internet Engineering Task-Force</td>
</tr>
<tr>
<td>IFMP</td>
<td>Ipsilon Flow Management Protocol</td>
</tr>
<tr>
<td>IPPE</td>
<td>IP Processing Element</td>
</tr>
<tr>
<td>IPv6</td>
<td>Internet Protocol Version 6</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ISR</td>
<td>Integrated Switch Router</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
</tr>
<tr>
<td>LANE</td>
<td>LAN Emulation</td>
</tr>
<tr>
<td>LDP</td>
<td>Label Distribution Protocol</td>
</tr>
<tr>
<td>LEC</td>
<td>LAN Emulation Client</td>
</tr>
<tr>
<td>LIS</td>
<td>Logical IP Subnet</td>
</tr>
<tr>
<td>LSP</td>
<td>Label Switched Path</td>
</tr>
<tr>
<td>MPC</td>
<td>MPOA Client</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-Protocol Label Switching</td>
</tr>
<tr>
<td>MPOA</td>
<td>Multi-Protocol Over ATM</td>
</tr>
<tr>
<td>MPS</td>
<td>MPOA Server</td>
</tr>
<tr>
<td>NHRP</td>
<td>Next Hop Resolution Protocol</td>
</tr>
<tr>
<td>NHS</td>
<td>Next Hop Server</td>
</tr>
<tr>
<td>NLANR</td>
<td>National Laboratory for Applied Networks Research</td>
</tr>
<tr>
<td>NNI</td>
<td>Network Network Interface</td>
</tr>
<tr>
<td>OSPF</td>
<td>Open Shortest Path First</td>
</tr>
<tr>
<td>OSPF - OMP</td>
<td>OSPF Optimal Multi-Path</td>
</tr>
<tr>
<td>PIM</td>
<td>Protocol Independent Multicast</td>
</tr>
<tr>
<td>PTI</td>
<td>Payload Type Identifier</td>
</tr>
<tr>
<td>PVC</td>
<td>Permanent Virtual Circuit</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RFC</td>
<td>Request For Comment</td>
</tr>
<tr>
<td>RIB</td>
<td>Routing Information Base</td>
</tr>
<tr>
<td>RIP</td>
<td>Routing Information Protocol</td>
</tr>
<tr>
<td>RSVP</td>
<td>Resource Reservation Protocol</td>
</tr>
<tr>
<td>SAR</td>
<td>Segmentation and Reassembly</td>
</tr>
<tr>
<td>SPF</td>
<td>Shortest Path First</td>
</tr>
<tr>
<td>SVC</td>
<td>Switched Virtual Circuit</td>
</tr>
<tr>
<td>TDP</td>
<td>Tag Distribution Protocol</td>
</tr>
<tr>
<td>UNI</td>
<td>User Network Interface</td>
</tr>
<tr>
<td>VC</td>
<td>Virtual Circuit</td>
</tr>
<tr>
<td>VCI</td>
<td>Virtual Circuit Identifier</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>VP</td>
<td>Virtual Path</td>
</tr>
<tr>
<td>VPI</td>
<td>Virtual Path Identifier</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

The exponential growth of the Internet motivates the development of faster networks at the core of the Internet where traffic aggregates. Transmission capacity has been increasing with the development of optical wavelength division multiplexing. However, packet forwarding speeds are lagging behind. This has led to development of faster and more scalable IP packet forwarding mechanisms modelled on the Asynchronous Transfer Mode (ATM) label swapping forwarding paradigm.

Label swapping forwarding significantly simplifies IP forwarding by using a direct lookup table instead of the more complicated longest-prefix match algorithm required for traditional IP forwarding. A flow identifier or label is attached to packets at the edge of a label switching network. Within label switches this label is used to lookup a table which contains the output port and the label to be used for the next hop. Not all packets can follow the label switched path, these packets will still require standard IP forwarding.

The most prominent of the IP label swapping forwarding proposals is Multi-Protocol Label Switching (MPLS) (Callon et al., 1997; Ahmed et al., 1997). However many other approaches have been proposed and developed in recent years including IP Switching (Newman et al., 1998), IP Navigator (Ahmed et al., 1997), Tag Switching (Rekhter et al., 1997), and Cell Switch Relay (Katsube et al., 1997). These protocols
operate over an underlying cell switched network, or add an additional packet header that contains labels.

There is a considerable deployment of ATM in backbone networks to take advantage of traffic engineering capabilities particularly the ability to create overlay networks. This thesis concentrates on the use of label switching protocols over an underlying cell based network (such as ATM). In particular we examine the use of these protocols in future network architectures such as:

- a highly hierarchical IP version 6 network with strict provider based addressing,
- more intelligent routing protocols that adapt to changing network conditions, and
- optically switched networks that have stringent buffer requirements.

1.2 Overview

This dissertation examines use of IP label switching forwarding protocols in the future Internet. Chapter 2 examines existing label switching protocols. We then provide a comprehensive review of label switching protocols used for best effort forwarding in Chapter 3. This is followed by three self contained chapters that examine different future network topologies. Chapter 4 examines the use of label switching protocols in hierarchical networks. The link between routing information and forwarding decisions is examined in Chapter 5. Chapter 6 looks at the implementation of label switching protocols in an optically switched environment. Chapter 7 concludes the thesis. The remainder of this section contains a more detailed a summary of each chapter.

Chapter 2 presents a critical review of current literature related to label switching protocols. We start by introducing label switching concepts and terms. This is followed by a description of the main protocols. The earliest forms of IP forwarding that use the label swapping paradigm are IP over ATM protocols such as LAN Em-
ulation (ATM-Forum, 1995a) and Multi-Protocol over ATM (Fredette, 1997), these protocols are described first. This is followed by descriptions of other protocols and concepts including: Threaded Indexing (Chandranmenon and Varghese, 1996), IP on ATM (Parulkar et al., 1995), IP Switching (Newman et al., 1998), Cell Switch Router (Katsube et al., 1997), Tag Switching (Rekhter et al., 1997), IP Navigator (Ahmed et al., 1997), Aggregate Route Based IP Switching (ARIS) (Woundy et al., 1997) and Multi-Protocol Label Switching (MPLS) (Viswanathan et al., 1998). A review of literature that compares the protocols is then presented. Chapter 2 concludes with a list of several deficiencies that we have uncovered in current literature.

A comprehensive classification of label switching protocols is presented in Chapter 3. Existing literature generally classifies protocols into into two categories: data driven and control driven. The classification presented in this thesis takes into account: cut-through trigger, flow definition, cut-through creation, path determination, forwarding sensitivity and forwarding decision. We also examine the inter-Autonomous System forwarding capability of label switching protocols. Each section of the classification is discussed in detail, particularly in relation to possible effects in the future Internet. This classification was published in (Boustead et al., 1999). Chapter 3 concludes by highlighting a list of the important problems and issues that will be investigated in the remainder of the dissertation.

Chapter 4 examines the use of label switching in a highly hierarchical IP version 6 Internet. We show that protocols which link label switching forwarding to routing table entries will perform significantly worse at gateway routers in such a network. This is due to the expected high levels of IP route aggregation, and leads to a high percentage of packets requiring network layer forwarding in switch boundary routers. We then present a hybrid label switching protocol, called Destination Site Label Switching (DSLS), that we have developed for this environment. This approach takes advantage of elements of control and data driven label switching. Control driven label switching is used within autonomous systems between interior gateway protocol routers (e.g. OSPF routers). Data driven label switching is used between gateway routers where route aggregation would introduce a problem. The data driven label switching approach uses a novel flow definition. Flows are defined
using the IP version 6 (IPv6) address hierarchy. A methodology is then presented which allows us to examine the use of this hybrid label switching approach in an IP version 6 network. We use traffic traces from a major Australian Internet service provider’s backbone and map subnet address and routing information from Internet address registries and routing databases from major servers around the world. The remainder of Chapter 4 discusses the implementation of this methodology and describes tests that examine the performance of our hybrid label switching approach. We show that DSLS Autonomous System boundary routers will switch greater than 99.8% of packets on a label switched path regardless of routing table aggregation. All packets within Autonomous Systems will not require network layer forwarding. The various aspects of this work was published in (Boustead et al., 1998a; Boustead and Chicharo, 2000e) and is currently under review for (Boustead and Chicharo, 2000c).

The interaction of label switching forwarding and future congestion based dynamic protocols, such as Open Shortest Path First (OSPF) Optimal Multi-Path is examined in Chapter 5. Several label switching protocols such as IP Switching and inter-AS DSLS do not link IP forwarding directly to routing protocol information. If congestion based dynamic routing protocols are used then routing tables are likely to change rapidly in response to network conditions. When routes change due to changing work load there are two actions that can be taken: tear-down all existing label switched paths, or maintain existing label switched paths and only forward new flows using the new routing information. The first alternative will result in a high percentage of network layer forwarding at the time of the route change. Chapter 5 examines the second approach. We examine the sensitivity of label switching protocols to route changes. We show that adjustments to packet threshold parameters does not significantly improve performance, however, the introduction of a maximum flow length was shown to have significant advantages. This work was published in (Boustead and Chicharo, 2000a).

Chapter 6 examines the use of label swapping forwarding protocols in an optically switched cell network. This chapter concentrates on determining the practicality of implementing protocols that use VC merge over optical cell switches. Previous studies (Widjaja and Elwalid, 1999) examining the impact of VC merge concentrate on
electronic switches which have large buffers (in comparison to optical buffers) and show that VC merge does not introduce a significant delay or buffer penalty. However, because the use of VC merge complicates the cell forwarding mechanism it may make it impractical for use in optical cell switches due to buffer and complexity constraints. Optical switch designs are discussed first. We then describe two scenarios that we envisage for the implementation of label switching protocols over these switches: the use of an adjunct electrical router, and optical reassembly. The methodology that we use for the comparison is then described. We believe that gaps between cells are important to consider when examining VC merge performance. Cell gaps may be introduced by cell level manipulation of traffic within switches, such as preferential queueing or traffic smoothing mechanisms. We develop a Markov model of a switch VC merge reassembly buffer with cell gaps. An existing DBMAP/D/1 output buffer model (Widjaja and Elwalid, 1999; Neuts, 1989) is modified to include cell gaps. This analytical model is used in conjunction with a discrete event simulation to examine the buffer requirements of VC merge in two different scenarios that lead to reduced output buffer usage. These scenarios are: maintaining a low switch utilisation to reduce output buffer usage, and use of control mechanisms to reduce buffer usage. We show that in both these cases the use of VC merge significantly increases the size of output buffers required. Aspects of this work were published in (Boustead and Chicharo, 2000d) and is under review for publication in (Boustead and Chicharo, 2000b).

Chapter 7 concludes the thesis with a summary of the major results obtained and identifies open research issues for future work.

1.3 Contributions

The contributions contained in this thesis are listed below. The section where this work is first discussed is also indicated.

1. Developed a classification of best effort unicast IP label swapping mechanisms. This classification serves as a framework for examining the advantages and
disadvantages of each approach. (Section 3).

2. Identified and examined a problem with MPLS and Tag Switching forwarding mechanisms in future networks with a high level of route aggregation, which is a likely scenario after the introduction of IP version 6. A high percentage of packets will require network layer forwarding at boundary routers in the core of the Internet (Section 4.4).

3. Developed a mechanism to examine network wide route aggregation using actual backbone traffic traces in conjunction with Internet wide address registry and routing database information (Section 4.7). Using this mechanism we examined the switching performance of data driven label switching with varying levels of flow aggregation (Section 4.8.2). This mechanism was also used to determine appropriate packet threshold values for highly aggregated data driven label switching protocols (Section 4.8.3).

4. Proposed a new hybrid label switching protocol called Destination Site Label Switching (DSLS) for use in an IP version 6 Internet (Section 4.5). This protocol combines the advantages of both control and data driven label switching protocols. The data driven component of this protocol uses a novel method of aggregating flows that uses the IP version 6 address hierarchy. In addition we provide results of a comprehensive examination of the switching performance of DSLS (Section 4.8).

5. Examined the abstraction of forwarding and routing mechanisms in an environment where routes change frequently due to congestion (Chapter 5). We show that data driven label switching is particularly insensitive to changes in underlying routing protocols, and this may be getting worse over time due to changing Internet flow characteristics (Section 5.6.1). We also show that the level of flow aggregation also has a large effect on the sensitivity of a protocol to underlying route changes. Changing flow detection parameters has little effect on the sensitivity to underlying route changes (Section 5.6.1.1). However, implementing a maximum flow length of 200 seconds substantially improves sensitivity while impacting little on the percentage of packets requiring network layer forwarding (Section 5.6.1.1). Use of a maximum flow length of
200 seconds is also shown to be effective for high levels of aggregation (Section 5.6.2).

6. The performance of label switching protocols in optically switched networks is examined. We propose and examine two scenarios for the implementation of label switching over such switches. These scenarios are: use of an adjunct electrical router for packets that require network layer forwarding, and performing network layer forwarding and VC merge in the optical switch fabric (Section 6.7).

7. Identified a problem with the use of VC merge in an optically switched environment where buffers are limited and forwarding complexity is a significant issue. VC merge requires additional buffering for packet reassembly, and increases the complexity of cell forwarding over traditional cell switching. If mechanisms are used to reduce buffer usage (using traffic smoothing techniques or low switch utilisation) then VC merge buffers can become a significant percentage of the total buffer requirement (Section 6.6). In fact, the introduction of traffic smoothing mechanisms can increase the VC merge buffer requirement.

8. Developed an analytical model to examine the effect of cell inter-arrival time on VC merge buffers. We also extended an existing output buffer model to allow for cell inter-arrival gaps. This model was validated against a discrete event simulation (Section 6.5.1).

1.4 Publications


Additional papers are currently under review:


Chapter 2

Literature Review

2.1 Introduction

This chapter introduces the current literature related to IP label switching protocols. Section 2.2 describes the major concepts related to IP label switching while Section 2.3 introduces the major label switching variants including IP and ATM integration protocols (Laubach and Halpern, 1998; ATM-Forum, 1995a; Luciani et al., 1998), MPOA (Fredette, 1997), IP on ATM (Parulkar et al., 1995), IP Switching (Newman et al., 1998), Cell Switch Relay (CSR) (Katsube et al., 1997), Threaded Indexes (Chandranmenon and Varghese, 1996), Tag Switching (Rekhter et al., 1997), Aggregate Route Based IP Switching (ARIS) (Woundy et al., 1997), IP Navigator (Ahmed et al., 1997), and MPLS (Viswanathan et al., 1998). General label switching studies are examined in Section 2.4. Section 2.5 summarises and examines several deficiencies in the literature.

2.2 IP Label Switching Concepts

IP label switching is defined as the use of a label swapping forwarding paradigm, similar to that used by ATM, to facilitate the forwarding of IP packets. Initial label switching approaches were designed to increase the speed of packet level forwarding by simplifying the forwarding mechanism. Examples of these early label switching protocols include IP over ATM (Parulkar et al., 1995), IP Switching (Newman
et al., 1996), and Cell Switch Router (Katsube et al., 1997). These protocols were designed assuming an underlying cell switched core. Later approaches take advantage of simplified forwarding in addition to certain traffic engineering advantages and are designed for both cell and packet switches. The major traffic engineering advantage introduced by the later approaches, such as MPLS, is the ability to create tunnels (Swallow, 1999). These tunnels can be used in a similar way that ATM VCs are used to create a different logical topology to the physical topology (the overlay model (Awduche, 1999)) which is a useful traffic engineering tool (Awduche, 1999).

This section introduces the major concepts related to label switching. Traditional IP forwarding and label switching forwarding are described first. This is followed by an examination of two alternatives for label encoding: utilisation of an underlying cell based ATM network, and the addition of a label field to the layer 3 header.

### 2.2.1 Traditional IP Forwarding

Traditionally forwarding of IP packets or datagrams has been guided directly by routing protocol information. A hop-by-hop forwarding paradigm is used where each router in the path uses local information as well as IP header information to determine the next-hop router on the path to the destination. This local information is contained within a table called the Forwarding Information Base (FIB) which is populated using information provided by routing protocols. There are several different terms that describe traditional IP forwarding that are used in literature and throughout this thesis: network layer forwarding, layer 3 forwarding, and routing of packets.

When a packet arrives at a router the destination address is extracted from the IP header. The FIB is then searched for the next hop for that particular packet. Entries in the FIB are identified by a route address and a prefix length for example: 130.110.10.0/24, where 130.130.10.0 is the route with a prefix length of 24 bits. This route represents a route to addresses between 130.130.10.0 and 130.130.10.255 ($2^{32-24}$ addresses). In order determine the correct output port to forward a particular packet a longest-matching-prefix lookup must be undertaken. This involves two steps: collecting a set of routes that match the address, and selecting the match with the longest prefix. For example, assuming that the FIB contains the following routes:
130.110.0.0/16, 130.0.0.0/8, 150.0.0.0/8. If we search this FIB for 130.110.10.0 then the set of matching routes is: 130.110.0.0/16, 130.0.0.0/8. The longest-matching-prefix route is 130.110.0.0/16.

Internet routers are segmented into domains. Routers within a domain fall under administrative control of a particular organisation and generally represent service provider or subscriber networks. This administrative domain is also termed an Autonomous System.

Similarly, routing protocols are also segmented. Intra-domain routing protocols, such as Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), distribute routing information within a single autonomous system. This routing information is used to forward packets across or within autonomous systems. Routers at the edge of autonomous systems, that forward packets to adjacent autonomous systems, are termed border or gateway routers. Inter-domain routing protocols distribute routing information between autonomous systems. An example of an inter-domain routing protocol is the Border Gateway Protocol (BGP). BGP routing information is used to determine forwarding paths across the Internet from source to destination autonomous systems.

### 2.2.2 Label Swapping Forwarding

Whereas traditional IP forwarding requires a longest-matching-prefix lookup to determine the next hop for a packet, label swapping forwarding uses a direct lookup mechanism similar to that used by ATM. To enable label swapping forwarding an additional field must be carried with the packet, this field is termed a label or a tag. This label is added to a packet when the packet first enters a label-swapping network. Switches within the network use the labels attached to packets and the incoming port number to lookup a table to determine the output port and a new label. The label field in the packet is then over-written (or swapped) with the new label.

Label swapping protocols do not forward all packets using labels in all switches. At some point traditional IP forwarding is required. When packets use label swapping forwarding this is described in several different ways in literature: following a label
switched path, data-link layer forwarding, layer 2 forwarding or following a cut-through route.

Label switching protocols differ substantially in the way they: bind labels to packets; create label switched paths between switches; and how they determine which packets will be forwarded by traditional means and which will follow a label switched path. These protocols will be described in more detail in Section 2.3.

2.2.3 Cell and Packet Label Switching

In order to facilitate label swapping forwarding a new header containing a label is added to the network layer packet, or an underlying cell based network (usually ATM) is used. The first approach is called packet based label switching. The second approach, that uses an underlying ATM network, is termed cell based label switching. This dissertation concentrates on cell based label switching however many of the results will also hold for packet based label switching. This section gives a brief introduction to packet and cell based label switching. The implementation details of different approaches is discussed in Section 2.3

2.2.3.1 Cell Based Label Switching

Cell based label switching utilises the standard ATM forwarding mechanism to facilitate label swapping forwarding. IP packets are segmented into ATM cells, and labels are stored in the cell's Virtual Channel Identifier (VCI) and Virtual Path Identifier (VPI) fields. In order to setup label switched paths the ATM switch VC table is manipulated to map incoming labels on incoming ports to outgoing labels on outgoing ports. Label switched paths are setup with the ATM signalling protocol or simplified signalling protocols.

IP packets are generally converted into ATM cells using ATM Adaptation Layer 5 (AAL5) encapsulation (Heinanen, 1993). In order to divide the packet into ATM cells the packet is first encapsulated, as shown in Figure 2.1, adding a cyclic redundancy check (CRC) as well as other fields. The encapsulated packet is then divided into 48 byte cells. A header is added to each of these cells. The cell header can be
Figure 2.1 AAL5 encapsulation

seen in Figure 2.2 for user network interfaces (UNI), for the network to network interface (NNI) four extra bits are used for the VPI field and the Generic Flow Control (GFC) field is removed. It is important to note that the cell header does not contain a cell sequence number to enable reassembly of the packet. Instead, the third bit of the PTI field is used to mark the end of the packet (EOP). If the EOP bit is set this denotes the last cell in the current packet, the next cell must then be the start of the next packet. Use of an End Of Packet (EOP) bit to denote the last cell in a packet necessitates that the order of cells within a VC must be strictly maintained since interleaving of cells will result in the inability to reassemble the individual packets. In order to ensure this does not happen cell streams can not be merged using the standard cell forwarding mechanisms.

Another alternative for encapsulation, although not a popular one, is the use of AAL3/4. This adaptation layer adds an additional header to the cell which contains a 10 bit cyclic redundancy check (CRC) with forward error correction capabilities, a two bit begin/middle/end packet field, and a multiplex ID (MID). This reduces the cell payload to 44 bytes. Although these additions allow easy cell interleave the additional overhead is considered to be too high (Armitage and Adams, 1995).

Network layer forwarding utilises the same underlying cell switched core. All packets that are being forwarded at the network layer are generally forwarded on the same VC between switches. At each switch the packets are reassembled, processed by a network layer forwarding engine which performs a longest-prefix routing lookup, and then segmented back into AAL5 cells and forwarded on the correct output port and VC. Network layer forwarding merges traffic streams onto one or more VCs go-
ing to the same next hop. AAL5 cell interleave is avoided because the packets are reassembled and inserted into the output stream as packets.

An examination of the efficiency of the transportation of IP packets over ATM cells is presented in (Armitage and Adams, 1995). The efficiency is measured in terms of effective cell utilisation (ECU) which represents:

\[
ECU = \frac{100 \times \text{Average bytes per packet}}{53 \times \text{Average number of cells per packet}}
\]

Traffic traces were used in (Armitage and Adams, 1995) to calculate ECU. For gateway traffic ECU was found to be 79.4%. Use of RFC1144 header compression (Jacobson, 1990) was shown, by (Armitage and Adams, 1995), to improve the efficiency resulting in ECU=88%.

Label switching protocols that support cell based label switching include IP on ATM (Parulkar et al., 1995), MPLS (Callon et al., 1999), IP Switching (Newman et al.,
1998), and CSR (Katsube et al., 1997). We also include approaches such as MPOA (Fredette, 1997) and LANE (Finn and Mason, 1996) with NHRP (Luciani et al., 1998) in our examination of IP label switching protocols. While these protocols are not traditionally termed label switching protocols, they meet our definition of label switching since they allow network layer forwarding plus label swapping forwarding (ATM cut-throughs).

2.2.3.2 Packet Based Label Switching

Packet based label switching protocols do not rely on an underlying cell based network. A label field is added to the packet. This label field is usually placed in a "shim" header inserted between network layer and data-link layer PDUs. The size and number of labels is not as restricted as cell based labels. Several protocols such as Tag Switching and MPLS allow multiple labels to be maintained.

Label switching protocols that support packet based label switching include MPLS, Tag Switching, Threaded indexes (Chandranmenon and Varghese, 1996), and IP Navigator (Ahmed et al., 1997).

2.3 IP Label Switching Protocols

This section examines protocols that use a label swapping forwarding paradigm to transport IP packets. We examine IP over ATM protocols such as MPOA and LANE first. Although not termed label switching protocols in literature these approaches fit in with our definition of IP label switching because they allow cut-through forwarding which enables forwarding of IP packets using the ATM virtual circuit identifier. Subsequently, we examine IP on ATM (Parulkar et al., 1995), IP Switching (Newman et al., 1998), Cell Switch Router (Katsube et al., 1997), Threaded Indexing (Chandranmenon and Varghese, 1996), Tag Switching (Rekhter et al., 1997), IP Navigator (Ahmed et al., 1997), Aggregate Route Based IP Switching (ARIS) (Woundy et al., 1997), and Multi-Protocol Label Switching (Viswanathan et al., 1998).
2.3.1 Classical IP over ATM (CLIP), LAN Emulation (LANE), Multi-Protocol Over ATM (MPOA) and the Next Hop Resolution Protocol (NHRP)

Classical IP over ATM (Laubach and Halpern, 1998), LAN Emulation (LANE) (Finn and Mason, 1996), and Multi-Protocol over ATM are methods of implementing an IP LAN over an ATM network. Classical IP over ATM (CLIP) was developed by the Internet Engineering Task Force (IETF), while LANE and MPOA were developed by the ATM forum. These protocols specify mechanisms to allow IP packets to be transported between hosts connected to the IP over ATM LAN, and between hosts and routers connecting the LAN to surrounding networks. The routers perform standard network layer IP forwarding. Both protocols use AAL5 to segment packets for forwarding on the underlying ATM network. By themselves CLIP and LANE do not fit into the definition of label switching provided earlier in this chapter since they do not allow creation of label switched paths, which bypass network layer forwarding. However, the use of these protocols in conjunction with the Next Hop Resolution Protocol (NHRP) does fit our definition of label switching because NHRP allows creation of a label switched path which bypasses IP network layer forwarding.

These protocols were designed to enable ATM to be utilised in a LAN environment. However, there are studies that examine the implementation of MPOA, LANE and NHRP in a backbone environment (Widjaja et al., 1999).

2.3.1.1 Classical IP Over ATM

Classical IP Over ATM (CLIP) (Laubach and Halpern, 1998) was developed by the IP over ATM working group of the IETF to provide LAN services over an ATM Network. This approach intends to be "a direct replacement for the 'wires' and local LAN segments connecting IP end-stations and routers" (Laubach and Halpern, 1998).

Classical IP over ATM introduces the concept of a Logical IP Subnet (LIS). A LIS network is shown in Figure 2.3. When a host transmits a packet to another host within a LIS an ATM VC must be created to the destination on which to forward the
packet. In order to forward a packet to a host outside the LIS it is placed on a point to point VC to a router. This router, which will be a member of more than one LIS (for example in Figure 2.3 the router belongs to LIS 1 and LIS 2), will use routing information to determine the next hop to the destination. In order to create point to point VCs to other hosts a mechanism is required to determine the ATM address of the destination host. This is done by an ATM Address Resolution Protocol Server (ATMARP).

The ATMARP service is provided by an ATMARP server (Laubach and Halpern, 1998). The ATMARP server contains a table of IP addresses and the corresponding ATM addresses of all hosts connected to the LIS. Each host must setup a VC to the ATMARP and send control packets to register its IP address and ATM address, this information is placed in the address resolution table. If a host needs to determine the ATM address of a host within the LIS an ATMARP_Request packet is sent to the ATMARP server. If the IP address is not found in the ATMARPs table then a negative reply is returned (ATMARP_NAK), otherwise an ATMARP_Reply packet with the corresponding ATM address is generated. In each LIS a mesh network is created using information from the ATMARP. This mesh network limits the size of the LIS since the number of VCs required is in the order of the square of the number of hosts.
2.3.1.2 LAN Emulation

LAN Emulation (LANE) (ATM-Forum, 1995a; Finn and Mason, 1996) mimics the implementation of IEEE 802.3 Ethernet LAN segments or IEEE 802.5 token ring LAN segments over an ATM network. A LAN implemented with LANE is termed an Emulated LAN (ELAN) and is capable of supporting multiple protocols including IP.

Each ELAN consists of a number of LAN Emulation Clients (LECs), which replace the Ethernet driver in the end stations, as well as a set of centralised services. The centralised services consists of several components including a LAN Emulation Configuration Server (LECS), a LAN emulation Server (LES), and a Broadcast Unknown Server (BUS).

The BUS enables broadcast to all LECs within the ELAN. Each LANE had a point to point connection to the BUS and the BUS has a point to multi-point multicast VC to all LECs. Any frame sent by a LEC to the BUS is forwarded on the BUS's multicast VC to all other LECs in the ELAN.

Instead of using the BUS for all LANE packet forwarding. LANE will forward a majority of packets on direct VCs from source LEC to destination LEC for traffic within the ELAN, or from source LEC to router LEC for traffic destined for anywhere outside the ELAN. In order to facilitate use of the direct VCs the LES maintains control VCs to all LECs and provides them with a host discovery service.

Although LANE forwards a majority of traffic within ELANs on point to point VCs all traffic destined for locations outside the ELAN are switched to a router which performs standard IP network layer forwarding. If LANE is used in conjunction with the Next Hop Resolution Protocol (NHRP) it is possible to bypass the routers with an ATM label switched path or cut-through.

2.3.1.3 The Next Hop Resolution Protocol

The Next Hop Resolution Protocol (NHRP) (Luciani et al., 1998; Cansever, 1998) was designed by the IP Over ATM (ION) working group of the IETF. NHRP pro-
vides, in essence, a inter-LIS address resolution mechanism which transmitting hosts can use to determine the ATM address of the destination. If the destination address resides on a non ATM network, such as an Ethernet LAN, NHRP returns the ATM address of the exit router of the ATM network. This allows the creation of cut-through routes or ATM label switched paths, that bypass inter-LIS routers, between hosts even if they don’t reside in the same LIS.

NHRP uses routing protocol information to determine mappings between the destination IP address and the destination ATM address. In order to describe the operation of the NHRP protocol we envisage an ATM network divided into LISs as shown in Figure 2.4. NHRP address resolution is provided by Next Hop Servers (NHS). In this example we assume that NHSs are located in routers connecting LISs. If source host S, which is connected to LIS 1, requests the ATM address of destination host D then it sends an address resolution packet to the Next Hop Server located at router A. The NHS A checks to see if it serves host B Otherwise, an address resolution packet is sent on the next hop to destination D, which is determined using IP routing information. In this case the next hop is router B. Router B receives the address resolution request, determines that the NHS serves D, obtains the ATM address host B, and sends an NHRP Resolution reply message back to host S. A local cache of the address resolution information is stored in each NHS to reduce the number of NHRP resolution packets that are required.

Whether to forward packets hop-by-hop, using the inter-LIS routers, or on an ATM cut-through route is a local decision made by the source host or the first router in the path. While waiting for creation of an ATM cut-through the source can (a) discard the packet, (b) buffer the packet until the connection is created, (c) forward the packet hop-by-hop to the destination (Luciani et al., 1998). The default option is (c) (Luciani et al., 1998).

In addition to use with CLIP, the NHRP protocol has been adopted by the ATM Forum and has been incorporated into Multi-Protocol Over ATM (MPOA).
2.3.1.4 MPOA

Multi-Protocol Over ATM (MPOA) (Fredette, 1997), which was developed by the ATM Forum, integrates LANE and a modified NHRP.

The components on an MPOA network can be seen in Figure 2.5 (derived in part, from (Fredette, 1997)). An MPOA host consists of an MPOA Client (MPC), and a LAN Emulation Client (LEC). The router consists of an MPOA server (MPS), LEC and IP routing functions. The MPC detects packet flows, determines which flows would benefit from creation of a cut-through flow, and creates an ATM cut-through VCs for these flows. The MPS includes a NHS (as described in Section 2.3.1.3) which resolves ATM addresses used to create the ATM cut-through VC. Packets that do not follow a cut-through route will be forwarded by LANE services as described in Section 2.3.1.2. These will follow a path determined by IP routing protocols.

The MPC uses the packet threshold technique in order to determine when a cut-through will be created. If the number of packets in the flow exceeds MPC-p1 within a certain time period (MPC-p2) then an NHRP resolution request is sent to the local MPS to resolve the ATM address of the destination IP address (where MPC-p1 and MPC-p2 are variables defined in (Fredette, 1997)). When the MPC gets a reply from
the MPS an ATM VC is created to the destination on which to forward subsequent packets. While waiting for the cut-through flow to be created packets within the flow are forwarded hop-by-hop by the routers between ELANs.

The scalability of MPOA in a Wide Area Network WAN Internet environment is examined in (Widjaja et al., 1999). This study uses traffic traces obtained from the National Laboratory for Applied Networks Research (NLANR) in conjunction with a detailed simulation model to examine the VC usage, VC setup rate, and the percentage of packets switched. The basic conclusion of this study was that, since the VC setup rate is linearly proportional to the traffic arrival rate, MPOA is not scalable to a core network environment.

2.3.2 IP on ATM

A novel method of forwarding IP packets over ATM cell switches was proposed in (Parulkar et al., 1995). Whereas the previous two examples used ATM signalling to create an end-to-end cut-through path for IP packets, this approach allows hop-by-hop cut-through creation and uses a much simplified signalling protocol.
At the core of the IP on ATM architecture is an ATM switch. The ATM switch is used to interconnect routing and line interface cards as seen in Figure 2.6. The routing interface cards perform IP routing lookups and forward packets to adjacent switch/routers or to other output ports on the same switch. The line interface cards are used to connect multiplexed hosts to the IP on ATM switch. Since we only consider backbone switches in this dissertation we do not discuss the line interface cards further.

A more detailed diagram of the routing interface card is shown in Figure 2.7. The routing card consists of a number of IP processing engines (IPPEs) chained together with ATM Port Interconnect Chips (APIC). The IPPEs implement routing and queueing strategies. The APIC supports: segmentation and reassembly (SAR); queueing; and is able to copy packets, if necessary, to the associated IPPE. The APIC maintains a VC table which contains VC numbers and a flag. If the flag is set to 1, for a particular VC, packets bypasses the IPPE. If the flag is set to 0 then packets on that VC require IP forwarding and are processed by the IPPE.

Short packet flows, which would not benefit from the creation of a cut-through, can be forwarded hop-by-hop using traditional IP forwarding techniques. Upon initiali-
Creation of a cut-through path is beneficial for longer flows and will reduce IP processing overhead. In order to facilitate cut-through forwarding, a number of PVCs are pre-setup between IPPEs in neighbouring switch/routers. The PVCs are deemed to be either active or inactive. If a packet is received, by a routing interface, on an inactive VC then the packet is processed by the IPPE. The IPPE, on the input card, performs the following tasks:

- A routing decision is made, determining the correct output port for the packet.
- Signalling messages are sent to the IPPE on the correct output port to obtain an unused VC to the next-hop switch/router. Signalling messages are sent to the ATM switch to configure the VC table to forward packets from the incoming VC to the negotiated outgoing VC.

- The packet is then forwarded on this new VC.

- The APIC on the input card is configured to bypass the IPPE for subsequent packets on this VC.

Once this operation has occurred in each switch in the path of a flow an end-to-end VC is created and IP forwarding is not performed in any of the transit switches.

Flows longer than 10ms were deemed in (Parulkar et al., 1995) to be suitable for creation of a cut-through path. IP on ATM has many similarities to another protocol called IP Switching (Newman et al., 1996) which is discussed in the next section. IP switching studies (Lin and McKeown, 1997; Newman et al., 1996; Newman et al., 1997; Newman et al., 1998) examine cut-through flow creation and tear-down mechanisms in greater detail.

### 2.3.3 IP Switching

Like the previously discussed IP on ATM approach, IP Switching (Newman et al., 1998) uses an ATM cell switch as the core switching mechanism. IP Switches support traditional hop-by-hop IP forwarding as well as cut-through forwarding where IP forwarding is bypassed on a cell switched path. The cut-through path is created hop-by-hop in a similar fashion to IP on ATM. Although there are a number of similarities with IP on ATM, IP Switching differs significantly in terms of switch architecture. In addition, IP Switching supports a more flexible mechanism to determine if flows will benefit from creation of a cut-through path.

#### 2.3.3.1 Basic Operation

The basic architecture of an IP Switch consists of an ATM switch and an IP switch controller. A block diagram of the IP Switch architecture can be seen in Figure 2.8.
A standard ATM switch is used, however ATM signalling protocols are removed and replaced with a much simpler General Switch Management Protocol (GSMP), which gives the IP Switch controller access to the switch's VC tables. IP packets are encoded using AAL5 encapsulation. The IP Switch controller is a high-end router with extensions to enable it to modify the switch VC tables. The switch controller is connected to a port on the ATM switch.

Figure 2.9 shows the path taken by packets that require network layer forwarding. In order to forward packets hop-by-hop using traditional IP forwarding default VCs are configured on switch start-up. If a packet is received on the default VC it is switched to the IP Switch controller where it is re-assembled and the forwarding decision is performed. The packet is then disassembled and forwarded on the default VC to the appropriate output port. In addition to forwarding packets the switch controller classifies packets into flows and determines if a cut-through path should be created for the remaining packets in the flow.

Creation of a cut-through can be divided into two stages as shown in Figure 2.10. In stage 1 (Figure 2.10 (a)), a cut-through is only created from the upstream IP Switch. A complete cut-through is only created when a cut-through is created between both
Consider a flow (f) which arrives on port x and is routed to port y as shown in Figure 2.9. In order to create a cut-through for flow f the switch controller selects a free VCI from port x (in this case VCI=A). The switch VC table is then altered to map VCI A to a free label on the switch controller port (VCI=A'). An IFMP message, called a redirection message, is then sent to the upstream node to instruct it to forward subsequent packets in the flow on VCI A. This is the end of stage one, as shown in the middle switch in Figure 2.10 (a). When the IP switch receives a packet on VCI-A it is switched to the switch controller. The switch controller re-assembles
the packet and uses a cached forwarding decision for VCI-A'. The packet is then forwarded onto the default VC used by hop-by-hop traffic on port y.

When the IP switch receives a redirection message from a downstream IP Switch to forward packets from flow f on VC B for port x the switch controller maps VCI A on port x to VCI B on port y. This creates a cell level cut-through and bypasses the switch controller and traditional IP forwarding as shown in Figure 2.10 (b).

2.3.3.2 Flow Detection and Classification

Flow classification and the determination of which flows will be forwarded hop-by-hop or on a cut-through flow is an important function of the switch controller. Design of these mechanisms has a significant bearing on the performance of IP Switching. Both flow classification and detection mechanisms are performed locally to each IP Switch. Flows are detected using information from IP/TCP/UDP headers. Two types of flows are defined (Lin and McKeown, 1997): port-pair (type 1), and host-pair (type 2).

Packets belong to the same port pair flows if the IP address, TCP/UDP port, type of service and time to live (TTL) fields are identical. Host pair flows include packets with the same IP addresses and TTL fields.

The cut-through decision mechanism determines which flows will benefit from creation of a cut-through path. Type 1 flows may use TCP/UDP application port numbers to determine which flows to cut-through. For example, the switch may create a cut-through for TCP traffic, which is likely to be a large number of packets, but not for Domain Name Server (DNS) requests which are likely to consist of no more than a few packets. A packet threshold technique is commonly used for host-pair flows (type 2). If a certain number of packets (P) are received in time t then a cut-through is created for the remaining packets. Cut-through routes are generally removed if no packets are received in that flow for a certain timeout period.

Studies examining the performance of IP switching are discussed in Section 2.4.2.
2.3.4 Cell Switch Router

A Cell Switch Router (CSR) (Katsube et al., 1997) is similar in architecture to an IP Switch and consists of an ATM switch connected to an IP Router. The CSR network supports both hop-by-hop and cut-through forwarding in a similar fashion to IP Switches. The major difference is in flow detection and cut-through creation. Because of the similarities between CSR and IP Switching, CSR will only be described briefly.

Flows are defined by source/destination IP address pairs, this is similar to type 2 flows used by IP switching. If a packet on this flow belongs to certain applications that are known to produce long traffic flows, then a cut-through path is created using a propriety signalling protocol called Flow Attribute Notification Protocol (FANP) which performs similar functions to IFMP. Packets belonging to this set of application flows are termed trigger packets. If a trigger packet is received in a flow a cut-through path is created for all subsequent packets in the source/destination pair (not just the packets with the same port number or application flow).

The study in (Katsube et al., 1997) used a corporate backbone trace obtained from Digital Equipment Corporation (DEC). Trigger packets were defined as any packets belonging to HTTP, Telnet, FTP or NNTP application flows. The percentage of packets switched on cut-through VCs was found to be 85%. Details on the number of VCs used and VC setup rate were not presented.

2.3.5 Threaded Indexes

A different approach to the use of a label swapping forwarding paradigm for IP forwarding, called threaded indexes, was presented in (Chandranmenon and Varghese, 1996). Threaded indexing does not rely on an underlying ATM switch and operates at the packet level. To enable label swapping forwarding a flow identifier, similar in function to an ATM VC identifier, is added to the IP header. Cut-through routes are linked directly to routing table entries and change dynamically as routing information changes. Creation of the label swapping path is independent of individual traffic flows and is only dependent on receipt of routing protocol packets. This is termed a
topology (or control) driven protocol in the literature. This methodology is also used in later protocols such as MPLS and Tag Switching for scalable high-speed unicast packet forwarding.

The operation of this forwarding mechanism is best explained with an example. Figure 2.11 shows an example of a packet following a threaded index path. When an unlabelled packet arrives at the first router (in this case router A) destined for 130.130.10.10 a standard IP look-up is performed and the most specific entry in the routing table is found. The routing table entry information includes an index (a unique value used to represent the entry), an address prefix, and a next hop index. The index is similar in concept to a VC Identifier, or an IP Switching label. In this case the value of the index is z, and the next-hop index is x. The next hop index is the index for the routing table entry that will be used to forward the packet in the next hop router (Router C). This value (x) is placed in the flow identifier field of the packet and is forwarded to router C. When the packet reaches router C the value in the flow identifier field x is used to index the correct entry in the routing table which is a simple O(1) operation. The new index (w) is then placed in the flow identifier field and the packet is forwarded to the correct output port.

In order to determine the correct value for the next hop index field routing protocols must be modified to propagate this information. When routers advertise routes to adjacent routers they must also advertise the index for that route. An example of this exchange by a distance vector routing algorithm, as used by the RIP (Routing Information Protocol) routing protocol, can be seen in Figure 2.12. Router R receives routing tables from both router A and router B. Router R selects the best route based on the lowest cost (shortest distance) and also obtains the index for the routing entry to use in R’s local routing table. A link state protocol such as OSPF can also be easily modified to include index information in link-state update control messages (Chandranmenon and Varghese, 1996).

The study in (Chandranmenon and Varghese, 1996) compares the performance of threaded indexes with cached IP lookup approaches in terms of the number of instructions required. The number of instructions required for a threaded index lookup was shown to be 8 instructions, which is significantly less than the 32 instructions
Figure 2.11 Operation of Threaded Indexes

Figure 2.12 Propagating indexes with routing protocol information
required for a Exclusive OR hash lookup.

2.3.6 Tag Switching

Tag Switching (Rekhter et al., 1997) is a label switching protocol that supports a forwarding mechanism which is similar to threaded indexing. This protocol supports a native packet mode, however, Tag Switching can also operate over and take advantage of an underlying ATM cell switched network. In addition Tag Switching supports a wider range of flow granularities including RSVP and multicast flows.

2.3.6.1 Architecture

The "label" required for label-swapping forwarding is termed a tag. This tag must be carried with each packet or cell. If an underlying ATM switch is used then tags can be placed in the VCI and VPI fields. In packet based tag switching the tag is placed in a "shim" header between the network layer and data-link layer headers.

Like threaded indexing, Tag Switching links cut-throughs to network layer routes. The creation of label switched paths (called tag binding) is independent of traffic flows and is only dependent on control messages. Tag binding information can be distributed by "piggy-backing" on existing control protocols, or by use of the Tag Distribution Protocol (TDP) (Doolan et al., 1997). Examples of protocols that can be modified to distribute tag-binding information and trigger creation of routes are: routing protocols, Resource Reservation Protocols (RSVP), and multicast protocols such as Protocol Independent Multicast (PIM) (Rekhter et al., 1997).

An advantage of using topology driven flows instead of data driven flows (such as IP Switching or MPOA) is that performance is independent of the traffic mix. If data driven flows are used then it is possible for a new "killer net-application" to change the performance of core network routers by changing the traffic mix. However, linking tags to routing table entries does not eliminate the need for a network layer forwarding component within a tag switch. IP route aggregation will cause the need for network layer forwarding (Rekhter et al., 1997).

Tag switching supports several different forwarding granularities depending upon
the control traffic used to bind tags. A tag may represent an aggregation of network routes, a multicast tree, or an individual RSVP flow. For the purpose of this dissertation we are only interested in scalable best effort unicast forwarding so we will examine destination based tag switching, which links tags to network layer routes, in more detail.

2.3.6.2 Destination Based Tag Switching

In order to support destination based tag switching the Tag Switch maintains a modified version of the Forwarding Information Base (FIB). The FIB contains routing information and associated tags. The tags represent a negotiated label switched path to the next hop for the address prefix in the routing entry.

Three different mechanisms for tag negotiation are supported to bind tags to address prefixes in the FIB (Rekhter et al., 1997): downstream, downstream on demand and upstream address allocation. When downstream allocation is used then the switch is responsible for binding tags that apply to incoming packets, on the other hand, when upstream allocation is used the switch is responsible for tags that apply to outgoing packets. If downstream on demand allocation is used then the downstream switch only allocates a tag if requested to do so by the upstream switch. Downstream allocation on demand is used when tag switching is implemented over an ATM network. When an underlying ATM network is not used downstream allocation is used for all types of tag allocation including RSVP, multicast and destination based unicast. Upstream allocation is supported but since it is not used (Rekhter et al., 1997) we will not discuss it further.

2.3.6.3 Tag Stack

Tag switching supports multiple labels in the form of a label stack. The tag stack is a useful mechanism that can be used for tunnelling. The top label is used for label swapping forwarding. A new label can be placed on top of the stack and will be used for forwarding until it is popped of the stack revealing the previous label. The label stack is placed in a "shim" header between the data-link and network layer headers. However ATM tag switching can only support a maximum of two labels, one in the
VPI field and one in the VCI field.

A use of two levels of labels was discussed in (Rekhter et al., 1997) that enables tag switches within an autonomous system to maintain only interior routing information. When a packet arrives at a Border Gateway Protocol (BGP) router on the edge of a transit autonomous system (AS) the label stack contains one label (A) that represents a tag switched path towards the destination. The label A is determined using BGP routing information. A new label (B) is then placed on the stack. This label represents a tag switched path to the egress router of the AS and is created using information from an interior routing protocols such as Open Shortest Path First (OSPF). The packet is switched through the AS using the label B. Label A is at the bottom of the stack and is not seen by the interior tag switches. When the packet reaches the exterior router label B is popped of the stack revealing the label A which is used to determine the next BGP hop router. Using this mechanism means that tag switches within an AS only needs to know interior routing protocol information.

2.3.6.4 Merging Flows

ATM Tag switching is limited in two aspects: the inability to merge VC streams, and the number of labels which can be stored in the cell header. Merging VCs is necessary to enable creation of destination based label switched paths.

Destination based label switched paths forward packets from different input VCs to the same output VC if they belong to the same routing prefix in the routing table. This is a significant problem when using ATM based Tag switching. Merging multiple input VCs into one output VC would cause cell interleaving and packet loss because of the use of ATM Adaptation Layer 5 (AAL5) (Rekhter et al., 1997). AAL5 only maintains an end of packet bit and no cell sequence numbers. There are two ways to overcome this problem: the use of multiple VCs for each routing prefix, or a modification of the cell switch to ensure that cell interleave does not occur. The use of a modified cell switch to prevent cell interleave is discussed in more detail in Section 2.4.1.

A common method of flow merge which uses multiple VCs is VP merge which is
used by IP Navigator (Ahmed et al., 1997). With this approach point-to-point VCs are created between ingress and egress nodes of the Tag Switching network. These VCs are aggregated in multi-point-to-point VPs to represent a destination based label switched path. Within the network VP switching is performed. Limitations of this approach are a poor utilisation of the total VC space and a limitation of the total number of VPs available (Viswanathan et al., 1998).

2.3.7 IP Navigator

IP Navigator (Ahmed et al., 1997) is an approach that has many similarities with Threaded Indexes, and Tag switching so this section will only describe the major differences. Whereas Tag switching supports multiple routing protocols IP Navigator is linked to and depends upon the use the OSPF intra-AS routing protocol. IP Navigator is designed to operate over an ATM network. When a packet arrives at the edge of an IP Navigator network a standard routing lookup is performed, the routing table entry contains a label which represents a label switched path to the egress router. Network layer forwarding is always required at the edge of an IP Navigator network or AS since label switched paths using intra-AS routing protocol information will always end at AS boundary routers that use the BGP routing protocol.

IP Navigator pre-establishes VCs between ingress and egress routers of the AS. Creation and modification of these VCs is triggered by receipt of OSPF routing updates. These VCs are multi-point to point VCs similar to those used by Tag Switching. In order to overcome the AAL5 cell interleaving problem multipoint to point VPs are created and interior routers perform VP switching. Within the VPs multiple VC numbers are used to differentiate packets from different flows. This is commonly called VP switching. Even though the number of VPs is limited, this solution is possible because the label switched paths are restricted by AS boundaries. For large autonomous systems OSPF areas can be used to limit the number of edge routers. IP Navigator VCs are terminated at OSPF area boundaries as well as AS boundary routers.
2.3.8 Aggregate Route Based IP Switching

Aggregate Route Based IP Switching (ARIS) (Woundy et al., 1997) utilises an under­
lying ATM network. The label switched paths are created, using routing information,
upon receipt of control messages. These label switched paths are created to forward
packets from ingress routers to an "egress identifier". One major difference between
ARIS and other control driven protocols is that creation of label switched paths is
initiated at egress routers.

Different levels of flow aggregation are made possible by several different definitions
of an egress identifier. Four types of egress identifier are presented in (Woundy et al.,
1997): destination IP prefix, egress IP address, OSPF router ID, and multicast pair.
Prevention of AAL5 cell interleave is supported by use of VP merging or use of a
VC merge capable switch.

An ARIS network consists of Integrated Switch Routers (ISRs). The ISRs maintain
three information tables: routing information base (RIB), forwarding information
base (FIB) , and VC information base. The RIB contains information calculated by
IP routing protocols and also identifies egress points. The FIB table is an extended
version of the standard router FIB and contains an egress identifier for every next
hop entry. Router FIBs generally contain a large number of IP destination prefixes
which are linked to a small number of next hop entries. ISRs contain a larger number
of next hop entries which contain egress, downstream VC, egress interface and next
hop entry. The VC information base is essentially the switch VC table.

Label switched path creation is initiated by egress routers. An example ARIS net­
work is shown in Figure 2.13. The egress router, which we name X, creates a VC
(VCI = e) to the upstream ISR A. The upstream router checks its FIB table for a
next hop entry which contains egress identifier X and sets the VC entry to e for those
entries. ISR A then creates VCs to its upstream neighbours and sets the VC field of
FIB entries which are destined for egress X. This continues until label switched paths
are created from ingress to egress nodes.
Figure 2.13 ARIS label switch path setup

2.3.9 Multi-Protocol Label Switching

Multi-Protocol Label Switching (MPLS) (Callon et al., 1999; Callon et al., 1997; Viswanathan et al., 1998) is a topology driven label switching protocol that is currently under development by the IETF MPLS working group. The aim of this working group is "to standardise a base technology that integrates the label-swapping forwarding paradigm with network layer routing" (Viswanathan et al., 1998).

The aims and motivating factors for the development of MPLS include simplifying the forwarding mechanism in IP networks by using label-swapping. MPLS is designed to forward a large percentage of packets on label switched paths while minimising the number of connections required. Another aim is to enable more flexible traffic engineering. In order to achieve these aims MPLS borrows ideas from many of the other label switching protocols, such as Tag Switching and ARIS, and integrates them into one protocol.
2.3.9.1 Label Encoding

MPLS label encoding is similar to that used by Tag Switching which is discussed in Section 2.3.6. The label is encoded in the VPI/VCI fields of ATM packets for ATM based MPLS. For packet based MPLS labels are encoded in a label header that is placed between the data-link layer and network layer headers. This label header can contain multiple labels, which are called a label stack. Encoding of this label stack is discussed in (Rosen et al., 1998). The most popular use of this label stack is the implementation of MPLS tunnels which will be discussed later in this section. Labels can be pushed onto the stack, popped off the stack, or swapped. Swapping labels is the mechanism by which packets follow a label switched path.

An example of the operation of a label stack can be seen in Figure 2.14. Two flows with different labels a and b are placed on the same tunnel c. At the start of the tunnel the label for tunnel c is pushed onto the label stack of each flow. The aggregated flows are forwarded in the MPLS tunnel. At the end of the tunnel the label for tunnel c is popped off the stack leaving the original labels.

2.3.9.2 Label Distribution and Forwarding

MPLS forwarding is similar to tag switching. The entire forwarding space is divided into Forwarding Equivalence Classes (FEC). A FEC is also referred to as a stream. Packets following a FEC have the same next hop. FECs usually correspond to routing table entries. Each FEC has an attached label which represents a label switched
path to the downstream switch. Packets require network layer forwarding at the first MPLS hop. The packet is then placed on an MPLS switched path.

Unicast label distribution is performed by the label distribution protocol (LDP). LDP uses routing protocol information to create LSPs. Two forms of label distribution are supported: independent and ordered (Viswanathan et al., 1998). The ordered approach initiates creation of the label switched path at the egress router. This is similar to the ARIS approach described in Section 2.3.8. The independent approach is similar to tag switching where label distribution is performed at each node.

Both downstream and downstream on demand label allocation is supported as used by Tag Switching. Downstream allocation is used for packet based MPLS. Downstream on demand is used by ATM based MPLS. A more detailed description of these approaches is described in Section 2.3.6.

Label allocation can be either ordered or liberal. Liberal allocation negotiates label switched paths with all neighbours even if they are not downstream nodes. This is useful if a large label space is available and allows faster reaction to routing changes. Ordered distribution only creates label switched paths to downstream nodes in order to minimise the use of labels.

2.3.9.3 Label Granularity

MPLS supports multiple flow granularities or FECs. Examples of supported granularities are (Viswanathan et al., 1998): IP Prefix, Egress Router and application flow. Use of IP prefix forwards all packets that belong to the same routing table entry on the same label switched path (as used by threaded indexes and Tag Switching). Egress router FECs forward all packets that exit the MPLS network at the same point on the same flow. The egress router concept is also used by ARIS, which defines different types of egress routers, and IP Navigator, which defines the egress as BGP edge and OSPF area egress routers. The finest level of granularity is application level flows which may be signalled by RSVP or a similar protocol.
2.3.9.4 Stream Merge

MPLS requires creation of multi-point-to-point label switched paths (like Tag switching). MPLS based on ATM requires a mechanism to allow merging of AAL5 encoded packets. This is done the same way as Tag Switching by using a modified cell switching mechanism (this is the preferred option (Viswanathan et al., 1998)) or VP merge. Both these mechanisms are discussed in more detail in Section 2.4.1.

2.3.9.5 Traffic Engineering

The process of controlling traffic flow to optimise performance in a network is termed Traffic Engineering (Xiao et al., 2000; Awduche, 1999). Traffic engineering encompass techniques that ensure efficient utilisation of network resources, planning of network capacity as well as reliable and expeditious packet forwarding through the network. There is currently significant effort examining the implementation of traffic engineering techniques leveraging on the MPLS label switching forwarding paradigm (Xiao et al., 2000; Awduche, 1999; Awduche et al., 1999; Swallow, 1999). A useful traffic engineering advantage made available by MPLS is the creation of Label Switched Path (LSP) tunnels (Swallow, 1999). These allow the creation of a different logical to physical topology. This is a useful traffic engineering tool, which is similar to the IP over ATM overlay model (Awduche, 1999), to aid in maximising the use of network elements and and reducing congestion points.

Current IP networks have limited traffic engineering capabilities. In particular routing and forwarding functions are inadequate (Awduche, 1999) leading to poor utilisation of network resources. Popular intra-AS routing protocols such as Open Shortest Path First (OSPF) select routes based on shortest path calculations using fixed link metrics. The routes are calculated independently of network congestion and other conditions. In addition the destination based forwarding, that is used by IP, tends to aggregate flows making traffic engineering difficult (Swallow, 1999). This type of route calculation and forwarding often leads to high utilisation and congestion in some nodes in the network (where shortest paths converge) while surrounding links may be under utilised. Underutilisation of alternate links when the shortest path is overloaded is an example of poor traffic engineering.
A popular existing technique for managing bandwidth in an IP Service Provider (ISP) network is use of an underlying layer 2 network, such as ATM, that supports VCs and traffic management (Awduche, 1999). This is termed the "Overlay Solution". The overlay technique enables a different logical topology to the actual physical topology. A simple example of the overlay model is shown in Figure 2.15. In this case the logical topology is a fully meshed network with each edge router of the ISP network a peer of all the other edge routers. The VCs interconnecting the routers are dimensioned and routed through the switched network to provide efficient utilisation of network resources.

MPLS can create LSP tunnels that can be used, in a similar way to VCs used in the Overlay Solution, to create a different conceptual network topology to the actual network topology (Swallow, 1999). LSP tunnels are explicitly routed label switched paths. In order to duplicate the overlay network shown in Figure 2.15 with LSP tunnels (instead of ATM VCs) a mesh of label switched paths must be created using explicit routes between all edge nodes (R1 to R6). These paths are then used as tunnels and bypass routing mechanisms at the intermediate nodes by using the label stacking mechanism as described in Section 2.3.9.1. When a packet enters the subnetwork at node R1 destined for R5 the label that represents the R1-R5 tunnel is pushed onto the packets label stack. The packet then follows the pre-engineered path through the intermediate nodes to R5. When the node reached R5 the label is removed from the
stack and the original label is then used to forward the packet to the next hop beyond R5. The term tunnel is used because the packets have been temporally "tunnelled" under normal IP routing mechanisms (Swallow, 1999). The tunnel thus appears to the routing protocol as one hop. LSPs can be created by signalling protocols such as RSVP (Awduche et al., 2000a).

There are also proposals for using constraint based routing protocols (Jamoussi, 1999; Xiao et al., 2000) to setup LSPs. With constraint based routing it will be possible for LSPs and LSP tunnels to be created based on QoS, traffic levels and other constraints. It is considered that the constraint-based routing functionality does not have to be part of the core MPLS functionality (Awduche, 1999).

2.3.10 Hybrid Label Switching Proposals

Three hybrid label switching protocols have been proposed: Flow aggregated traffic driven label mapping (FATDLM) (Nagami et al., 1999), smart IP switching (Lloyd and O'Mahony, 1998), and our approach (called Destination Site Label Switching - DSLS) which will be presented later in this thesis. All three approaches were developed independently. Smart IP switching and flow aggregated traffic mapping are both essentially traffic driven approaches similar to IP switching except the traffic driven flows are defined using routing table information. Our approach (DSLS), which will be introduced later in this thesis, incorporates control driven label switching within autonomous systems as well as traffic driven label switching to incorporate label switching in the AS path across the network. DSLS uses an new approach to aggregation on the AS path. Instead of aggregating traffic based on routing information we aggregate traffic based on the IPv6 address hierarchy.

2.3.10.1 Flow Aggregated Traffic Driven Label Mapping

FATDLM (Nagami et al., 1999) combines features of data driven and control driven label switching. Flows are defined by entries in routing tables, however these flows are created only when triggered by traffic flow. This has the advantage of significantly reducing the VC requirement in the case where routing tables are large. The trigger for creation of flows differs from that used by DSLS. A flow is created upon
receipt of a HTTP, FTP, TELNET or NNTP packet.

A simulation performance evaluation of FATDLM is presented in (Nagami et al., 1999). This study uses traffic from the WIDE backbone in Japan with traffic rates of 1.2Mb flowing into the AS and 0.9Mb flowing out of the AS. Destination network addresses were obtained using a local routing table from within their AS to define flows (2806 entries), as well as a core router table (50903 entries). The results for the core local routing table indicated 99% packets cut-through, and 99 VCs used. For the core routing table the percentage of packets cut-through was 86% with 542 VCs used. Other granularities were also examined including source to destination, source to destination-prefix, source-prefix to destination, source-prefix to destination-prefix and destination address.

2.3.10.2 Smart IP Switching

Smart IP Switching (Lloyd and O’Mahony, 1998) is similar in many aspects to Flow Aggregated Traffic Driven Label Mapping in that it uses routing information to define traffic driven flows. In addition it still allows host-pair traffic driven label switched paths similar to those used by IP Switching (Newman et al., 1998).

A network emulation package, called Vendor Independent Network Control Entity (VINCE), was used to simulate the Smart IP switching mechanism (Lloyd and O’Mahony, 1998). A 23 router network based on the Irish IP Research Network (HEANET), was simulated using this package. The test network simulated a normal IP network, an IP Switching network and a smart IP switching network. A test traffic trace was "reverse engineered" from a 8869 packet FIX-WEST (a major Internet backbone node) trace from NLANR. The routing information was estimated from the network topology. The results of this test indicated that IP switching would switch 54% of packets on a label switched path, while 99.2% of packets would be switched by smart IP switching with 1.2% control packet overhead. The results obtained for IP Switching significantly underestimates the performance indicated by other studies (some of which use FIX-WEST traffic traces) (Lin and McKeown, 1997; Newman et al., 1996; Newman et al., 1997; Newman et al., 1998) and our studies discussed later in this thesis (Boustead et al., 1998b; Boustead et al., 1998a; Boustead et al.,
2.4 General Label Switching Studies

This section discusses general label switching studies. Most of the literature examines scalability in terms of the state storage space (or number of VCs) required. We concentrate on two groups of studies: VC merge analysis and packet threshold parameter choice. VC merge techniques are a common method of improving VC scalability. There are several studies examining the performance of VC merge implementations in a traditional Internet environment. The selection of parameters for data driven label switching techniques also affects scalability in terms of VC usage as well as the percentage of packets switched.

2.4.1 VC Merge Evaluation

VC merge mechanisms, for cell switches, have been proposed for use in Label Switching protocols such as MPLS, Tag Switching, and ARIS. VC merge is required for MPLS to allow destination based label switched forwarding while maintaining VC scalability. VC merge also enables the creation of multipoint to multipoint VCs (Venkateswaran et al., 1997; Grossglauser and Ramakrishnan, 1997). Both these applications drive the implementation of VC merge capable switches. These mechanisms increase the complexity of the basic cell level forwarding mechanism to ensure AAL5 cell sequence integrity of the merging streams.

Similar VC merge mechanisms have been proposed for use in MPLS (Rosen et al., 1999; Widjaja and Elwalid, 1999) as well as for ATM multipoint-multipoint multicast VCs (Venkateswaran et al., 1997; Grossglauser and Ramakrishnan, 1997). The VC merge mechanism utilises the end of packet bit (EOP) in the ATM cell header (see Figure 2.2) to determine packet boundaries. An example of the VC merge mechanism in (Widjaja and Elwalid, 1999) can be seen in Figure 2.16. This mechanism is placed at the output port of a cell switch switch. Each cell received that belongs to a particular packet is placed in a reassembly buffer for that packet. When the last cell belonging to the packet is received the whole packet is then transferred to the output
buffer.

The performance of VC merge is examined in (Widjaja and Elwalid, 1999) using simulation and analytical techniques. The main findings of this study were that the addition of the VC merge module shown in Figure 2.16 added minimal overhead in terms of additional buffering required for the VC merge buffers when compared to the output buffer requirements. In particular when the burstiness of the traffic increases the additional buffer penalty buffers becomes less significant.

2.4.2 Flow Analysis

The performance of data driven protocols, such as IP Switching, CSR and IP on ATM, depends greatly on traffic characteristics. This section examines literature that investigates flow based traffic characteristics. A general study of flows, that is not directly related to label switching, is presented in (Claffy et al., 1995). More specific studies (Ke et al., 1998; Lin and McKeown, 1997; Newman et al., 1996; Newman et al., 1997; Newman et al., 1998) examine different flow granularities as well as the effect of varying packet threshold parameters.

Data driven label switching uses the concept of packet flows. A flow, as defined in (Claffy et al., 1995), contains packets that meet a particular flow specification. The
flow is active as long as the packets are separated by a time less than a flow timeout value. The flow definition is generally taken from the IP, TCP and UDP headers. A common example of a flow definition is all packets with the same source and destination IP addresses. An diagram of this flow definition can be seen in Figure 2.17.

The flow based study presented in (Claffy et al., 1995) used Internet traffic traces from LAN networks, university backbones, and an Internet backbone. There are several important conclusions of this work that impact on label switching designs. It was found that for host-pair flow timeouts between 16 and 128 seconds proved to be an appropriate tradeoff between router processing and the number of VCs required. Examining the backbone traces with a 64 second flow timeout showed that a large percentage of flows (approximately 60%) were shorter than one second. The TCP/UDP port number was shown to provide a good indication of expected flow duration.

A study of IP switching (Newman et al., 1998) examines two types of flow specifications called type 1 and type 2 flows. Type 1 flows use IP source, IP destination, as well as TCP/UDP port numbers to define flows. Type 2 flows use only IP source and
destination addresses to define flows.

Type 1 flows are examined in (Newman et al., 1996) using a trace driven simulation. The traffic trace was obtained from the FDDI ring connecting the San Francisco bay area to the Internet on the 29th September 1995 over a time period of five minutes. A port pair flow classification was used with a timeout of 60 seconds. This timeout value was chosen because a previous study (Claffy et al., 1995) has shown that a flow timeout in the order of 60 seconds represents a reasonable compromise between the number of VCs required and the probability that deleted flows will become active again soon. The simulation study in (Newman et al., 1996) examines the characteristics of port pair flows and determines the port numbers of flows that would benefit from creation of a cut-through. This decision was based on the number of packets in the flow before it timed out. Following is a list of the applications that were deemed suitable for a cut-through:

<table>
<thead>
<tr>
<th>Application Name</th>
<th>Application Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP in IP</td>
<td>TCP ftp.data (20)</td>
</tr>
<tr>
<td>TCP telnet (23)</td>
<td>TCP gopher (70)</td>
</tr>
<tr>
<td>TCP http (80)</td>
<td>TCP nntp (119)</td>
</tr>
<tr>
<td>TCP netbios (139)</td>
<td>TCP login (513)</td>
</tr>
<tr>
<td>TCP cmd (514)</td>
<td>TCP audio (1397)</td>
</tr>
<tr>
<td>TCP AOL (5190)</td>
<td>TCP x.11</td>
</tr>
</tbody>
</table>

The IP Switch was then simulated with cut-throughs being created for the above well known port numbers, flows with different port numbers were forwarded hop-by-hop. This resulted in 84% of packets and 91% of bytes being switched on a cut-through. The rate of cut-through creation was 92 per second. The average VC requirement was 15500. This compares to 422 cut-throughs created per second and an average usage of 42000 VCs if all flows were cut-through.

Type 2 flows were also examined in (Newman et al., 1996) with the same traffic trace as in the previous example. A packet threshold was used to determine which flows
were cut-through. The packet threshold time parameter $t$ was set to 60 seconds while the packet threshold $P$ was varied between 1 and 100 packets. A flow timeout of 60 seconds was also used. The results showed that with a packet threshold of $P=13$, the connection setup rate was approximately the same as the type 1 case examined previously. However the performance in terms of the percentage of packets switched was slightly better with 87% of packets switched and 92% of bytes. Using a cost function with varying cost weightings for the cost of cut-through and hop-by-hop forwarding it was determined that the optimum packet threshold $P$ lay in the region of 3-20 packets.

A later examination of type 2 flows (Newman et al., 1998) found that the optimum value, in terms of processing load and control messages for the switch controller, for the packet threshold $P$ was 10 packets. In addition (Newman et al., 1998) examined the switching performance of type 2 flows with two additional backbone traces. Each trace was taken, over a 10 minute period, on the 28th of February (one in the morning and one in the evening) from the same FDDI ring discussed above. Traces were also obtained from a corporate backbone. The additional FDDI backbone traces produced similar results to the previous study. However, the corporate traces switched significantly fewer packets on cut-throughs. The percentage of packets switched varied between 70% and 80%.

A study presented in (Lin and McKeown, 1997) examines type 1 and type 2 flows. VC usage and switching performance is examined for each case. Three different cut-through decision mechanisms were used: the packet threshold mechanism; port number, as used in the (Newman et al., 1996) study; and protocol number. Cut through decision based on protocol number is simple, a cut-through is created for TCP flows while UDP flows are always forwarded hop-by-hop. The performance evaluation shows that type 2 flows using the packet threshold cut-through mechanism always outperformed the other approaches in terms of both VC usage and percentage of packets and cells switched.
2.5 Summary

This chapter has examined the major label switching protocols and proposals in conjunction with performance evaluation and comparative literature. These label switching protocols use substantially different and diverse mechanisms to route, create and determine which packets will flow on label switched paths. Below we pinpoint several deficiencies that we can see in current literature. The next chapter provides a comprehensive classification of the approaches discussed here.

2.5.1 Deficiencies in Existing Literature

- A complete classification of label switching protocols which highlights the main functional differences between the many different label switching approaches has not been published. Existing classifications divide protocols into only two groups: Data driven and topology driven (White, 1998; Lloyd and O'Mahony, 1998).

- There are several papers discussing the use of constraint based routing protocols in conjunction with MPLS. An examination of constraint based routing protocol in conjunction with data driven label switching protocols does not appear to have been examined in the literature.

- Data driven label switching abstracts routing from forwarding decisions. No examination has been performed examining the effect of this abstraction in terms of response to route changes, particularly at high levels of aggregation.

- Label switching proposals concentrate on providing label switching forwarding within autonomous systems. There does not appear to be investigations of inter-AS label switching in current literature.

- It is unclear how control driven label switching techniques, such as MPLS, will perform in hierarchical networks. Future hierarchical networks such as IPv6 will enable higher levels of routing table aggregation due to a strict provider based address structure. This will effect label switching protocols that link label switched paths to routing table entries.
• The introduction of IPv6 with a strict address hierarchy and provider based address­ing will provide additional information that can be used to aggregate label switched flows. No investigation has been performed examining the use of hierarchical addressing information, available in the IPv6 aggregatable global unicast format, in order to facilitate aggregation in data driven label switching.

• Label switching protocols are designed for use in contemporary electronic switches, where memory is comparatively cheap and abundant. It is unclear how they will perform in an optically switched environment. In particular it is unclear how the complicated MPLS VC merge forwarding mechanism will effect performance in this environment.
Chapter 3

Classification of Scalable IP Over Cell Packet Forwarding Mechanisms

3.1 Introduction

As discussed in Chapter 2, there are many approaches for carrying IP over ATM including Multi-Protocol Over ATM (Fredette, 1997) (MPOA), Multi-Protocol Label Switching (MPLS) (Callon et al., 1997; Ahmed et al., 1997), several proposals similar to IP Switching (Newman et al., 1998), and a hybrid approach proposed by the author called Destination Site Label Switching (DSLS) (Boustead et al., 1998a). Label switching protocols are generally classified as traffic driven (White, 1998) (also called data driven (Lloyd and O'Mahony, 1998)) or topology driven (White, 1998) (also called control driven (Lloyd and O'Mahony, 1998)) in literature. Traffic driven protocols use packet flows to trigger creation of label switched paths and includes protocols such as IP Switching, and CSR. Topology driven label switching protocols link label switched paths to routing information. This simple classification does not discriminate between all the major functional differences between approaches. This chapter intends to provide a more complete classification of these approaches to establish the major functional differences. We concentrate on scalable unicast forwarding mechanisms over an underlying cell switched network. A three-category classification is described which includes: cut-through creation; path determination; and forwarding mechanism. This framework is then used to highlight the advantages and disadvantages of each approach.
As a result of the classification and associated discussion we find several areas of work which are fruitful for further examination. In particular we find that the popular MPLS table linked forwarding paradigm may suffer performance problems when routes within IP routing tables are highly aggregated. This leads us into the next chapter which examines label switching in highly aggregated IP version 6 networks in more detail. In addition we postulate that the VC merge mechanism may present problems in an optically cell switched environment.

This chapter is organised as follows. The next section introduces the entire classification. Section 3.3 examines the cut-through classification. The forwarding granularity classification is then discussed in Section 3.4. Path determination is discussed in Section 3.4. Section 3.6 briefly presents our hybrid label switching approach which will be examined in more detail in the next Chapter. Finally, Section 3.7 concludes.

### 3.2 Classification

This section provides a summary of the complete label switching classification that we describe in this chapter. The classification, which is shown in Table 3.1 is divided into three areas: cut-through mechanism; path determination; and forwarding mechanism. Each section of the classification is divided into several subsections to enable us to differentiate between the major label switching proposals. Each section is described briefly below. The remainder of this chapter will elaborate and give examples.

The cut-through forwarding mechanism classification examines the decision mechanisms that are used to determine how and when a cut-through or label switch path is created. Cut-through creation consists of three parts: cut-through trigger, flow definition and creation. The cut-through trigger determines which flows will be forwarded by IP and which flows will bypass IP and are forwarded directly on an ATM cut-through. The cut-through trigger can either be based on a packet threshold, or the receipt of a particular type of packet. The flow definition determines how packets are grouped together for determining if a cut-through is created. Flow definition can be divided into source/destination and destination based flows. Source/Destination
### Table 3.1: Complete classification of Scalable IP over Cell Packet Forwarding Mechanisms

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Inter-AS</th>
<th>Table Linked</th>
<th>Flow Linked</th>
<th>Non-Aggregated</th>
<th>Aggregated</th>
<th>ATM only</th>
<th>IP Only</th>
<th>IP &amp; ATM</th>
<th>Ordered</th>
<th>Independent</th>
<th>Destination</th>
<th>Source/Destination</th>
<th>Packet Threshold</th>
<th>Packet Type</th>
<th>Forwarding Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP &amp; ATM</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MPOA</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>CSK</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>IP Switch</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MPLS &amp; Tag Switch</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Single-AS</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Optional</td>
<td>Inter-AS</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

*MCI supports IP Switching mechanisms, used only if VC merge is unavailable.*
flows are defined using both the source and destination addresses and possibly including TCP/UDP port numbers. Destination based flows use the destination address (as well as other information which may include routing table entries) to define flows. Cut-through creation can be either independent or ordered. When all routers within the cut-through path are responsible for determining if a cut-through is created this is called independent creation (Viswanathan et al., 1998). When one router in the network (typically an ingress or egress node) determines if a cut-through is created this is called an ordered creation (Viswanathan et al., 1998).

The forwarding mechanism is divided into three categories: flow granularity, flow decision and inter-AS forwarding capability. Forwarding granularity is either aggregated or non-aggregated. Non-aggregated forwarding does not require VC merge and maintains separate VCs for flows between ingress and egress nodes of the MPLS network. Aggregated label switching protocols combine VCs that are destined for a particular node which is usually an egress node. The forwarding decision indicates what information is used to forward packets on a label switched path and is either flow linked or table linked. When flow linked forwarding is used the label switched paths are tied to flow information. Table linked forwarding protocols link cut-through routes to routing table entries and forwarding decisions are independent of flows. The last part of this classification is inter-AS forwarding capability which differentiates between label switching protocols that can create label switched paths over Autonomous System boundaries.

Path determination is the mechanism by which a route is determined for a packet through the network. MPOA uses a combination of IP routing protocols and the ATM Forum’s Private Network to Network Interface (PNNI). Other label switching approaches use only IP routing protocols. Both of these approaches are examined and compared. Of particular interest is the inherent scalability and functional differences. The scalability of PNNI is compared with a combination of BGP and OSPF with IPv4 and IPv6. PNNI and IP routing protocols (including IPv6, provider based addressing, and BGP4) both allow address summarisation to enable them to scale well in terms of routing table sizes. In terms of functionality PNNI offers extra functional advantages over standard OSPF such as QoS routing with multiple met-
Table 3.2 Cut-through classification

<table>
<thead>
<tr>
<th>Cut-Through Trigger and Creation</th>
<th>Packet Type</th>
<th>Flow Definition</th>
<th>Source Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Packet Threshold</td>
<td>Independant</td>
<td>Ordered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination Based</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPLS, TAG</td>
<td>ARIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP Navigator</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DSLS (Intra-AS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smart IP Switch</td>
<td>Smart IP Switch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DSLS (Inter-AS)</td>
<td>DSLS (Inter-AS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP Switch</td>
<td>MPOA</td>
</tr>
</tbody>
</table>

Current work on OSPF extensions including QoS, and multi-path routing is also examined.

3.3 Cut-Through Mechanism

The cut-through mechanism classification differentiates label switching protocols in terms of the different mechanisms that are used for creation of label switched paths. We examine several characteristics of different cut-through creation mechanisms including: the different triggers used to initiate creation of label switched paths; flow definitions; and which switches participate in making the decision to create the cut-through. The classification is presented in Table 3.2 which is divided into four quadrants examining the main differences in terms of packet trigger and flow definition.

3.3.1 Packet Threshold Cut-through Creation

The packet threshold technique is a popular mechanism that is used to determine if a label switched path should be created. If a certain number of packets are received within a certain time-frame then a cut-through is created. The cut-through paths remain active while packets are flowing on the connection, however, if the label switched path remains idle for a certain time period then it is removed. IP Switching (Newman et al., 1998), MPOA (Fredette, 1997), Smart IP Switching (Lloyd and O’Mahony, 1998) and Flow Aggregated Traffic Driven Label Mapping (FATDLM)
(Nagami et al., 1999) all use the packet threshold technique for the cut-through trigger. The packet threshold part of Table 3.2 is further divided into two types of flow definition which will be discussed below.

### 3.3.1.1 Source Pairs

IP Switching and MPOA both use source and destination addresses to group packets into flows as well as the packet threshold cut-through mechanism. The main difference between them is where the decision to cut-through is made. The ingress node of the MPOA network decides if an end-to-end cut-through is necessary (ordered cut-through creation). On the other hand, every switch in an IP switching network is involved in creating its local segment of the cut-through route (independent cut-through creation).

An IP Switching flow contains all the packets with identical source and destination IP addresses. IP Switching can operate in different modes that use additional information such as the type of application generating the flow. This information is obtained from the TCP/UDP header. However, studies (Lin and McKeown, 1997) show that using only address pairs performs better. The decision to assign a VC to a flow and to cut-through packets at the ATM level is based on a packet threshold parameter. This decision is made locally to each switch and is thus ordered cut-through creation. The first datagrams in a flow are forwarded hop-by-hop by IP to the destination. If the number of datagrams in the flow exceeds the packet threshold value then a label is assigned to the flow and subsequent packets in the flow are forwarded solely by ATM. The label represents a VC between the IP switch and adjacent IP switches and is set-up by a proprietary signalling protocol.

MPOA (Fredette, 1997) uses the IP address of the source and destination MPOA Client (MPC) to define flows. The decision to create a cut-through between the source and destination MPC is based on a packet threshold technique similar to that used by IP Switching. Before the threshold is exceeded packets are forwarded hop-by-hop via IP routers between Emulated LANs to the destination. Once the decision to create a cut-through has been made the ingress MPC generates a NHRP (a modified version of the IETF NHRP protocol (Luciani et al., 1998)) request is used to
determine the ATM address of the destination MPC. The ingress MPC creates an
ATM VC to the egress node using the ATM PNNI signalling protocol. This is an
ordered cut-through creation since one node initiates the complete label switch path
creation.

3.3.1.2 Destination Based

Destination based approaches use only the destination address from the IP header
to determine flow membership. Examples of protocols that use destination based
flow definitions with a packet threshold cut-through mechanism include: Smart IP
Switching (Lloyd and O'Mahony, 1998), and Flow Aggregated Traffic Driven Label
Mapping (FATDLM) (Nagami et al., 1999).

Both Smart IP Switching and FATDLM define flows using both IP destination ad­
dresses as well as routing table entries. Label switched forwarding is linked to rout­
ing table entries in a similar way to MPLS and label switch path creation is indepen­
dent in each switch. However, unlike MPLS these cut-through paths are only created
when triggered by traffic flow. These approaches are described in more detail in
Section 2.3.10.

3.3.2 Packet Type Cut-Through Creation

Tag Switching, Multi-protocol Label Switching (MPLS), Aggregate Route Based
IP Switching, IP Navigator and Cell Switch Relay are examples of label switching
techniques that trigger label switch path creation on the receipt of a certain packet
type. All approaches listed above use the destination address to classify flows except
Cell Switch Relay which uses source pairs.

Cell Switch Relay (CSR) uses both source and destination addresses to classify flows
as well as independent cut-through creation. CSR is similar to IP Switching except
that instead of using a packet threshold mechanism to determine if a label switch
path should be created this decision is based on the receipt of packets from particular
protocols. If a packet is received from a protocol that is known to produce long flows
then the cut-through is created.
Table 3.3 Forwarding mechanism classification

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Aggregated</th>
<th>Non Aggregated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Table Linked</td>
<td>Flow Linked</td>
</tr>
<tr>
<td>Inter-AS</td>
<td>MPLS</td>
<td>ARIS</td>
</tr>
<tr>
<td></td>
<td>Tag Switching</td>
<td>IP Navigator</td>
</tr>
<tr>
<td></td>
<td>Smart IP Switch</td>
<td>DSLS(Intra-AS)</td>
</tr>
<tr>
<td></td>
<td>FATDLM</td>
<td>DSLS (Inter-AS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPOA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP Switch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSR</td>
</tr>
</tbody>
</table>

Tag Switching and MPLS use the destination address as well as routing information to classify cut-through flows. The trigger for the creation of a Tag Switching, or MPLS cut-through is either the receipt of a standard IP routing protocol packet advertising a new route, or a Label/Tag Distribution Protocol (LDP or TDP) packet (Doolan et al., 1997). Unicast best effort MPLS and Tag switching cut-through creation is generally performed independently at each switch. However, both these protocols also support the ordered approach.

Aggregate Route Based IP Switching (ARIS) is similar to Tag Switching and MPLS. Cut-through flows are classified using the destination address as well as routing information. However, the cut-through creation is ordered. It is the egress node that initiates the creation of label switched paths.

3.4 Forwarding Classification

The forwarding classification is divided into three sections: forwarding granularity, forwarding decision, and inter-AS forwarding capability. Forwarding granularity is further divided into aggregated and non-aggregated. Forwarding decision examines the the link between routing information and forwarding. The inter-AS forwarding capability indicates which protocols support creation of label switched paths across autonomous system boundaries.
3.4.1 Forwarding Granularity

We group packet forwarding protocols into two groups, non-aggregated and aggregated forwarding, depending upon the need for AAL5 cell stream merging support within switches. Non-aggregated approaches, include IP Switching and MPOA do not require VC merge. Aggregated approaches such as MPLS and Tag Switching require VC merge reassembly buffers to maintain AAL5 cell sequence integrity.

Non-aggregated forwarding techniques maintain separate virtual circuits (VCs) for source/destination pairs as shown in Figure 3.1. Examples include IP Switching, MPOA and CSR. These approaches generally define flows using source and destination addresses. Some approaches such as IP Navigator, MPLS and Tag Switching allow destination based flows while using non-aggregated forwarding by using VP merge.

VP merge maintains separate VCs between ingress and egress points. These VCs are then aggregated into a multi-point to point VP. VP switching is then used within the network. The multi-point to point VP is essentially a tunnel. VP merge is seen as an unscaleable flow merge alternative for MPLS (Viswanathan et al., 1998) due to the large VC space required and the limited VP space available. The preferred approach is to use VC merge which is described below. However, VP merge is seen as a viable alternative for some approaches, such as IP Navigator, which restrict label switched paths to autonomous system boundaries. For large OSPF autonomous systems VP merge can be made scalable by further restricting label switched paths to OSPF areas.
Aggregated forwarding techniques merge one or more VCs from different input ports to a single VC on the output ports as shown in Figure 3.2. The merging of VCs leads to the necessity for packet reassembly at VC merge points. This is due to the use of AAL5 which uses only an end of packet bit in the last cell of a segmented packet for delineation. If cells belonging to AAL5 encoded packets are interleaved then the packets can no longer be re-assembled. Examples of aggregated approaches include Tag Switching, hop by hop routing, MPLS and IP Navigator. VC merge is discussed in more detail in Chapter 2.4.1.

### 3.4.2 Forwarding Sensitivity

The label switching forwarding decision can be characterised into two groups: table linked forwarding and flow linked forwarding. With table linked forwarding each individual packet is forwarded based upon the current state of the routing table. Flow linked forwarding forwards all packets belonging to the same flow along the same route. The forwarding decision is made at the start of the flow and all subsequent packets follow that route.

Table linked protocols include: ARIS, IP Navigator, Smart IP Switch, FATDLM, Tag Switching and MPLS. All these protocols link forwarding to routing information in the same way. A traditional router maintains a table called the Forwarding Information Base (FIB) which is calculated from routing information. The FIB contains a list of routing prefixes and associated next hop information. This table is used to determine which next hop router to send each packet to. Table linked label switching
protocols add an extra field to this table which is a pointer to a label database. In the case of cell based label switching this is a pointer to a VC table entry. The VC represents a label switched path to the next hop router. When all switches in the network link VCs to their routing table entries label switched paths will be created that follow IP routes. When a packet arrives at the edge of a label switching network the FIB is searched for the longest matching prefix. The label associated with this prefix is placed in the cell’s VCI/VPI field and the packet is forwarded on a label switched path following the IP route through the network.

The advantage of table linked forwarding is rapid adaptation of cut-through routes to changes in the calculated best route to the destination for datagrams. One problem with table linked forwarding cut-throughs is that it is necessary to ensure that route changes do not take place mid-packet. This would cause cells from one AAL5 packet to be sent on different paths that would cause cell reordering and loss of the packet. If routing table changes take place frequently a practical implementation must ensure that all cells belonging to the same packet are delivered along the same route through the network. Another problem may be caused by different packets from the same TCP flow following different paths to the destination. This may cause packets being received by the TCP source out of sequence and cause performance problems.

Flow linked forwarding creates a label switched path for each flow. Once the label switched path is created all subsequent packets follow that path. Examples of protocols that support flow based forwarding are: MPOA, CSR and IP Switching. MPOA performs flow linked forwarding of cut-through datagrams. A modified NHRP protocol is used to determine the destination ATM address for a particular datagram flow and an end-to-end VC is then signalled and routed using PNNI. The datagrams are then forwarded at the cell level to the destination on a fixed path. IP Switching and other data driven label switching techniques also perform flow-based forwarding. A cut through is created via individual switches for particular flows by obtaining forwarding information from the routing table and creating a cut-through VC from a upstream label switch to the appropriate output port. Once the label switch path is created subsequent packets in the flow follow that fixed path.

The advantage of flow linked forwarding is that the sequence of delivery of TCP
packets is ensured. A problem of flow linked forwarding is the inability to quickly adapt to changes in network conditions due to congestion or network failure.

3.4.3 Inter-AS Forwarding

This section examines the use of label switching protocols in a wide area Internet. We consider the use of label switched paths across multiple Autonomous Systems. We examine in particular the percentage of packets that require network layer forwarding at boundary routers between autonomous systems. Table 3.3 shows which protocols are able to create label switched paths that cross autonomous system boundaries.

Inter-AS forwarding does not appear to be addressed comprehensively in literature. Some protocols such as IP Navigator and ARIS create label switched paths to egress nodes. Once at the egress nodes it is necessary to perform network layer forwarding to determine the next hop (or the next label switched path). Protocols such as IP Switching, CSR, MPOA, and Smart IP Switching will switch the same percentage of packets, on label switched paths, across network boundaries as in other routers. It is more complicated, however, to examine the performance of MPLS and Tag Switching at boundary routers. MPLS and Tag Switching are able to create cut-through routes between autonomous systems since label switched paths can be linked to BGP routing information. The percentage of packets that follow label switched paths at boundary routers is going to be affected greatly by the level of route aggregation.

3.4.3.1 The Effect of Flow Aggregation on MPLS and Tag Switching

In this section we briefly examine the performance of Tag switching and MPLS at gateway routers along an autonomous system path. In particular we discuss the effect of IP routing table aggregation on this performance.

Binding of tags to entries in the FIB does not eliminate the need for network layer forwarding (Rekhter et al., 1997). Traditional IP forwarding must be performed at the entry into a Tag Switching network to untagged datagrams. Additional IP forwarding is required since routing tables do not necessarily maintain a route to all possible destinations due to aggregation within IP routing tables. An example of this can
Figure 3.3 IP route aggregation

be seen in Figure 3.3, where packets flowing from hosts connected to networks A and B are destined for hosts connected to networks E and F. The routing tables (or FIBs) at gateway routers in A and B are searched to find the longest prefix match for the destination IP address. A tag is obtained from the TIB entry linked to the corresponding FIB entry. The packets are then placed on a cut-through using that tag. Since in this example, the routes to E and F are aggregated in the routers in networks A, B and C, packets going to E and F will have the same tag and therefore traverse the same cut-through. When the packets reach network D which has separate entries for networks E and F, IP forwarding is necessary to determine the next hop for packets. Thus it is not possible to cut-through a packet from source to destination if route aggregation exists on the path. In fact, some implementations, for example IP Navigator (Ahmed et al., 1997), forward all packets at the network layer at ingress and egress nodes of ASs.

The MPLS working standard proposes an alternative to routing packets at the end of aggregated IP routes. This involves negotiation between Tag Switches at points of IP route aggregation and de-aggregation to enable set-up of label stacks. The
example network in Figure 3.4 shows a three level hierarchical network with aggregation nodes (A,B), de-aggregation nodes (C,D,E), and associated tag stacks. At aggregation point A the current top label is discarded and the label for the associated de-aggregation point (3) is pushed on the stack. The label representing the switched path to the IP next hop (1) is then placed on the top of the stack. This is then repeated at node B. At the de-aggregation points the top label is removed from the stack and the underlying label is used to switch the packet to the next de-aggregation point or to the destination. The problem with this approach is that aggregation nodes such as A and B need to know where the corresponding de-aggregation point is in order to negotiate a unique label. Current routers do not contain sufficient information to resolve such a situation. This method would require a complicated protocol to determine the topology to the destination for packets before they can be routed. It is anticipated that this would introduce long connection set-up delays. In addition, large amounts of hierarchy and route information is required to be stored in Tag Switches. For the cell switched case only two labels are available so using label stacks to forward packets through a hierarchical network is not possible.
### Table 3.4 Path determination

<table>
<thead>
<tr>
<th>Use of IP and ATM Routing Protocols</th>
<th>Use of only IP Routing Protocols</th>
<th>Use of only ATM Routing Protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPOA (PNNI)</td>
<td>MPLS IP Switching</td>
<td>MPOA (I-PNNI)</td>
</tr>
<tr>
<td>Classical IP Over ATM</td>
<td>HLS</td>
<td></td>
</tr>
</tbody>
</table>

### 3.5 Path Determination

Label switching over ATM and MPOA require the ability to forward datagrams hop-by-hop by routers as well as create cell level "cut-through" routes that bypass Segmentation And Reassembly (SAR) and IP forwarding. This requires the use of IP routing protocols, PNNI or some other protocol to create cell-level "cut-through" routes. This section classifies label switching protocols with respect to their use of routing protocols. The use of currently available routing protocols such as PNNI, BGP and OSPF will be discussed as well as future routing protocols including Integrated PNNI and QoS capable IP routing.

IP over ATM approaches can be divided into two distinct groups depending upon use of routing protocols. This classification can be seen in Table 3.4. This section examines the advantages and disadvantages of the different approaches. The major considerations are scalability, and functionality.

The use of IP routing protocols in conjunction with PNNI is the ATM Forum’s preferred method and is the basis for MPOA. IP routing protocols are used for routing hop by hop traffic between routers. Cut-through routes are created using NHRP to map the destination IP address to an ATM address. PNNI routing is used to determine the path to a destination and PNNI signalling is then used to create a cut-through VC. Other label switching approaches use IP routing protocols to determine the path for cut-through as well as hop-by-hop packets and do not use PNNI routing or signalling. Integrated-PNNI (Jeffords, 1996) takes the opposite approach and tries to simplify IP over ATM by removing IP routing protocols and use PNNI for both hop-by-hop as well as cut-through packets.
3.5.1 Scalability

The ATM Forum’s PNNI routing protocol is widely considered to be highly scalable and functional. It provides a many level hierarchical routing structure with address summarisation at each level to maintain small routing databases and reduce routing convergence times as well as comprehensive QoS routing support. There has been much work (Bass, 1997) over the last few years which has substantially improved the scalability of BGP. Prior to the introduction of the current version of the Border Gateway Protocol (BGP4) and Classless Inter-Domain Routing (CIDR) the Internet was viewed as an unstructured and non-hierarchical interconnection of routing domains or ASs (Bass, 1997). With the introduction of BGP4, CIDR provider-based IP address allocation, and variable bit subnet masking, it became possible to reduce the size of routing tables by aggregating routes. This is similar in concept to the address summarisation of PNNI. However, CIDR and BGP4 have some problems. CIDR does not require domains to renumber if they change to a different provider (Fuller et al., 1993). This has the potential to significantly reduce gains in routing table aggregation over time. In addition, the IPv4 address space is still becoming exhausted, just at a slower rate. The introduction of IPv6 will overcome the problem of address renumbering resulting in a scalable IP routing solution.

3.5.2 IP Routing Extensions

IP routing protocols are divided into exterior (e.g. BGP) and interior routing protocols (e.g. OSPF, RIP). Exterior routing protocols have a global scope and determine a path between administrative boundaries or ASs. Interior routing protocols only have scope within a particular AS. This section describes some of the extensions that have been proposed for IP routing protocols including QoS routing support, utilisation based dynamic routing and multi-path routing support.

There are currently several proposals for adding QoS extensions to the OSPF interior routing protocol. OSPF is a link-state routing protocol similar to PNNI. This similarity enables easy extension of OSPF to allow QoS routing. Opaque Link-State Advertisement (Opaque-LSA) can be used to flood the network with QoS related information. A proposal in (Guerin et al., 1997) discusses the use of OSPF with Opaque-
LSA and QoS path selection algorithms. Another interesting proposal, called OSPF Optimised Multipath (OSPF-OMP), adds congestion sensitive multipath forwarding capabilities.

Popular IP routing protocols such as OSPF are insensitive to changes in network conditions such as congestion, and are only capable of providing equal cost routes to destinations based on fixed weightings defined at each hop. Previously there have been attempts at adding delay based dynamic routing to the ARPANET's SPF routing protocol (Glazer and Tropper, 1990). These approaches were shown to provide effective routing based congestion control and substantial improvements in delay and throughput. However, these approaches did suffer problems due to course granularity of routing changes leading to severe oscillation (Villamizar, 1998). An IETF working group is working on a proposal to allow best effort traffic to be dynamically routed around congested links. This proposal called OSPF Optimised Multi-Path (OSPF-OMP) (Villamizar, 1998) works by propagating information about load levels in links across the AS. The flooding of this information is facilitated by use of Opaque LSA. The link-state information includes link load, packet loss, and link capacity. Using this information each router can use the internal link state topology map to determine the lowest-cost route to the destination according to current network conditions and the utilisation of links within the network. A finer granularity of control is facilitated by the ability to forward a proportion of packets on multiple links.

Adding QoS capabilities or dynamic congestion based routing to the BGP inter-AS routing protocol will be more difficult (Crawley et al., 1998). Whereas changes to interior routing algorithms can be implemented easily, as they only have scope within administrative domains, substantial changes to BGP would require Internet wide router upgrades. This scale of upgrade has historically been difficult in the Internet, for example the necessary upgrading from IPv4 to IPv6 is still progressing very slowly.

Use of dynamic routing leads to a situation where the best or lowest cost route in terms of delay, utilisation or loss will change as network conditions change. Use of packet-based forwarding will take advantage of these changes in network conditions.
and forward packets on the appropriate route. However, the connection oriented flow-based approach will react more slowly to these changes.

3.6 Hybrid Label Switching: Destination Site Label Switching

A hybrid solution for use with IPv6 called Destination Site Label Switching (DSLS) is proposed by the authors in (Boustead et al., 1998a). This approach is discussed in more detail in Chapter 4, however, to enable a more complete classification it is forward referenced here. DSLS was designed to operate efficiently in hierarchical IPv6 environments, by reducing the number of packets that need to be routed at hierarchical boundaries (Boustead et al., 1998a). DSLS uses both nonaggregated table linked label switching, and aggregated packet threshold label switching. Two levels of labels are used and stored in the ATM cells VPI and VCI fields. Non-aggregated table linked label switching is used within ASs to set-up VP paths from ingress to egress nodes, this can be seen in Figure 3.5. The aggregated packet threshold label switching technique is used for VC paths to the destination using information from the IPv6 header. Aggregating traffic based upon address hierarchy is possible because IPv6 will have a strict provider based addressing structure. The purpose of this approach is to maintain the ability of many nodes to switch all packets (as in Tag switching), while eliminating the effect of IP route aggregation on performance by not assigning labels based on routing table entries on the AS path.

3.6.0.1 Classification

Table 3.1 shows how DSLS fits into the classification. Both inter-AS DSLS and intra-AS DSLS are described.

DSLS uses both packet threshold and packet type triggers. A packet threshold trigger is used for inter-AS forwarding while packet type trigger is used for intra-AS forwarding. Destination based flow definition is used for intra-AS. Inter-AS forwarding also uses destination based flows defined by the destination-site address if VC merge is available. If VC merge is unavailable then flows are defined by both source-site
and destination-site addresses. DSLS uses independent cut-through creation.

The hybrid DSLS approach includes both flow and table linked forwarding. Flow linked forwarding is performed by gateway or BGP label switches between ASs. Table linked forwarding is used within ASs. The use of flow linked forwarding on the AS path is not a significant disadvantage since as discussed in Section 3.5 path-vector BGP is unable to take advantage of dynamic multi-path forwarding and therefore routes are likely to remain static. However, DSLS can take advantage of future multi-path dynamic algorithms within ASs because of its use of packet-based forwarding.

### 3.7 Discussion

This chapter presents an overall classification of a wide range of IP over cell based networks that summarises the major differences of each approach. The overall classification can be seen in Table 5.

MPOA uses packet threshold source/destination cut-throughs which require unscalable state information storage (Widjaja et al., 1999). In addition each switch/router
in the network must forward a percentage of packets at the network layer (approximately 15% (Newman et al., 1998; Lin and McKeown, 1997)). An advantage of MPOA is the use of PNNI signalling for cut-throughs, which is advantageous for QoS connections. However, the flow-based forwarding nature of MPOA results in a slow adaptation to changing routing tables.

MPLS uses control driven destination based cut-throughs for scalable unicast forwarding. VC usage is scalable since the maximum number of VCs required is limited to the number of routes in the routing table. Forwarding performance is exceptional within ASs as all packets will be forwarded at the cell level. Between ASs a high level of forwarding may be required, depending on the level of IP routing table aggregation. The IP navigator variant forwards 100% by the network layer at AS gateway routers. Tag switching will have a similar problem in a highly hierarchical network. The table linked forwarding nature of MPLS allows rapid adaptation to changes in the IP routing table.

The DSLS hybrid label switching approach uses data driven label switching between ASs with flows based on the destination-site address to reduce VC usage and improve forwarding performance (This will be investigated in more detail in Chapter 4). A control driven approach similar to MPLS is used within ASs to eliminate the network layer forwarding requirement. DSLS uses table-linked forwarding within ASs, and is able to take advantage of future OSPF routing protocol extensions such as OSPF-OMP. Aggregated flow linked packet threshold forwarding is used between ASs so the inter-AS path will be independent of IP route aggregation. The aggregated flow based forwarding between ASs will be insensitive to route changes and work must be done to examine the performance in conjunction with dynamic congestion sensitive routing protocols.

3.8 Conclusions

The classification presented in this chapter highlights the major functional differences between label switching protocols. In addition this work leads to several areas of interest which are studied in the remainder of this thesis.
• In this chapter the relationship between routing table aggregation and table driven label switching protocols such as MPLS is discussed. We suggest that increasing routing table aggregation will shorten MPLS label switched paths and lead to a higher network layer forwarding requirement. The introduction of IPv6 will increase routing table aggregation further due to the strict provider based hierarchical address allocation. An investigation into the performance of label switching protocols in an IPv6 environment is deemed an open issue. In Chapter 4 we examine this issue in detail. In addition we develop and examine the performance of a hybrid label switching protocol, called DSLS, that takes advantage of additional information provided in an IPv6 environment.

• Flow based label switching protocols, such as IP Switching and MPOA, do not link forwarding directly to IP routing table information. The performance of these protocols in networks where routing tables change frequently (possibly due to congestion or loading information) has not been examined in literature thus far. This issue is examined in Chapter 5. The response of flow based protocols to routing table changes is examined. In addition we examine the effect of route aggregation on flow lengths and the response to routing changes. This study is of additional interest due to the inter autonomous system component of the proposed DSLS protocol which uses highly aggregated flow based label switching.

• MPLS relies upon VC merge for scalable operation in a cell based network. Use of VC merge entails a substantially more complicated packet forward mechanism than simple cell forwarding. Studies have shown the VC merge mechanism does not add significantly to the buffer requirements (Widjaja and Elwalid, 1999). However, this is not necessarily the case in an optically switched environment where buffer space is limited and output buffer reduction techniques may be used. An examination of VC merge in optical cell switches is considered an open question and is examined in detail in Chapter 6.
Chapter 4

Label Switching in Hierarchical Networks

4.1 Introduction

Label switching combines the flexibility and robustness of IP routing with ATM’s label swapping forwarding paradigm. Routers are under strain due to an increase in the size of internal routing tables resulting from the dramatic increase in the size of the Internet over the last few years (Bass, 1997), and the need for higher packet forwarding rates. Label switching protocols such as MPLS (Viswanathan et al., 1998) and IP Switching (Newman et al., 1998) attack this problem by reducing the number of datagrams that need to be routed by the network layer. Forwarding decisions are ‘cached’ in fixed length labels that are carried with the data. In the case of cell based label switching, this fixed length label is the VCI field in the ATM cell. Some approaches (e.g. Tag Switching) do not restrict themselves to ATM, they are designed to run over all types of data link layer. This thesis concentrates primarily on label switching over cell based networks.

The use of label switching reduces the load on traditional IP forwarding techniques. This does not, however, address all of IP’s scalability problems. In order to improve the scalability and stability of IP it is necessary to reduce the size of IP route tables, this is done by aggregating routes. Instead of core routers containing routes to all destination Autonomous Systems (AS), routes need to be aggregated to less specific
destinations representing groups of ASs. This is being done with the introduction of Border Gateway Protocol version 4 (BGP4), Classless Inter-Domain Routing (CIDR) and the future introduction of IP Version 6 (IPv6). IP route aggregation is on the rise because of the advances in IP scalability and will increase further with the introduction of IPv6. It is important to consider how this effects label switching protocols.

In this chapter we examine the use of label switching in a wide area environment where packets are forwarded on a label switched path within Autonomous Systems as well as between them. In particular we examine the effect of IP routing table aggregation on label switching proposals. We predict that high levels of route aggregation will cause significant problems for Tag Switching and MPLS. This is due to the link between labels and routing tables used by both these approaches. Label switching protocols such as IP Switching do not link labels to routing tables, however they require approximately 15% (Newman et al., 1998; Boustead et al., 1998a) of packets forwarded at the network layer in each label switch and have large flow-state storage requirements.

We propose a hybrid label switching solution called Destination-Site Label Switching (DSLS), which was briefly introduced in Chapter 3, that will switch a high percentage of packets on label switched paths independent of IP routing table aggregation. The proposed protocol uses components of both packet threshold destination based label switching as well as MPLS style routing table linked forwarding. Data driven flows are defined on the AS path (inter-AS path) and routing table linked flows are defined within ASs (intra-AS path). The inter-AS flows use an innovative method of aggregation that takes advantage of the strict provider based nature of IPv6 address allocation, and aggregates traffic based on the IPv6 address hierarchy. Aggregating traffic based on address hierarchy enables a high level of aggregation of label switched paths which is independent of any underlying routing table aggregation. The intra-AS flows take advantage of MPLS style routing table linked flows. In order to use both these label switching techniques in one protocol we use two levels of labels.

This chapter examines the performance of DSLS using traffic traces to determine the performance in terms of the percentage of packets that follow the cut-through path,
the amount of state storage space required, as well as the connection set-up rates. In order to examine the performance of DSLS a methodology has to be determined to allow trace traffic to aggregated based on the IPv6 destination site address. In initial experiments we use the old class based address structure to estimate destination site. However it is unclear how accurate this method of aggregation is. In order to get a better idea of the performance of DSLS we developed a novel approach of mapping subnetting information over a packet level trace. The subnetting, or hierarchy, information is obtained from Internet address registries, routing arbiter databases, and route servers. This information is then mapped on top of the packet level trace before anonymisation of traces (renumbering of IP addresses for privacy reasons). We use this methodology to collect hierarchical traffic traces from a key backbone node in a major Australian Internet service provider’s network.

We use the hierarchical trace to examine the performance of DSLS in a wide area Internet environment. Our simulation studies show that use of the proposed protocol reduces the packet forwarding requirement to below 0.15% in AS gateway routers with 100% of packets being switched within ASs. In addition to an examination of DSLS we use the trace to examine aggregated packet threshold protocols in general. We also examine the effect of flow aggregation and variation of packet threshold parameters on performance.

This chapter will first introduce examples of table linked label switching protocols. Hierarchical routing and IP version 6 are discussed in Section 4.3. Section 4.4 examines the impact of hierarchical routing on label switching. The new concept of destination-site based label switching (DSLS) is introduced in Section 4.5. Section 4.6 examines the performance of DSLS assuming a class based address structure. A mechanism for examining DSLS and aggregated packet threshold forwarding without the assumption of a class based address structure is presented in Section 4.7. Detailed results of the simulation study are presented in Section 4.8. Section 4.10 concludes this study.
4.2 Label Switching Approaches

In order to examine the effect of IP routing table aggregation on label switching protocols we classify them into two groups, those that link forwarding decisions to routing table entries and those that do not. This section examines both these types of label switching protocols. We show that linking forwarding to routing table entries results in a situation where network layer forwarding is sensitive to IP route aggregation, particularly in routers at the edge of autonomous systems.

4.2.1 Table Linked Label Forwarding

The concept of table linked forwarding was described in Chapter 3. In addition the protocols that use table linked forwarding (IP Navigator (Ahmed et al., 1997), ARIS (Woundy et al., 1997), MPLS (Viswanathan et al., 1998) and Tag Switching (Rekhter et al., 1997)) were described in detail in Chapter 2. In this section we examine the use of label switched forwarding on the autonomous system path. An autonomous system path follows a route across the Internet determined by gateway routing protocols such as the Border Gateway Protocol (BGP). Some label switching protocols such as IP Navigator depend upon OSPF and do not create label switched paths based on BGP information. Tag switching and MPLS are label switching techniques that use table linked forwarding based on BGP information.

MPLS and Tag Switching link label switched paths to routing table entries. This is done by maintaining an IP Forwarding Information Base (FIB). A FIB is populated using information from routing protocol messages, and is similar to the routing table in a standard IP router. A label is assigned to each FIB entry and a label switched path is negotiated with adjacent switches as discussed in Section 2.3.6 and Section 2.3.9. The first hop tag switch performs network layer forwarding to find the correct entry in the FIB as well as the associated label. The label is placed in the VCI or VPI entry in the cells, and the datagram is then forwarded through the ATM switch using this label. Subsequent label switches will have pre-setup bindings between this tag and a tag for its next hop router to the destination. Thus the datagram will be switched towards the destination.
Binding of tags to routing table does not eliminate the need for network layer forwarding (Viswanathan et al., 1998). IP forwarding must be performed at the entry into a MPLS or TAG switching network to untagged datagrams. Additional IP forwarding is required since routing tables do not necessarily maintain a route to all possible destinations due to aggregation within IP routing tables.

We will now examine the extra network layer forwarding required due to IP routing table aggregation when label switching on an autonomous system path. In order to explain the effect of route aggregation we provide an example network in Figure 4.1, and we assume sources in AS1 and AS2 are transmitting packets to AS5 and AS6. If we assume that there is no route aggregation in the BGP tables then BGP routers A to G will contain routes to 120.99.130.0/24 and 120.100.5.0/24 and associated label switched paths. Packets being forwarded from AS1 or AS2 will follow a label switched path to AS5 and AS6. If routers A, B, C and D do not contain routes to 120.99.130.0/24 and 120.100.5.0/24 but contain an aggregated route to 120.0.0.0/8. Then packets from AS1 and AS2 will follow the same label switched path to router E. At router E network layer forwarding is required to split packets from the one label switched path to the next hop. In the non-aggregated case network layer forwarding is required at gateway routers for customer networks AS5 and AS6 and no network layer forwarding is required in the core of the network. When flows are aggregated then additional network layer forwarding is required in the core network (AS4) as well as the customer networks AS5 and AS6. This example illustrates how IP route aggregation can increase network layer forwarding requirements in the core network.

Both Smart IP Switching (Lloyd and O'Mahony, 1998) and Flow Aggregated Traffic Driven Label Mapping (FATDLM) (Nagami et al., 1999) also link forwarding to routing prefixes. However, these protocols only create the cut-through path when traffic flows on the route. These protocols will have significantly lower VC usage. However since they still link cut-through routes to routing table entries they will still be affected by IP routing table aggregation.

Tag switching and MPLS support multiple labels. These multiple labels are termed a "label stack" (Viswanathan et al., 1998). Labels can be pushed on the top of the stack, and can be popped off the stack. It is suggested that the label stack be used
in hierarchical networks to remove the need for network layer forwarding at de-aggregation points (Viswanathan et al., 1998) (such as the edge of AS4 when the routes are aggregated in Figure 4.1). However the number of labels in cell based label switching is limited to two. This was discussed in more detail in Section 3.4.

### 4.2.2 Other Label Switching Protocols

Ipsilon's IP Switching (Newman et al., 1998), Cell Switch Router (Katsube et al., 1997) (CSR), IP on ATM (Parulkar et al., 1995), and MPOA (Fredette, 1997) are all examples of label switching techniques that do not link cut-through forwarding to routing table entries. Because these protocols do not link forwarding directly to routing table entries they are not affected directly by routing table aggregation.

Studies examining the performance of IP Switching (Newman et al., 1996) show that core Internet routers will switch approximately 85% of packets on label switched paths. A study on the performance of MPOA (Widjaja and Elwalid, 1999) indicates that between 70% and 85% (depending upon the available VC space) of packets will
be forwarded on ATM cut-through. These protocols will perform the same at gateway routers and routers internal to autonomous systems.

### 4.3 Hierarchical Networks

A major motivating factor driving the development of label switching protocols was a desire to simplify IP forwarding allowing it to scale more easily to the higher forwarding capabilities required. Label switching uses a much simplified direct lookup instead of a longest-prefix lookup on a large FIB required by traditional IP forwarding. This overcomes part of the problem that is caused by large IP routing tables but does not attempt to solve other problems related to large IP routing table sizes, such as route stability, and processing requirements for route calculation. Therefore it is still necessary to develop scalable IP routing protocols. Most proposals to develop scalable IP routing use a hierarchical routing structure and IP route aggregation.

Prior to the introduction of the current version of the Border Gateway Protocol (BGP version 4) and Classless Inter-Domain Routing (CIDR) the Internet was viewed as an unstructured and non-hierarchical interconnection of routing domains or Autonomous Systems (AS) (Bass, 1997). In the early 90’s with the Internet growing rapidly and router table sizes doubling approximately every 9 months (Rekhter and Li, 1996) the unstructured nature of the Internet was becoming a problem. IP routers were running out of table space, and "grinding to a halt processing routing updates" (Bass, 1997). With the introduction of BGP4 and CIDR provider-based IP address allocation with variable bit subnet masking, it became possible to reduce the size of routing tables by aggregating routes. However, CIDR and BGP4 have some problems. CIDR does not require domains to renumber if they change to a different provider (Fuller et al., 1993). This has the potential to significantly reduce gains in routing table aggregation over time. In addition, the IPv4 address space is still becoming exhausted, just at a slower rate. The need for a more hierarchical network with smaller route tables and more address space has partially motivated development of IPv6.
4.3.1 IP Version 6

IPv6 uses a 128-bit address to identify interfaces, not hosts as in IPv4. Several types of addresses are supported. This chapter will concentrate on the use the best-effort unicast address called "aggregatable global unicast address" (Hinden and Deering, 1998). This address format was designed to "facilitate scalable Internet Routing" (Hinden et al., 1998). The address format has a fixed structure as shown in Figure 4.2 and is organised into a three level hierarchy: Public Topology; Site Topology; and Interface Identifier. Provider based address allocation is mandatory, non-provider based address such as AS2 in Figure 4.1 are not allowed. This means that when customers change service providers they must change their network address. This results in strict hierarchical provider based address structure which will allow high levels of IP route aggregation.

The public topology consists of a two level hierarchy of service providers with a Top-Level Aggregation Identifier (TLA ID) and a Next-Level Aggregation Identifier (NLA ID). The TLA ID is initially to be restricted to 13 bits which translates to 8192 routers in the core IPv6 network. This was done to constrain core routing table sizes. However, there is provision for expansion by 8 bits if necessary (Hinden et al., 1998). The NLA ID is 24 bits long and allows for a flat or hierarchical allocation of the NLA address space. The Site-Level Aggregation Identifier (SLA ID) is 16 bits long. The SLA ID is used by an individual organisation to define its local address hierarchy,
and subnets.

The strict provider based address allocation used by IPv6 means that all the customers of a particular service provider must be a subset of the providers address space. For example if an Internet service provider (ISP) A's IPv6 address is:

```
1e:10:fe:00:00:00:00:00/48
```

Then the following customer’s site address must be provided by ISP A:

```
1e:10:fe:01:00:00:00:00/64
```

If this customer changes service providers then they must also change their network number (in this case the first 48 bits of their address). Provider based addressing allows routes to all networks provided by a service provider to be aggregated to the service providers address. This allows high levels of aggregation.

### 4.4 Impact of Hierarchical Routing on Label Switching

The level of aggregation of IP routes will effect the performance of table linked routing protocols such as MPLS, Tag switching and FATDLM. At the end of aggregated IP routes all packets must be routed or a hierarchical label stack is required. The level of route-aggregation will therefore impact greatly on the performance of such protocols. A non-hierarchical network would result in very few packets being routed in the core of the network, but non-hierarchical routing algorithms have inherent scaling problems related to route stability (Hinden et al., 1998), router processing required for large tables, and processing time required to perform IP forwarding. This means that a future Internet needs to have high levels of route aggregation. IPv6 has a strict three level hierarchy for unicast traffic with a top level of only 8192 routers. It is suggested in (Hinden et al., 1998) that a table size of around 8000 entries should be maintained in the IPv6 Internet core to optimise performance of IP routing protocols. With this level of aggregation, a high percentage of packets would need to be routed by table-linked protocols at the boundary between the top level of the hierarchy (TLA) and the next level (NLA).
Table linked label switching protocols perform well within an autonomous system, and are capable of switching 100% of packets on label switched paths. However the performance of such protocols between autonomous systems for an Internet wide path will depend upon the level of IP route aggregation. This will particularly be the case when IP version 6 is implemented with it's strict provider based, hierarchical address structure. The remainder of this chapter examines our approach which is designed to perform well in this environment by using table linked label switching within ASs and flow linked label switching between ASs.

DSLS (as shown in Figure 4.3) uses both control driven label assignment, and data driven label assignment. Two levels of labels are used, and stored in the ATM cell’s VPI and VCI fields. Table linked forwarding is used within ASs to setup VP paths from ingress to egress nodes. Flow linked forwarding is used for VC paths to the destination using information from the IPv6 header. The purpose of this approach is to maintain the ability of many nodes to switch all packets (as in Tag switching), while eliminating the effect of IP route aggregation on performance by not assigning labels based on routing table entries for inter-AS traffic.
Flow linked label assignment is applied at egress routers of ASs. A label is negotiated between ingress and egress nodes across VP pipes. A destination based label is assigned, based on the IPv6 destination site network. Using the aggregatable global unicast address format with fixed hierarchical address fields, it will be possible to define flows based on IPv6 destination-site address. Cut-through routes will be created using a packet threshold technique similar to that used by IP Switching. Due to statistical multiplexing of traffic destined for the same routing domain we predict that a high percentage of packets will be switched through egress routers of the core ASs to the gateway router of the destination routing domain. An alternate form of flow linked inter-AS forwarding is also investigated in this chapter that aggregates packets between source-site and destination-site address. This mode of operation is advantageous since it does not require VC merge capability, however, it will require more VCs and will switch less packets on label switched paths.

Within ASs table-linked forwarding is used to create VP pipes from ingress to egress nodes. An example of this can be seen in Figure 4.3. Routers internal to transit ASs forward all transit packets at the data-link layer. Only routers at the edges of transit ASs will be required to forward a small percentage of packets at the network layer. These VP pipes are destination based VPs from multiple ingress nodes to the one egress node. Using destination based VPs removes the need for hardware VC merge internal to ASs.

Unlike other data driven label switching techniques such as IP Switching, destination site label switching will not allow a single cut-through from source to destination. It is necessary to route all packets in the ingress router at the destination site. This is not a significant disadvantage since most large organisations maintain IP firewalls around their sites and packets need to be forwarded by IP at this point irrespective of which label switching technique is used.

In summary DSLS uses a combination of table-linked forwarding as well as labels linked to the IPv6 address-hierarchy. Hierarchical routing aims to reduce the size of inter-AS routing protocols, and will greatly effect AS gateway routers. Routing table sizes within ASs are restrained by the size of the administrative domain and thus do not require high levels of route aggregation (however in the case of a particularly
large administrative domain it can be split into smaller segments such as OSPF routing areas to reduce routing table sizes as discussed in Section 2.3.7). DSLS takes advantage of the strict provider based IPv6 addressing and uses address-hierarchy labels between AS gateway routers and thus overcomes the problem of IP route aggregation. Table-linked routing protocols are still used within ASs to take advantage of high packet switching rates. The next section examines the performance of inter-domain DSLS.

4.6 Preliminary Examination of DSLS

This section presents preliminary results on the performance of DSLS at AS boundary routers that we presented in (Boustead et al., 1998a). We were particularly interested in the percentage of packets forwarded on label switched paths at DSLS boundary routers.

A traffic driven discrete event simulation was used in conjunction with traffic traces from the National Laboratory for Applied Networks Research (NLANR\(^1\)). The simulation modelled the flow detection and cut-through creation in a single edge switch. The packet threshold technique was used for cut-through creation. Flows were defined by the destination network and source/destination pairs (IP Switching). The anonymised NLANR trace did not contain accurate destination subnet information, however it did contain the obsolete class based address structure. The traffic trace used was generated 28th of February 1996 which is several years after the abolition of class based addresses and the introduction of classless addressing. However, this information was used to give an indication of performance.

A summary of the results is presented in Figure 4.4. DSLS was found to use significantly less VCs (6000 VCs) than source/destination IP Switching (14000 VCs) and destination-based (9500 VCs) approaches due to a higher level of granularity. The percentage of packets forwarded by DSLS was 4% which was also significantly less than the other approaches (14.9% for IP Switching).

\(^1\)www.nlanr.net/PMA
These are only preliminary results. The accuracy of determining subnets using outdated class based information is suspect. In addition the traces used were several years old. However, it gives an indication of the performance of DSLS. The next section describes a more complete performance examination of DSLS using current subnet information.

### 4.7 Framework for Performance Evaluation

In order to examine the performance of label switching protocols in a hierarchical Internet it is necessary to obtain packet level information as well as network hierarchy information. This allows aggregation of individual source/destination flows into aggregated flows which may be defined by source and destination subnets or higher levels of aggregation. This section describes the collection of traffic traces that contain this information and the subsequent processing of these traces.
4.7.1 Hierarchical Trace Collection

Four traces were obtained from a major Australian Internet backbone. These traces contained 40 million packets each. The traces were obtained during work hours on weekdays at 11am 19/10/99, 2pm 12/11/99, 3pm 16/11/99 and 4pm 18/11/99. The trace durations were 3387, 1380, 1510 and 1608 seconds respectively. Trace bit rates vary from 65Mbps for the short trace to 11.8Mbps for the longest trace. The traces contained packet level and hierarchical address information.

For each address in the trace address hierarchy information was obtained from a hierarchical address database. The hierarchical address database contains information information from major international and local from route servers, routing arbiter databases, and Internet address registries. Local copies of these databases were obtained before collection of each trace. This information is used to add subnetting information to packet level traces.

Internet address registries maintain databases of registered sub-nets and are used to obtain more specific destination subnets than provided in route information. We collect Internet number entries with subnet sizes 4 bits or greater. An example of an Internet registry entry from the Asia-Pacific Network Information Centre (APNIC) is:

<table>
<thead>
<tr>
<th>Internet number</th>
<th>202.6.91.0 - 202.6.91.255</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>National Library of Australia</td>
</tr>
<tr>
<td>Description</td>
<td>Parkes Place</td>
</tr>
<tr>
<td>Description</td>
<td>Canberra ACT 2600</td>
</tr>
<tr>
<td>Source</td>
<td>APNIC</td>
</tr>
</tbody>
</table>

Route information is also collected for the address hierarchy. Route information gives larger aggregates, and is used to examine the relationship between aggregation levels and label switching performance. Route-servers maintain a superset of routing information required by peering service providers. We collect "route" objects from these databases. An example of a collected entry follows:
<table>
<thead>
<tr>
<th>Route</th>
<th>202.61.224.0/19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Paradox Digital non portable CIDR space</td>
</tr>
<tr>
<td>AS Number</td>
<td>AS7586</td>
</tr>
<tr>
<td>source</td>
<td>APNIC</td>
</tr>
</tbody>
</table>

Routing-Arbiter Databases contain collections of route information that is used for operational purposes by network service providers. An example of the entries collected is:

- B 170.143.0.0/16 via 198.32.176.25
- B 207.25.252.0/24 via 134.24.88.55
- B 204.179.85.0/24 via 134.24.88.55
- B 204.145.119.0/24 via 134.24.88.55

In the above database entries "B" represents a BGP routing entry. The second column is the only information we are interested in and represents aggregated addresses in core routers.

Address and route information is collected from the databases in Table 4.1 into our hierarchical database. After removal of many duplicate entries the database contains 321805 entries from the Internet address registries, and 135827 entries from the routing databases. The IP addresses in the packet level trace are anonymised by the re-numbering method whereby the first IP address in the trace is renumbered to 1, the next address is renumbered to 2 and so on. Before the addresses are anonymised the hierarchical database is searched for all matching entries, these entries are then anonymised and stored in conjunction with the packet level trace. An example of a hierarchy determined for a particular address can be seen in Figure 4.5. The first entry represents the anonymised IP source or destination address in the packet trace. The second and third entries represent an anonymised class based subnet address. Class A addresses have an 8 bit mask, Class B addresses have a 16 bit mask and Class C addresses have a 24 bit mask. The remaining columns contained subnet addresses and bitmasks starting with the smallest subnets (largest bit mask) on the left. We provide more information about the hierarchy information collected in Appendix A.
### Table 4.1 Hierarchy information

<table>
<thead>
<tr>
<th>Internet Number Registries</th>
<th>Routing Arbiter Databases</th>
<th>Route Servers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia Pacific NIC ftp.apnic.net</td>
<td>European Internet Registry ftpripe.net</td>
<td>Planet Online Europe route-server.as5388.net</td>
</tr>
<tr>
<td>Taiwan NIC ftp.apnic.net</td>
<td>Australian NIC ftp.apnic.net</td>
<td>ATT Cerfnet route-server.cerf.net</td>
</tr>
<tr>
<td>American Registry for Internet Numbers rr.level3.net/pub/rr</td>
<td>Japan NIC ftp.apnic.net</td>
<td>ATT WorldNet Service route-server.ip.att.net</td>
</tr>
<tr>
<td>Australia NIC ftp.apnic.net</td>
<td></td>
<td>Route Views Oregon Internet Exchange route-views.oregon-ix.net</td>
</tr>
</tbody>
</table>
4.7.2 Determining Site Address

The site address is determined as the most specific match in the hierarchy database that has a bit mask of less than 28 bits (this represents a 6 bit subnet). For example the most specific match for the example in Figure 4.5 is 1286\20 where 1286 is an anonymised address and 20 is the CIDR bit-mask. This site address is in most cases represents a smaller subnet than the most specific route contained in a full unaggregated routing table because of the addition of address registry information to give more specific customer sites. Levels 2 to N shown in Figure 4.5 represent aggregations higher than the site-address.

4.7.3 Simulation

A traffic driven discrete event simulation was used and was similar to that used in Section 4.6 however the Australian traces were used in conjunction with the above procedure for determining the destination site address. The simulation modelled the flow detection and cut-through creation in a single edge switch. The packet threshold technique was used for cut-through creation. Flows were defined by the destination network and source/destination pairs (IP Switching). Results were obtained for all Australian traces. Confidence intervals (95%) were obtained for the results using the method of batch means and the student distribution (Averill M. Law, 1982).
The length of the batches was set to equal at least 10 times the transient period.

4.8 Results

4.8.1 Performance of Destination-Site Label Switching

The hierarchical trace allows us to examine the performance of destination-site label switching. We examine both destination-site, and source-site/destination-site flows. These results are compared with source/destination (IP Switching) flows.

The site addresses are defined by the most specific match in the hierarchical database. A histogram of the level 1 bit-mask (see Figure 4.5 for definition of level 1) sizes for all the addresses contained in the trace is shown in Figure 4.6. This gives an indication of the destination-subnet subnet size distribution.

The performance of DSLS as well as IP Switching in terms of the percentage of
packets switched is shown in Figure 4.7. IP Switching forwards 13.5% of packets at the network layer. Aggregating traffic between source and destination sites reduces this to 8%. The highest level of aggregation shown is destination-site (this is used by DSLS) where all packets destined for a single site address are aggregated, this results in less than 2% of packets requiring network layer forwarding.

VC usage is shown Figure 4.8. The number of VCs required for source/destination flows (IP Switching) was 16500 VCs. The number of VCs required for DSLS destination-site flows was 5000 VCs which is 70% lower than IP switching. The number of VCs required for site-subnet pair flows was 14000 VCs which is only 15% lower.

The VC setup rates shown in Figure 4.9 show a more significant difference. The required VC setup rate for IP switching was 165 VCs/sec. The setup rate required for DSLS destination-site label switching was 20 VCs/sec (87% less). The VC setup rate for site-subnet pair flows was 102 VCs/sec which is 36% lower than the VC setup rate for IP switching.
Figure 4.8 Average number of VCs required

Figure 4.9 Connection setups required per second
The results for percentage of packets routed are an average of all four traces with 95% confidence intervals were generated by using the batching method. The VC results are also a combination of results from each of the traces. In order to average the VC results they were first normalised with respect to the packet-rate of each trace, the results are then scaled to the average packet rate of the traces for readability.

4.8.2 Higher Levels of Aggregation

The previous section examined use of site address to determine stream aggregation. In this section we examine the relationship between aggregation and label switching performance. In order to obtain this relationship aim aggregation levels (bitmasks) were selected between a bit-mask of 16 to 32. For each packet in the trace the hierarchy database is searched for a subnet that contains the address with a bit-mask closest to the aim bit-mask. A bit-mask size of 32 represents source-destination flows, a bit-mask size of 21 (see Figure 4.6) represents the site address as examined in the previous section. A bit-mask size of less than 21 represents flow aggregation greater than site-address.

Figure 4.10 examines the relationship between aggregation level and percentage of packets forwarded at the network layer. The x-axis in Figure 4.10 shows average bit-mask size. It can be seen that increasing aggregation to site-aggregation significantly increases performance. Further increasing aggregation particularly for destination-site flows does not significantly improve performance. These results are for the first trace, results for the remaining traces show a similar trend.

4.8.3 Parameter Adjustment

Choice of the packet threshold parameter has a significant impact on label switching performance. The level of flow aggregation is also likely to have a significant impact on the optimum choice of packet threshold parameters. This section examines the effect of aggregation on sensible packet threshold choice.

The results presented in Figure 4.10 use a packet threshold of 10 packets in 60 seconds to trigger creation of cut-through flows. This is frequently used for non-
aggregated source-destination flows (Lin and McKeown, 1997) as a compromise between high percentage of packets switched and low VC usage. Higher aggregations may have a different optimum packet threshold value. In order to examine this we use the hierarchical trace to plot a surface examining the relationship between aggregation, packet threshold and average VC usage. We generate the results as shown in Figure 4.10 over a range of packet threshold values from 0 to 50.

Figure 4.11 (a) examines the relationship between aggregation, average VC usage and packet threshold for the first trace. The aggregation level is expressed in terms of bit-mask size, a small bit-mask represents a high level of aggregation. It can be seen that at low aggregations (bit-mask of 32) VC usage is sensitive to the packet threshold parameter. Choosing a packet threshold of below 10 has a significant impact on the number of VC's used. At high aggregation levels the average number of VCs used is relatively insensitive to packet threshold choice. This indicates that a lower packet threshold choice is sensible for highly aggregated label switching.

A surface plot examining the effect of packet threshold on the percentage of packets requiring network layer forwarding is shown in Figure 4.11 (b). This graph clearly
Figure 4.11 Effect of aggregation on (a) the average VC requirement and (b) the network layer forwarding requirement
shows that at high levels of aggregation (low bitmask size) the network layer forwarding performance is also less sensitive to changes in the packet threshold value. However, a packet threshold less than 10 still results in a significant improvement.

These surface plots, while showing interesting trends, are difficult to read in terms of actual performance. We now present graphs showing different levels of aggregation over a range of packet threshold values.

The performance of data-driven aggregated label switching protocols with reduced packet threshold is shown in Figure 4.12. Reducing the packet threshold to 5 lowers the percentage of packets switched for the destination subnet case from 1.38% to 0.8% a reduction of 42% which results in 6% increase in average VC usage. A further decrease in packet threshold to 1 reduces the network layer forwarding requirement to 0.13% with minimal increase in VC usage.

If VC merge is not available then DSLS uses site-pair flows. Site-pair flows with a packet threshold of 5 requires network layer forwarding for 4.3% of packets while maintaining a VC usage significantly below that required by IP Switching.

4.9 Discussion on the Choice Packet Threshold Parameters

We show that a packet threshold of 1 is advantageous for highly aggregated flows. A packet threshold of 1 creates a cut-through path for every packet flow. However, it is important to note that this is still a data driven flow and the cut-through path will timeout after the flow timeout period (usually 60 seconds). The use of the flow timeout is important to constrain the number of cut-through paths that are active at any one time.

As discussed in Section 2.4.2 the results for flow based label switching protocols are highly dependent on traffic characteristics. The introduction of applications that result in short flows will reduce performance significantly. This will also be true for aggregated flow based label switching. However, we believe that it will have a much less significant effect for aggregated flows because the trace driven studies in this
chapter have shown that aggregation based on destination site address significantly reduced the number of short flows. Thus, the introduction of a new application that results in many new short flows will effect non-aggregated flows much more significantly than aggregated flows.

4.10 Conclusions

In order for IP to cope with future expansions of the Internet a strict hierarchical address structure and high levels of route aggregation are needed (Bass, 1997). Table

Figure 4.12 Subnet aggregation with reduced packet threshold
linked label switching approaches, such as Tag switching, will not perform well in this environment. Large numbers of packets will need to be routed at the boundaries between TLA and NLA layers of a hierarchical IPv6 network. We propose a label switching approach, referred to as DSLS, that is independent of BGP routing table aggregation and takes advantage of the strict provider based IPv6 address structure. DSLS uses table-linked forwarding within ASs and a packet threshold destination-site based label assignment at AS boundaries. Routers within ASs will switch all packets.

A methodology is developed to examine the performance of DSLS using IP version 4 traffic traces and information from Internet address registries, route servers and route arbiter databases. In addition to examining the performance of DSLS we use this methodology to examine the performance of aggregated label switching protocols in general.

We then examine the relationship between aggregation and label switching performance. We show that destination-site aggregation significantly improves performance. The relationship between aggregation and the packet threshold parameter is also examined. We show that a packet threshold of below 5 significantly reduces the network layer forwarding requirement with only a small penalty in terms of extra VCs required. We show that DSLS will forward less than 0.15% of packets at the network layer in gateway routers.
Chapter 5

Flow Based Forwarding and Frequently Changing Routing Tables

5.1 Introduction

Label switching protocols such as MPLS (Callon et al., 1997), MPOA (Fredette, 1997), IP Switching (Newman et al., 1998), Cell Switch Relay (CSR) (Katsube et al., 1997), Destination Site Label Switching (DSLS), IP Navigator (Ahmed et al., 1997) and Tag Switching (Rekhter et al., 1997) are all designed to provide IP services over an ATM network. However, these protocols differ substantially in the way in which they route and forward packets. This chapter examines the relationship between forwarding and routing decisions in these protocols. Forwarding is not always explicitly linked directly to routing decisions. This chapter investigates label switching protocols used in conjunction with adaptive routing protocols that react to congestion or other network conditions and lead to dynamically changing routing protocols.

We group label switching protocols into two categories: Flow Based Cut-through (FBC) forwarding, and Table-Linked Cut-through (TLC) forwarding. FBC forwarding makes the the routing decision when the cut-through flow is created and subsequent packets in the flow will follow the same path. FBC protocols include IP Switching (Newman et al., 1998), Cell Switch Relay (CSR) (Katsube et al., 1997), MPOA (Fredette, 1997) and Destination Site Label Switching (DSLS) on the inter-AS path. TLC paths are linked directly to routing protocol information. These cut-
through paths are created and modified upon receipt of routing protocol messages. TLC capable protocols include MPLS (Callon et al., 1997), IP Navigator (Ahmed et al., 1997) and Tag Switching (Rekhter et al., 1997). These two groups of protocols will perform differently in situations where the underlying routing protocol changes frequently.

Currently popular IP routing protocols (including OSPF and BGP) are insensitive to changes in network conditions such as congestion, and only provide shortest path routes to destinations based on fixed weightings defined at each hop. A common problem with these routing protocols is congestion in shortest path to the destination while longer alternate routes may be under utilised (Aukia et al., 2000). This is essentially a traffic engineering problem (Aukia et al., 2000). The use of overlay networks is a popular way of attacking this problem (Wang and Zheng, 1999). An alternate solution to this problem is the use of congestion sensitive multi-path routing (Widjaja and Anwar, 1999; Villamizar, 1998). Currently such a protocol is under development by the OSPF working group of the IETF. This approach is called OSPF Optimised Multipath (OSPF-OMP) applies dynamic routing for best effort services to allow routing of datagrams on multiple paths around heavily loaded links (Villamizar, 1998).

In Chapter 4 we proposed using aggregated flow based label switching on the AS path. An obvious disadvantage of FBC forwarding is that packet forwarding is not directly linked to routing information. This situation was discussed in more detail in Section 3.4. The introduction of congestion sensitive multipath routing will produce a situation where the lowest cost route to a particular destination may change frequently depending on network load. This chapter examines the effect of frequently changing underlying routes on FBC label switching performance.

We investigate the performance of FBC forwarding mechanisms with backbone traffic traces. The time response to a single route change in the underlying routing protocol is examined first. The characteristics of the traces are then examined to explain the poor performance of FBC forwarding. Finally, we examine the effect of changing FBC flow detection parameters and the inclusion of a maximum flow length to improve the time response while minimising the number of cells forwarded
Flow Based Forwarding and Frequently Changing Routing Tables

at the IP layer. The above methodology is applied to two case studies. The first study examines historical traces from a major U.S.A. backbone node (Federal Internet Exchange West - FIX-West) between 1995 and late 1997. The second case study examines several traces obtained in 1999 from a major Australian backbone node. This second study also examines the effect of aggregation on the performance of FBC forwarding.

We find that FBC performs even worse than expected from published flow length distributions (Claffy et al., 1995). In the worst case example considered for non-aggregated flow based forwarding it took over 1200 seconds to start forwarding 50% of the packets on the new route after a route change was initiated. For aggregated flows the performance was significantly worse. We showed in particular that the response of destination-subnet flows is unsuited to congestion sensitive multipath routing. We find that implementing a maximum flow length does not result in a significant reduction in the number of switched cells while dramatically improving the response time.

Section 5.2 describes FBC forwarding and TLC forwarding. Section 5.3 discusses current advances in dynamic routing protocols. We examine flow studies to predict the sensitivity of flow based forwarding in Section 5.4. The trace driven experiments are described in Section 5.5. Results comparing FBC and TLC forwarding are presented in Section 5.6. Section 5.7 presents the conclusions of this chapter.

5.2 Cut-through Forwarding Classification

This section briefly describes flow and table based label switching.

5.2.1 Flow Based Cut-through Forwarding

Flow based cut-through forwarding (FBC) abstracts the forwarding mechanism from routing table information. When a cut-through flow is created the routing decision is made for that flow, a fixed cell-level cut-through path is created and all subsequent cells in that flow follow this fixed route. Flow based cut-through protocols include:
Flow Based Forwarding and Frequently Changing Routing Tables

IP Switching, Cell Switch Relay (CSR), MPOA and Destination Site Label Switching (DSLS) on the inter-AS path. In this section we briefly examine the most popular FBC protocols (IP Switching and MPOA) as well as the flow based forwarding component of our Destination Site Label Switching (DSLS) proposal.

MPOA (Fredette, 1997) uses data driven techniques to determine if a cut-through flow should be created. The IP address of the source and destination MPOA Client (MPC) is used to define flows. The decision to create a cut-through between the egress to the MPOA network and destination MPC is based on a packet threshold parameter. The first packets in a flow, before the threshold is exceeded, are forwarded hop-by-hop by IP routers between Emulated LANs to the destination. The ingress node of the MPOA network determines if an end-to-end cut-through is necessary. If the number of datagrams in the flow exceeds a threshold value within a certain time period a cut-through is created. All subsequent packets follow this flow.

IP Switching also uses a packet threshold to determine if a cut-through path should be created. When the threshold is exceeded a cut-through path is created between itself and an adjacent IP Switch. When this happens in all IP Switches along the path an end to end cut-through is created. As with MPOA, once the cut-through VC is created all subsequent packets on that flow follow this fixed route independent of underlying routing table changes.

DSLS, as described in detail in Section 4.5, consists of a flow linked component as well as a table linked component. The flow based forwarding component of DSLS creates label switched paths between gateway routers belonging to different autonomous systems. The mechanism to create label switched paths is similar to IP switching, however, the flow definitions are significantly more aggregated. Flows are defined using information from the IPv6 address structure, specifically, the destination and possibly source site addresses. Two types of flows definitions are defined: destination site flows if VC merge is available; and source/destination site pairs if VC merge is not available. These flow definitions are more highly aggregated than IP switching and MPOA flows. Since flows are more highly aggregated it is expected that DSLS will respond significantly slower to underlying routing table changes.
5.2.2 Table-linked Cut-through Forwarding

Table linked cut-through forwarding links cell level cut-through paths directly to routing table information. Examples of protocols which support TLC forwarding are MPLS, Tag Switching and IP Navigator. This mechanism is described in more detail in Section 3.4.2.

5.3 Dynamic Routing

This section examines scenarios where underlying routing protocols may change frequently due to changing network conditions. The aim of congestion sensitive routing is the efficient utilisation of network resources. Currently popular routing protocols such as OSPF use fixed metrics for the determination of shortest paths. This can result in poor overall network utilisation. Packets transiting from ingress to egress nodes of a network will follow the shortest path even if it is highly congested and alternate lower cost routes are underutilised. Congestion sensitive routing protocols have been under investigation since the development of the ARPANET routing protocol. We first examine the ARPANET implementation and the problems that were encountered. We then discuss other multipath protocols and proposed protocols that address this issue such as Cisco’s IGRP, an extension to OSPF called Optimised Multipath (OSPF-OMP) and MPLS Adaptive Traffic Engineering (MATE). In this chapter we do not propose to analyse the performance of or extend on these protocols. This chapter intends to examine the performance of flow based label switching forwarding in terms of its speed of reaction to routing table changes.

5.3.0.1 ARPANET Delay Sensitive Routing

The first delay sensitive routing protocol was designed and implemented for use in the ARPANET in 1969. At its largest the ARPANET consisted of several hundred packet switches (Streenstrup, 1995) which were called IMPs (Interface Message Processors).

The initial ARPANET routing algorithm was based upon a distributed Bellman Ford
algorithm. The shortest path was chosen by summing link metrics on each path to the destination. This metric was calculated by adding a fixed constant to the instantaneous queue length. Routing information was periodically exchanged between neighbours every 2-3 seconds (Khanna and Zinky, 1989). This protocols produced routes that were unstable, and far from optimal. This was blamed upon the instantaneous calculation of delay (Khanna and Zinky, 1989). Some of these issues were addressed in the next major update which was called ARPANET SPF (shortest path first).

The ARPANET SPF algorithm (McQuillan et al., 1980) was implemented in 1979. The metric used to calculate the shortest path was based on delay averaged over 10 seconds. The delay measurement consisted of queueing, transmission and propagation delays (Streenstrup, 1995). Changes in delay triggered flooding of link state advertisements to other IMPs, the minimum time between these link state advertisements was 10 seconds. This updating of link delay information enabled routing protocols to route packets around congested nodes thus improving network utilisation. Over the lifetime of ARPANET SPF other changes were proposed and implemented (Glazer and Tropper, 1990; Khanna and Zinky, 1989) in attempts to overcome stability problems.

5.3.0.2 Interior Gateway Routing Protocol (IGRP)

Cisco's IGRP is a proprietary form of the RIP (Routing Information Protocol) distance vector routing algorithm (Streenstrup, 1995). The additional features provided by IGRP that are of most interest to this discussion are: a composite link metric that allows multiple link costs including delay; and the ability to split traffic over multiple routes.

The composite link metric can be seen in Equation 5.1. C is the the minimum bandwidth of all links in the route. D is the route delay which can be comprised of switching delay, transmission delay and propagation delay. The reliability of the route is represented by R. The relative importance of bandwidth and delay on the calculation of the link metric is determined by the parameters \( k_1 \) and \( k_2 \). A study in (Low and Varaiya, 1993) examines the effect of the parameters \( k_1 \) (traffic insensitive weighting)
and $k_2$ (traffic sensitive weighting) on convergence. It was shown that if the traffic sensitive component is given a high weighting then convergence is not achieved and routes "will oscillate between two worst cases" (Low and Varaiya, 1993). This indicates that a finer granularity of congestion based changes will improve stability.

\[
\frac{k_1}{C} + \frac{k_2}{D}/R \tag{5.1}
\]

Many routing protocols such as OSPF and RIP allow multipath forwarding (or load balancing) on equal cost routes to destinations. IGRP allows multipath forwarding on similar cost routes. A route of cost (C) is deemed to have a similar cost to the shortest path cost (P) if $C < vP$ where $v$ is called the variance parameter (usually set to 1 (Streenstrup, 1995)). The analytical study in (Low and Varaiya, 1993) shows that the use of multipath forwarding has a significant stabilising effect on the performance of IGRP. Multipath forwarding is the main aim of OSPF Optimised Multipath (OSPF-OMP) that will be discussed next.

5.3.0.3 OSPF-OMP

The IETF OSPF working group is working on an extension to OSPF to allow best effort traffic to be dynamically routed around congested links. This proposal, called OSPF Optimised Multipath (OSPF-OMP) (Villamizar, 1998), works by propagating information about load levels in links across the Autonomous System. The flooding of this information is facilitated by use of OSPF's Opaque Link State Advertisement option (Opaque LSA) (Coltun, 1998). The link state information includes: link load, packet loss, and link capacity. Using this information each router can use the internal link state topology map to determine the lowest-cost routes to the destination according to current network conditions and the utilisation of links within the network. OSPF-OMP also supports multipath forwarding on unequal paths and divides the traffic between these paths according to known traffic loads. A source/destination hash approach is used to determine which packets will follow each path, this ensures that packets in the same source destination flow will follow the same path. OSPF-OMP insures stability by providing an algorithm that allows gradual adjustments while allowing fast adjustments when required (Villamizar, 1998).
Another new approach to ensuring effective usage of network resources by using delay sensitive multipath forwarding is MPLS Adaptive Traffic Engineering (MATE) (Widjaja and Anwar, 1999). This approach uses the tunnelling capability of MPLS to enable the setup of multiple label switched paths between the same ingress and egress routers of an autonomous system. The main aim of MATE is distribute packets arriving at the ingress node on these multiple label switched paths to reduce congestion.

M label switched paths are created between every ingress/egress node pair as shown in Figure 5.1. These paths represent the best M paths calculated by link state routing protocols. Once these paths have been established MATE attempts to distribute load on these paths so that congestion is minimised. Traffic can be divided between paths using the same source/destination IP address hash as used by OSPF-OMP (Villamizar, 1998).

The update period proposed for MATE is 5 minutes. This large update interval was chosen because a study (Thompson et al., 1997) showed that Internet traffic is quasi-stationary over at least 5 minutes.
In summary, this section examined the major protocols and proposals that allow congestion sensitive routing. This examination highlighted several desirable features of such protocols to enable stability: the availability of multipath forwarding capability (particularly on non-equal cost paths); and avoidance of large step changes improves stability.

5.3.1 Label Switching and Dynamic Routing

This section examines the issues related to the use of flow-linked switching in conjunction with multipath congestion sensitive routing protocols such as OSPF-OMP. In particular we are interested in allowing multipath forwarding while maintaining packets on a label switched path. We are also interested in the response of flow based forwarding to changes in the underlying path, and changes in the percentage of packets forwarded on the multiple paths.

Consider the example network shown in Figure 5.2. In order to use the host pair hash method of splitting traffic between path a and path b it is necessary to examine IP addresses when forwarding each packet - this implies network layer forwarding. If we wish to preserve the label switched paths and still divide packets between the multiple paths it is necessary to perform the multipath forwarding decision at the start of flows, and forward packets on that path for the duration of the flow. However flows are not always defined on source and destination address. We propose an alternate method of splitting traffic between multiple paths. Instead of using host pairs we propose to use label switching flows which may be defined many ways including host pairs and destination site address (used by DSLS). However it is unclear how quickly such an approach will adapt to changes particularly if a destination-site flow aggregation is used.

Another important factor to consider is flow aggregation. In the previous section it was stressed that the stability of congestion sensitive routing was a fine granularity of forwarding adjustments. This facilitated in MATE and OSPF-OMP by dividing traffic between paths using source-destination pairs. The use of destination-site aggregated flows significantly increases this adjustment granularity.
The minimum time between congestion based routing advertisements gives an indication of how quickly flow based forwarding label switching protocols must be able to adapt to changes in routing tables. In the early ARPANET the minimum time between changes was 2-3 seconds (Khanna and Zinky, 1989). This was increased to 10 seconds in (Khanna and Zinky, 1989) for stability reasons. Later protocols such as IGRP suggest a minimum adjustment period of 90 seconds, while MATE suggests 5 minutes. These larger adjustment periods were also chosen for stability reasons.

5.4 Previous Studies of Aggregated Flows

An overview of the previous studies on flows and flow aggregation is given in Section 2.4.2. This section will concentrate on (Claffy et al., 1995) which presents an examination of data driven flows based on several flow granularities. Several packet level traffic traces were examined with localities ranging from backbone to a laboratory Ethernet. This study does not consider label switching, however, the examination of source-pair flows and aggregated flows is of interest to label switching. The key results we are interested in are packet threshold and flow timeout requirements for aggregated and non-aggregated flows. The flow length distributions presented in (Claffy et al., 1995) are also useful to enable us to predict sensitivity to underlying route table changes.
In (Claffy et al., 1995) packet level traces traces are analysed for multiple aggregation levels:

- host pairs (hp),
- destination host (dh),
- destination network (dn),
- network pairs (np), and
- source Host (sh).

Except for the source-host flows all of these aggregation levels are interesting for the study of label switching. The aggregation part of the study examined one backbone trace which contained traffic injected into the NSFNET from the Urbana-Champaign FDDI ring. Traffic flowing in the opposite direction was filtered out. The trace was collected on the 23rd of March 1993. This trace collection date was before deployment of BGP-4 and Classless Inter-Domain Routing (CIDR), this indicates that the destination network addresses were determined by using class-based addresses structure.

Flow length distributions for the different aggregation levels are shown in (Claffy et al., 1995). For non-aggregated host-pair flows 50% of flows are shorter than 2 seconds while over 99% of flows are less than 1000 seconds in length. The highly aggregated destination network case consists of 50% of flows shorter than 11 seconds with approximately 97% of flows less than 1000 seconds in length.

If we examine the flow length distributions provided by (Claffy et al., 1995) in terms of the IGRP 90 second minimum update time we find that 90% of host-pair flows are shorter than 90 seconds and 74% of destination subnet flows are less than 90 seconds. Assuming a uniform average bit-rate for all flows this indicates that host-pair flows will react to a route change and forward 90% of packets on the new route within 90 seconds for the aggregated case the percentage of packets is reduced to 75%. However it is unlikely that small flows will have the same average packet size as long flows so a more thorough investigation is required. It is also indicated in (Claffy et al., 1995) that these results are likely to change significantly as new
Internet applications, and traffic types are introduced.

5.5 Methodology

In order to examine the sensitivity of FBC forwarding mechanisms backbone traffic traces along with flow detection code were used. Of most interest was the response to changes in the underlying routing table. Examining the performance of TLC is trivial since a change in routing protocols will by definition change the forwarding decision. The purpose of these tests is to quantify how long it takes FBC forwarding to adapt to routing table changes, and to identify possible ways to improve the response.

Two sets of traces were obtained. The dates, duration and size of these traces is shown in Table 5.1. The first set of traces are from the western USA Federal Internet Exchange (FIX-West) over a 3 year period. These traces were obtained from the National Laboratory for Applied Network Research (NLANR) website\(^1\). Since flow based forwarding is dependent on traffic characteristics, it is useful to examine traffic over a range of years to look for any trends. Four additional traces were obtained from a major Australian telecommunication provider’s backbone in late 1999. These traces contained additional post-CIDR subnetting information and were used to examine the performance of aggregated flow based forwarding. The trace collection method and details of the calculation of destination site addresses described in Chapter 4.

Figure 5.3 shows how the traces were processed. We collected trace characteristics of the whole trace, then passed it through an FBC forwarding mechanism which forwards the packet on route 1 or route 2 depending upon the routing table. The "routing table" is set to forward all packets on route 1 until a route change trigger is received after which the routing table changed to forward all packets on route 2.

The FBC forwarding mechanism uses a packet threshold for flow detection in order to end flows due in inactivity. Optimal flow timeout values for host pair flows is shown to be between 16 and 128 seconds in (Claffy et al., 1995) as a tradeoff between

\(^1\)www.nlanr.net
Table 5.1 Traffic traces used

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Date of Trace</th>
<th>Duration</th>
<th>Number of packets</th>
</tr>
</thead>
<tbody>
<tr>
<td>19950621.1500UT</td>
<td>Fix West</td>
<td>21/6/1995</td>
<td>1625</td>
<td>132202033</td>
</tr>
<tr>
<td>19960228.2145UT</td>
<td>Fix West</td>
<td>28/2/1996</td>
<td>770</td>
<td>79583186</td>
</tr>
<tr>
<td>19970109.1751UT</td>
<td>Fix West</td>
<td>9/1/1997</td>
<td>1084</td>
<td>7302564</td>
</tr>
<tr>
<td>19971120.1945UT</td>
<td>Fix West</td>
<td>20/11/1997</td>
<td>1154</td>
<td>400000000</td>
</tr>
<tr>
<td>Australian Trace 1</td>
<td>Australian Backbone</td>
<td>19/10/1999</td>
<td>3387</td>
<td>40000000</td>
</tr>
<tr>
<td>Australian Trace 2</td>
<td>Australian Backbone</td>
<td>12/11/1999</td>
<td>1380</td>
<td>40000000</td>
</tr>
<tr>
<td>Australian Trace 3</td>
<td>Australian Backbone</td>
<td>16/11/1999</td>
<td>1510</td>
<td>40000000</td>
</tr>
<tr>
<td>Australian Trace 4</td>
<td>Australian Backbone</td>
<td>18/11/1999</td>
<td>1608</td>
<td>40000000</td>
</tr>
</tbody>
</table>

Figure 5.3 Traffic trace analysis
Flow Based Forwarding and Frequently Changing Routing Tables

memory and processing requirements. In addition (Lin and McKeown, 1997) and (Newman et al., 1996) both cite 60 second flow timeouts as a reasonable choice for data driven label switching. We used a flow timeout of 60 seconds for the initial examination, however, we varied this in later tests to examine the effect on sensitivity to routing table changes. A maximum flow length was also implemented whereby flows are ended if the flow-time exceeded x seconds even if they are active. This facility was only used in the last experiment described in each case study.

5.6 Traffic Analysis

This section presents the results of our examination of flow based forwarding in an environment where the underlying routing table changes frequently. The results are divided into two case studies. The first case study examines historical fix-west traces obtained between June 1995 and November 1997. The second case study uses the Australian traces and examines the effect of aggregation on route change sensitivity. In each case study the time response due to a single route change is examined. The trace characteristics are then examined to explain the results for FBC forwarding. Finally flow detection parameters are altered to determine the effect on the results observed.

5.6.1 Case Study 1: Historical Examination of Fix West 1995-1997

The time response to a single route change is shown in Figure 5.4 (a) for the four different backbone traffic traces. All packets are forwarded on route 1 (see Figure 5.3) for the first 100 seconds. After 100 seconds the routing table is changed to route 2. The results are significantly worse than our preliminary examination using distributions from (Claffy et al., 1995), as discussed in Section 5.5. In the best case (June 1995 trace) 50% of packets are forwarded on the new route (route 2) 150 seconds after the route change while 85% is reached after 1350 seconds. The November 1997 trace did not even reach 50% of packets on route 2 when the trace ended 1200 seconds after the route change.
Figure 5.4 Time response to a single route change (a) over whole trace duration and (b) for the first 60 seconds after the route change.
Flow Based Forwarding and Frequently Changing Routing Tables

Figure 5.4 (b) magnifies the first 90 seconds (90 seconds is the minimum IGRP update period). The oldest three traces (June 1995 to January 1997) produce similar results with approximately 45% of packets being forwarded on route 2 after 90 seconds. The latest trace forwarded only 24% of packets on the new route after 90 seconds.

In order to explain these results it is necessary to examine the trace characteristics. Of interest is the relationship between the length of flows and the average size of packets within these flows.

Figure 5.5 shows the characteristics of the traces in more detail. Figure 5.5 (a) to (d) compares the percentage of packets on route 2 after the route change with the percentage of total flows for each trace. A significant difference can be seen between the percentage of flows and the percentage of packets in each case, particularly the later traces. Flow length statistics are shown in Figure 5.5 (e) to (h). The dashed lines in Figures 5.5 (e) to (h) show the distribution of flow lengths. The histogram in same graphs show the percentage of packets contributed by all of the flows whose length are within the range of each histogram bin. These graphs show the contribution of different length flows to the total number of packets and flows. We can see that each trace has a large percentage of short flows which contribute a significant proportion of packets, and a very small proportion of large flows that also contribute significantly to the total percentage of packets. This characteristic gets more exaggerated in the later traces particularly the 20/11/97 trace where only 8% of packets exceed 750 seconds, however, this small percentage of flows contributes over 57% of total packets. In the best case the 21/06/95 trace 4% of flows are longer than 750 seconds and contribute 23% of packets.

These results indicate that traffic characteristics at the FIX-West Internet exchange are changing over time. The ratio of short to long flows has remained similar. However, the long flows (greater than 750 seconds) are contributing a higher percentage of the total packets in the later traces. This explains the deviation of our sensitivity results from those expected from (Claffy et al., 1995) which used traces from 1993. These results also indicate that flow based label switching protocols, which make routing decisions only at the start of a flow, will react slowly to underlying routing
Figure 5.5 Flow characteristics
Flow Based Forwarding and Frequently Changing Routing Tables

Table changes in current traces. The next section will examine more up to date traces collected from a different point in the Internet. The remainder of this section will examine methods of reducing flow lengths without reducing significantly the performance of label switching protocols in terms of the percentage of packets switched.

We first examine changing the flow timeout parameter. The second method we examine takes advantage of the results presented in Figure 5.5 which shows that a small percentage of long flows contributes significantly to the total number of packets. If we introduce a maximum flow length this will ensure a significantly faster response to a route change while interrupting only a small percentage of the flows.

5.6.1.1 Altering Flow Detection Parameters

In the previous section we show that the poor sensitivity of FBC forwarding is linked to a small number of long flows that contribute a significant proportion of total cells. In this section we examine possible ways of improving the sensitivity of FBC forwarding to route table changes. We examine two possible alternatives: reducing the flow timeout value to below 60 seconds; and fixing an upper bound on the flow length.

A flow timeout of 60 seconds is commonly used for FBC forwarding (Newman et al., 1998). If the flow timeout is reduced then the number of long flows will also be reduced improving sensitivity to route changes. However, this will also have the undesirable side-effect of increasing the percentage of packets forwarded at the network layer. We examine the effect of reducing the flow timeout value by repeating the experiment in the previous section with flow timeouts of between 1 and 60 seconds. The results of this experiment can be seen in Figure 5.6 (a) which shows the percentage of packets on link 2 versus flow timeout 200 seconds after the route change. In the worst case (trace: 20/11/97) the effect of reducing the flow timeout from 100 seconds to 10 seconds is an increase in percentage of packets from 20% to 40% on the new route 200 seconds after the route change. The other traces experience a similar magnitude improvement. The penalty of this performance improvement is a reduction in the number of packets switched. This reduction can be seen in Figure 5.6 (b) and becomes quite significant at flow timeouts below 10 seconds.
Flow Based Forwarding and Frequently Changing Routing Tables

Figure 5.6 Vary flow timeout
We now investigate our proposal of limiting the maximum length of flows to improve the sensitivity of FBC forwarding to route table changes. Using the flow length characteristics shown in Figure 5.5 we predict that implementing a Maximum Flow Length (MFL) will substantially improve sensitivity while resulting in a minimal reduction in the percentage of packet switched.

In the following experiments we examine the performance when the MFL is introduced. As with the previous experiments all flows are forwarded on route 1 at the start of the trace. An MFL is applied to all flows. When this maximum flow length is reached the cell based cut-through created for the flow is ended, and re-created, if necessary, according to the normal flow detection criteria. At 200 seconds from the end of the trace we change the routing table and start forwarding new flows on route 2 to examine the sensitivity of MFL flows to a route change.

Figures 5.7 (a) and (b) show the real-time response to a single route change with an MFL of 100 seconds and 200 seconds respectively. With an MFL of 100 seconds 50% of packets are forwarded on route 2 after 10-30 seconds. With an MFL of 200 seconds 50% of packets are forwarded on route 2 after 10-105 seconds. After 90 seconds 95% of packets are forwarded on route 2 for a MFL of 100. The results for the November 1997 trace in Figure 5.7 (b) are problematic because the MFL is large compared to the total length of the trace and the simulation does not get a chance to settle down to equilibrium. The longer Australian traces shown in the next section are more indicative of performance with MFL=200.

The cost of the introduction of an MFL is a reduction in the percentage of packets forwarded on cut-through routes leading to a corresponding increase in the number of packets forwarded at the network layer, this can be seen in Figure 5.8. With an MFL of 100 the cost is a 2%-3.4% reduction of packets on cut-through routes. An MFL of 200 leads to a significantly lower penalty of between 1% and 1.4%.

### 5.6.2 Case Study 2: Hierarchical Aggregation

This section examines the effect of aggregation on forwarding. We use the hierarchical Australian traces described in Chapter 4 and examine the sensitivity of site pair,
Figure 5.7 Vary maximum flow length
destination-site, and destination address flows to changes in routing tables. We use the same methodology as used in the previous section to examine the traces.

The time response to a single route change for the four Australian traces is shown in Figure 5.9. Three of the four traces exhibit similar performance. The shortest trace shows an unexplained step change at 1350 seconds, and is not used to generate results. In the remainder of this chapter we use trace 1 because it is significantly longer than the other traces.

As expected aggregation significantly effects sensitivity to underlying routing table changes. This can be seen in Figure 5.10 (a) where 50% of packets are forwarded on the new route 80 seconds after the route change for source-pair flows this increases to 200 seconds for site-pair flows and 580 seconds for destination based flows. Destination subnet flows do not reach 50% of packets flowing on the new route during the 3500 second trace duration.

Figure 5.10 (b) shows the response curve for the first 90 seconds (minimum IGRP update time). After 90 seconds the percentage of packets flowing on link 2 is 38% for source-pairs, 34% for subnet pairs, 30% for destination address and 9% for destina-
Flow Based Forwarding and Frequently Changing Routing Tables

Figure 5.9 Time response to route changes for each Australian trace with no aggregation

tion subnet. Of most interest is source-pairs (MPOA and IP Switching), destination-subnet (DSLS with VC merge), and subnet-pairs (DSLS without VC merge). Clearly the response of destination-subnet flows is unsuited to congestion sensitive multipath routing even for the 5 minute update intervals used by MATE. We now investigate the effect of implementing maximum flow lengths on aggregated flows to improve the performance.

The effect of introducing an MFL can be seen in Figure 5.11 (a) for MFL=100 and Figure 5.11 (b) for MFL=200. With MFL set to 100 at least 50% of packets are forwarded on the new route within 30 seconds from the route change for all levels of aggregation except for destination site. Using an MFL of 200 increases the time to forward at least 50% of packets on the new route to 40 seconds. The percentage of packets forwarded on the new route 90 seconds after the route change is 90% for MFL=100 and 70% for MFL=200 for all levels of aggregation except destination site.

The results for destination site shown in Figure 5.11 (a) and (b) are slightly erratic when compared to the other levels of aggregation. This is caused by the length of
Figure 5.10 Time response to route changes for different levels of aggregation (a) over whole trace and (b) after 60 seconds.
Figure 5.11 Maximum Flow Length (a) MFL=100 (b) MFL=200
destination site flows in relationship to the trace length. We believe with longer traces the destination site flows will give similar results to the other levels of aggregation.

The network layer forwarding requirement for flow based forwarding with MFL=200 is shown in Figure 5.12. These results can be compared with the results with no MFL in Chapter 4 to determine the network layer forwarding penalty. This graph is reproduced in Figure 5.13 for convenience. We also examine the packet threshold parameter. The packet threshold will have a significant impact on the forwarding penalty because it determined the number of packets that will require network layer forwarding before a label switch path is re-established after the MFL is reached. In Chapter 4 we determine that use of a low packet threshold is beneficial for destination site flows. However many studies have shown that a packet threshold of 10 (Newman et al., 1998) for source pair flows is a reasonable choice.

For source pair flows (with a packet threshold of 10) using an MFL of 200 increases the network layer forwarding requirement from 13.7% (see Figure 5.13) to 15.5% (see Figure 5.12) which is a 13% increase. For the highest aggregated flows (also with a packet threshold of 10) defined by destination site address the forwarding re-
requirement increases from 1.2% to 2% with an MFL of 200, this represents a 66% increase. As discussed above a lower packet threshold may be used for highly aggregated flows with little VC usage penalties. At a packet threshold of 5 destination site flows require network layer forwarding for 1% (compared to 0.8% with no MFL) of packets with an MFL of 200. Using a packet threshold of 1 the network layer forwarding penalty of using MFL=200 increases the network layer forwarding requirement from 0.11% to 0.13% of packets which is an increase of 18%.

5.7 Conclusions

In this chapter we described a series of experiments designed to investigate the performance of flow-based forwarding used in conjunction with dynamic routing protocols. We propose using label switching flows to split traffic between multiple paths. In this section we examine how sensitive this approach is to changes in underlying routing tables. We examine two different case studies. The first case examines historical traces from FIX-West from 1995 to late 1997. The second case study examines the effect of flow aggregation on performance. This study concentrates on the performance of DSLS destination-site flows.
Firstly, each study examined time response curves depicting the reaction to a single change in the lowest-cost route using several traffic traces. Flow-based cut-through forwarding was shown to react significantly slower than expected, by examination of published flow length distributions (Claffy et al., 1995), to changes in routing tables. The time lapse varies between 140 and over 1200 seconds (depending on trace used) before more than 50% of flow-based packets are forwarded on the new route after a route change. This problem was seen to get progressively worse between June 1995 and November 1997 indicating a possible trend in changing traffic characteristics. The reason for this unexpectedly slow reaction to route changes was shown to be caused by a small percentage (between 4% and 8%) of long flows that contribute significantly to the total packet rate (18-57%). The second part of this chapter examined the effect of aggregation on the performance of flow based forwarding. Aggregation was shown to significantly effect performance. We showed in particular that the response of destination-subnet flows is unsuited to congestion sensitive multipath routing even for the 5 minute update intervals used by MATE.

Flow detection parameters for flow based cut-through forwarding were examined to determine their effect on the sensitivity to route changes. We found that reducing the flow timeout parameter improved performance slightly if a flow timeout of 10 seconds or less was selected. However, this also results in a significant reduction in the percentage of packets forwarded over a cell level cut-through and therefore an increase in the network layer forwarding requirement. The introduction of a maximum flow length was shown to result in a significant improvement in the sensitivity to route changes for all the traces from 1995 to 1999. Selecting a maximum flow length of 200 seconds for source-pair (IP Switching and MPLS) increases the number of packets requiring network layer forwarding by 2% (of total flows) in the worst case examined while improving the response to routing table changes. The percentage of packets forwarded on the new route 90 seconds after the route change (90 seconds is the update interval for IGRP) was 95%. Destination-site flows showed significant improvement with the addition of an MFL of 200. Approximately 90% of packets follow the new route 90 seconds after the route change with a small increase in the network layer forwarding requirement.
Chapter 6

Packet Forwarding in Optically Switched Networks

6.1 Introduction

Optical transmission capacity is increasing dramatically with the introduction of Wavelength Division Multiplexing (WDM). In order for future switching architectures to keep up with projected transmission capacities there has been much interest in the development of optical switching technology. The synchronous time-slotted nature of many proposed optical switching architectures means that a cell based datalink layer is likely. It is therefore important to consider packet over optical cell switch forwarding alternatives.

The current protocols for IP over cell in the electrical domain include Multi-Protocol Label Switching (MPLS) (Callon et al., 1997), IP Switching (Newman et al., 1998) and Multi-Protocol Over ATM (MPOA) (Fredette, 1997). Protocols such as MPOA and MPLS were designed and optimised for use with traditional electrical switching/routing technology. The nature of optical switching hardware necessitates different design criteria for an IP over cell protocol. It is particularly important to reduce complexity, and to minimise the need for buffering. This chapter investigates the possible implementation of these protocols over optical cell switches.

We examine the forwarding mechanisms of several IP over cell approaches including
MPOA, MPLS and IP Switching. Of particular interest is the amount of buffering required per port for VC Merge, packet reassembly and output buffers. In existing electronic switches the additional buffer required for VC merge has been shown to be dwarfed by the size of output buffers that are required for switches operated at high utilisation (Widjaja and Elwalid, 1999). In the optical domain where buffering is scarce the size of the required output buffers may not dwarf reassembly buffers, especially when techniques are used to minimise output buffer requirements. We use analytical and simulation techniques to model the output and reassembly buffers. Two scenarios are examined that minimise output buffer usage. Firstly we examine operating switches at low utilisation. The second scenario examines use of a traffic smoothing methodology proposed in (Moser and Melliar-Smith, 1996). In each of these cases we find that reassembly buffers become a significant component of buffer requirements. It is concluded that label-switching protocols that do not use VC merge require significantly fewer cells of optical buffer. A backbone traffic trace driven simulation is then used to verify the results with non-synthetic traffic.

Label switching techniques that do not require VC merge such as IP Switching generally use a packet threshold to determine when to create a cut-through. Before the packet threshold is met packets on a particular flow must be forwarded by the network layer at each hop on a common VC. We examine two approaches to enable network layer forwarding: use of an adjunct electronic router; and reassembly in the optical switch. The first approach forwards the first few packets of a flow in the adjunct electrical router, when a flow is established the subsequent packets are forwarded by the optical switch. The second approach assumes that the first few packets in a flow are reassembled for packet forwarding within the optical switch. This packet forwarding may take the form of MPLS style table-linked label switched forwarding which requires packet reassembly, or a simplified network layer forwarding mechanism. For the second case we find that the use of aggregated packet forwarding protocols such as MPLS requires on average twice the number of cells for reassembly buffers than non-aggregated protocols such as IP Switching. This is an extension of our work presented in (Boustead and Chicharo, 2000d).

In the next section two broad categories of optical buffering are described, their
suitability for VC merge is then discussed. Section 6.3 describes packet over cell protocols, and classifies them into two groups: aggregated forwarding, and non-aggregated forwarding. The factors effecting buffer usage are described in Section 6.4. This section also describes mechanisms to reduce output buffer requirements. The methodology behind our performance comparison is described in Section 6.5, this includes analytical and simulation model descriptions. Section 6.6 compares buffer usage of aggregated versus non-aggregated protocols. Network layer packet forwarding options are described in Section 6.7. Section 6.8 concludes.

6.2 Optical Switching

Optical switching is seen as a way of enabling switching to keep pace with the dramatic increases in optical transmission capacity, particularly with the introduction of WDM and DWDM. This section briefly introduces technologies that will allow optical networking including all optical wavelength routing (Ramamurthy and Mukherjee, 1998), and optical cell/packet switching. This thesis concentrates on cell switching, therefore, optical cell switches and buffers are described in more detail. However, it is important to discuss how cell switching and wavelength routing technologies fit into the network architecture.

6.2.1 Wavelength Routing

WDM is increasing transmission capacities with the capability to transmit on multiple wavelengths (or channels) on a single fiber simultaneously. The number of parallel channels demonstrated varies between 2 and 32, with recent announcements of up to 100 channels (Elmirghani and Mouftah, 2000b). With the success of WDM transmission there has been much effort applying WDM technology and devices to networking rather than just transmission (Elmirghani and Mouftah, 2000b). The technology that holds most promise for networking is wavelength routing, and wavelength converters. The use of these devices will allow the creation of wavelength switched paths which forward packets on an all optical path bypassing conversion to electrical signals, packet/cell forwarding and IP processing.
A wavelength routed network forwards data based on the wavelength of the channel (Ramamurthy and Mukherjee, 1998). Wavelength routing is particularly useful in order to create a different logical topology than the actual topology. Protocol layers higher than the physical layer only see the logical topology they will not be aware of the actual topology. An example of this concept is shown in Figure 6.1. The physical topology shown in Figure 6.1 (a) does not contain connections between A and C. Without wavelength routing packets travelling between A and C must be transmitted to B, converted to the electrical domain, and be switched to C. If wavelength switching is used with the wavelengths configured as shown in Figure 6.1 (a) then packets travelling on the wavelength switched path from A to C remain in the optical domain along the path and require no electronic processing at B. The network layer at A sees a direct connection to C, in addition the network layer and data link layer at B are unaware of the transit traffic from A to C. It is important to note that with simple wavelength switching the same wavelength must be used for the entire wavelength switched path.

The use of wavelength converters (Elmirghani and Mouftah, 2000a) in conjunction with wavelength routing makes wavelength switching more flexible. Wavelength converters can convert the data on an incoming wavelength to a different wavelength on the output port. The function of a wavelength converter in a WDM network can be seen in Figure 6.2. The simple wavelength switching approach is shown in Figure 6.2
(a) where the same wavelength must be used for the entire wavelength switched path. However switch D in Figure 6.2 (b) contains a wavelength converter so a different wavelength can be used on each segment of the path. It is interesting to note the similarity of the wavelength paths shown in Figure 6.2 (b) and MPLS label switched paths, however, instead of labels being swapped wavelengths are swapped. Because of this similarity members of the MPLS working group are examining the use of MPLS over wavelength routed networks with wavelength converters, this approach is called Multi-Protocol Lambda Switching (MPλS) (Awduche et al., 2000b).

The availability of wavelength conversion and wavelength routing switches does not remove the need for packet switches. Packet switches and wavelength routing switches will coexist (Danielsen et al., 1998a). In order to aggregate traffic onto a particular wavelength packet switching is needed. Packet switching is also necessary to deaggregate flows of packets on a wavelength. It is likely that packet switching will reside at the edge of autonomous systems with wavelength switching within the core (Danielsen et al., 1998a).

A MPλS (Awduche et al., 2000b) network consists of Label Switch Routers (LSRs) and optical cross-connects (OXC). The OXC is defined as an optical path switching element that connects optical channels (or wavelengths) on an incoming port to optical channels on an outgoing port. The LSRs are traditional electronic label
switches. The basic idea of MP\text{\&}S is to use the MPLS mechanisms to setup wavelength switched paths within the core of a network and use electronic LSRs at the edge of the network to forward packets on the correct wavelength switched path across the network. MP\text{\&}S is currently in the early stages of development and has many open issues and is an interesting area for future work.

6.2.2 Slotted and Unslotted Optical Switches

Optical cross connects, and wavelength routed networks are already widely demonstrated (Danielsen et al., 1998a) and are an accepted part of a future backbone network architecture (Yoshimura et al., 1999). Optical packet switches are much further away from commercial reality, however, they still have an important role in future optical networks replacing the electronic packet/cell switches at the edge of the future wavelength routed core networks (Danielsen et al., 1998a).

Optical switches can be described as slotted or unslotted (Yao and Dixit, 2000). A slotted optical switch operates on fixed size cells. Cell arrival is synchronous and fiber delay lines are generally used for contention resolution (Yao and Dixit, 2000). These types of switches have been studied extensively in literature (Yao and Dixit, 2000). Slotted networks have been popular because the slotted nature of the switch means that the chance of contention is less than in unslotted switches (Yao and Dixit, 2000). We examine these switches in more detail in Section 6.2.3. Unslotted switches (Danielsen et al., 1998b) can operate on packets of different sizes, and packet arrival does not have to be aligned. The absence of alignment circuitry simplifies switch design and results in a cheaper and more robust switch (Yao and Dixit, 2000).

We are interested in cell switching so we examine slotted switches in more detail in the next section. Slotted switching is assumed in the remainder of this chapter. A cell size of 53 bytes (same as ATM cell) is assumed. The choice of a 1500 byte cell would reduce the maximum transmission utilisation from 88% (Use of 53 byte cells and RFC1144 header compression (Armitage and Adams, 1995)) to 20% assuming an average Internet packet size of 300 bytes.
6.2.3 Optical Cell Switch and Buffer Designs

There are many proposed optical cell switch designs. This section briefly introduces several single stage and multiple stage optical switches. Two broad categories of buffers are considered: feed-forward and feedback (Hunter et al., 1998b). Cells entering a feed-forward buffer pass through a fixed number of optical delay lines. Feedback buffers have the capability of feeding cells back through delay lines multiple times.

A simplified block diagram of a feed forward switch design from (Masetti et al., 1996) is shown in Figure 6.3. When the cell arrives at the switch, the header is decoded. As the header is being decoded the data portion of the cell is stored optically in a delay loop for a fixed period. The electronic control determines which output port the cell is destined for. The cell is then converted to a wavelength by a tunable wavelength converter depending upon which output port the cell is destined for. The cell is then switched by electronically controlled fast optical switches to the appropriate delay line as determined by the electronic control. The delay is chosen to simulate an output buffered switch. The buffer delay lines contain wavelength division multiplexed (WDM) cells destined for different outputs. Upon exiting the delay line cells enter a wavelength de-multiplexer which routes cells to output ports depending upon wavelength.
An example of an optical switch that uses the feedback optical buffering concept is the fiber loop switch (Hunter et al., 1998b) in Figure 6.4. The switch buffering consists of a single cell period loop of fiber. Utilising WDM the capacity of the buffer is $m$, where $m$ is the maximum wavelengths available. When a cell enters the switch the header is converted to the electrical domain and used by the electronic control circuit that co-ordinates the optical switching and buffering. The optical data component of the cell is converted to a spare wavelength or "memory location" and enters the fiber loop. The electronic control maintains the cells in the loop for a time equal to a traditional output buffered switch. The optical data component is then switched to the appropriate output switch. There are other proposed feedback buffered switches such as the Shared Memory Optical Packet switch (SMOP) described in (Karol, 1993). The later approach uses delay loops of different lengths.

When a cell enters a switch with feed-forward buffers the exact buffering time must be known in order to place the cell in the correct delay line. The feedback buffer has the advantage of being able to re-circulate the cells, and may allow the reassembly required by label switching approaches that use VC-merge. However, it is important to determine how much extra buffering is required by VC merge. The next section will examine label switching approaches and discuss buffer usage.
6.3 Packet Forwarding Techniques

Current packet over cell forwarding techniques are designed to improve the scalability of electronic switch routers. IP forwarding is bypassed, for a large percentage of packets, by dynamically created cell switched paths. There are several different protocols that have been developed to do this including MPOA, MPLS, and IP Switching. These protocols have substantially different mechanisms for creating these cell switched paths. This section examines these mechanisms. Of major interest is the amount of buffering required.

Reassembly buffers are required by packet forwarding mechanisms for two purposes: IP forwarding, and to ensure ATM Adaptation Layer 5 (AAL5) cell sequence integrity. In Section 3.4.1 we group packet forwarding protocols into two groups, non-aggregated and aggregated, depending upon the need for cell stream merging. This concept is important to this chapter and is discussed in more detail below. Diagrams from Section 3.4.1 explaining both concepts are reproduced in Figure 6.5 and Figure 6.6 for convenience. Non-aggregated approaches, include IP Switching and MPOA, and require reassembly buffers only for IP forwarding. Aggregated approaches such as MPLS and Tag Switching require reassembly buffers for VC merge to maintain AAL5 cell sequence integrity.

6.3.1 Non-Aggregated Forwarding

Non aggregated forwarding techniques maintain separate virtual circuits (VCs) for source-destination pairs as shown in Figure 6.5. Packet reassembly is only required for the packets that are forwarded at the network layer. Examples include IP Switching, and MPOA.

IP Switching (Newman et al., 1998) and MPOA (Fredette, 1997) determine if a cut-through flow should be created based on the level of traffic flow. The main difference between them is where the decision to cut-through is made. The ingress node of the MPOA network decides if an end-to-end cut-through is necessary. On the other hand, every switch in an IP switching network is involved in creating its local segment of the cut-through route.
IP Switching uses both IP source and destination addresses to define flows. IP Switching can operate in different modes that use additional information such as the type of application generating the flow. This information is obtained from the TCP/UDP header. The decision to assign a VC to a flow and to cut-through packets at the ATM level is based on a packet threshold parameter. The first datagrams in a flow are forwarded by IP hop-by-hop to the destination. If the number of datagrams in the flow exceeds the packet threshold value then a label is assigned to the flow and subsequent packets in the flow are forwarded solely by ATM. The label represents a VC between the IP switch and adjacent IP switches and is set-up by a proprietary signalling protocol.

MPOA (Fredette, 1997) uses the IP address of the source and destination MPOA Client (MPC) to define flows. The decision to create a cut-through between the source and destination MPC is based on a packet threshold technique similar to that used by IP Switching. Before the threshold is exceeded packets are forwarded hop-by-hop via IP routers between Emulated LANs to the destination. Once the decision to create a cut-through is made then a modified version of the Next Hop Resolution Protocol (NHRP) protocol is used to determine the ATM address of the destination MPC and an ATM VC is created.

These approaches make similar use of reassembly buffers. Reassembly buffers are used solely for packet level forwarding since cut-through flows are defined depending on source and destination addresses. However MPLS and other aggregated forwarding techniques require additional reassembly buffers in order to maintain sequence integrity of cells in the aggregated streams.
6.3.2 Aggregated Forwarding

Aggregated forwarding techniques merge one or more VCs from different input ports to a single VC on the output ports as shown in Figure 6.6. The merging of VCs leads to the necessity for packet reassembly at VC merge points. This is due to the common use of AAL5 packet encoding, which uses only an end of packet bit in the last cell of a segmented packet for delineation. If cells belonging to AAL5 encoded packets are interleaved then the packets can no longer be reassembled. Examples of aggregated approaches include Tag Switching, MPLS and IP Navigator.

Tag Switching (Rekhter et al., 1997) and Multiprotocol Label Switching (MPLS) (Callon et al., 1997) are both examples of aggregated label-switching techniques. Both approaches are similar in many ways and will be described together.

The trigger for the creation of a Tag Switching cell level cut-through is either the receipt of a standard IP routing protocol packet advertising a new route, or a proprietary tag (a tag is the Tag Switching term for a label) distribution protocol packet (Doolan et al., 1997). The main component of a tag-switching network is a Tag Switch. A Tag Switch maintains a Forwarding Information Base (FIB), and a Tag Information Base (TIB). The FIB is populated using information from routing protocol messages, and is similar to the routing table in a standard IP router. The TIB is essentially the switch VC table. Tag Switches bind all entries in the FIB with tags in the TIB. The first hop Tag Switch performs network layer forwarding to find the correct entry in the FIB. The associated tag in the TIB will then be placed in the cell’s VPI field, and the datagram is forwarded through the ATM switch using this tag. Subsequent Tag
Switches will have previously set-up bindings between this tag and a tag for its next hop router to the destination. Thus, ATM will switch the datagram to its destination.

6.4 Factors Effecting Buffer Usage

6.4.1 Optical VC Merge

In order to perform VC merge, as required by aggregated protocols, a feedback optical buffer must be used. A feed-forward buffer can only be used if the buffer exit time can be calculated when the cell enters the switch. We envisage using electronic control of optical fiber loop buffers to simulate VC merge buffers. This would be performed in a similar way to which the output buffer is simulated using the central fiber loop buffer and electronic control. This electronic control to "simulate" an output buffer was described in more detail in Section 6.2. The MPLS label is encoded in the optical header. Only one label is available, it does not make sense to use label stacks.

If VC-merge is performed in the optical domain then the most important metrics to measure are related to buffer usage. It is important to reduce the size of buffers, and to reduce the time that cells spend in buffers.

6.4.2 Cell Inter-arrival Time

In a simple best-effort VC merge network all cells will be transmitted back to back within a packet. This study examines the effect of the introduction of cell intervals within a packet. These cell intervals may be introduced by cell level manipulation of traffic within switches, such as preferential queueing or traffic smoothing which is discussed in the next section.

 Preferential queueing mechanisms such as round robin queueing will spread cells. Consider the case where the aggregated best effort MPLS traffic is given a one fifth of total bandwidth. Using round robin scheduling the best effort queue will be serviced one out of every cell service times.
6.4.3 Traffic Burstiness

In order to operate backbone switches at high utilisation the number of cells required for a low overflow probability is in the thousands per port. For example the Cisco BXM-622 broadband switch module offers buffer sizes of up to 1,024,000 cells per switch module (obtained from www.cisco.com). The size of optical buffers provided by optical switches is significantly smaller. In (Hunter et al., 1998b) switches with a optical buffer capacity 100 cells or greater are considered to have large buffers. Larger buffers can be created by cascading switches (Hunter et al., 1998b), however, we consider the use of single switches. The total buffer requirement for the packet forwarding protocols discussed consists of output and reassembly buffers. This section examines methods to reduce output buffer usage. One alternative is to operate the switch at low utilisation with existing traffic characteristics. Other approaches, including the one described in (Moser and Melliar-Smith, 1996), involve alteration of traffic characteristics to smooth aggregate streams.

The output buffer minimisation approach proposed in (Moser and Melliar-Smith, 1996) uses traffic policing in the network core to smooth traffic and reduce contention in the switch. A period $p$ is placed on the incoming flows of each port $n$. During the period $p_n$ only a fixed number of packets can be transmitted by the downstream switch. The maximum buffer required is shown to be $p - 1$ cells where $p$ is the least common multiple of $p_1 \ldots p_n$. This approach is similar to time-division multiplexing (TDM) circuit switching. Buffer usage is minimised in the optical network core. This method effectively pushes buffer requirements to the edge of the optical network where electronic switches with large buffers will be needed.

A simple example of this approach is a $p \times p$ port switch where $p$ incoming streams transmit one cell per period of length $p$ cell transmission times. The maximum buffer size required is then equal to $p - 1$ cells. This effectively spreads each packet over $p \times l$ cells, where $l$ is the packet length in cells. We examine this case later using simulation results.
6.5 Methodology

This section describes the analytical and simulation tools used to examine the performance of aggregated and non-aggregated protocols. Of most interest is output, reassembly and total buffer usage for switches where output buffers have been minimised. The aim is to determine the penalty of VC merge in this scenario.

6.5.1 Analytical Model

The analytical model used for this study consists of a reassembly buffer model and an output buffer model. The output buffer model is based on (Widjaja and Elwalid, 1999), however, it is extended to include the realistic situation where gaps are introduced between cells within packet boundaries. The analytical models in (Widjaja and Elwalid, 1999) assumed back to back cells.

6.5.1.1 Reassembly Buffer Model

The Markov chain used to model the reassembly buffer with deterministic cell inter-arrival times is shown in Figure 6.7. The average packet size is $\frac{1}{b}$, and average off time is $\frac{1}{a}$. The top row of the Markov state diagram indicates changes in reassembly buffer level $n$ from 1 to $\infty$. In the subsequent rows the reassembly buffer stays constant. These states represent the delay between cell arrivals. This model is solved analytically. We then aggregate $N$ reassembly state maps by convolution of steady state probability vectors.

The transition matrix $A$ for the Markov chain is determined as:
Figure 6.7 Markov chain for reassembly buffer with cell interval of $g \times dt$
The steady state probabilities \( \pi \) are found to be:

\[
\pi_0 = (1 - a)\pi_0 + b \sum_{i=1}^{\infty} \pi_{i,G} \\
\pi_1 = \pi_{1,1} = \ldots = \pi_{1,G} = a\pi_0 \\
\pi_i = \pi_{i,1} = \ldots = \pi_{i,G} = (1 - b)\pi_{i-1,G} \\
= (1 - b)\pi_{i-1} \quad \forall i > 1
\]

also

\[
\pi_0 + \sum_{i=1}^{\infty} \left( \pi_i + \sum_{j=1}^{G} \pi_{i,j} \right) = \pi_0 + \sum_{i=1}^{\infty} (G + 1)\pi_i = 1
\]

so

\[
\sum_{i=1}^{\infty} \pi_i = \sum_{i=1}^{\infty} \pi_{i,G} = \frac{1 - \pi_0}{G + 1}
\]

substituting Equation 6.6 into Equation 6.2 gives

\[
\pi_0 = \frac{b}{aG + a + b}
\]

The steady state reassembly buffer depth probabilities can then be determined by summing cell arrival and wait-state probabilities:
\[ P(n) = P\{\text{state} = n\} + \sum_{i=1}^{G} P\{\text{state} = n, i\} \]  \hspace{1cm} (6.8)

so

\[ P\{n = 0\} = \pi_0 = \frac{b}{(aG + a + b)} \]  \hspace{1cm} (6.9)

\[ P\{n = 1\} = \pi_1 + \sum_{j=1}^{G} \pi_{1,j} = (G + 1)a\pi_0 \]  \hspace{1cm} (6.10)

\[ P\{n = i\} = \pi_i + \sum_{j=1}^{G} \pi_{i,j} = (G + 1)(1 - b)\pi_{i-1} \quad \text{for } i > 1 \]  \hspace{1cm} (6.11)

Utilisation can be calculated by summing the steady state probabilities of the cell arrival states and multiplying by the number of aggregated streams \( N \):

\[ \rho = N \sum_{i=1}^{\infty} \pi_i = \frac{Na}{a(G + 1) + b} \]  \hspace{1cm} (6.12)

The average off time parameter \( a \) is calculated by:

\[ a = \frac{b\rho}{N - G\rho - \rho} \]  \hspace{1cm} (6.13)

6.5.1.2 Output Buffer Model

This section examines the output process of the reassembly buffer with cell inter-arrival times. The model from (Widjaja and Elwalid, 1999) was modified to include the realistic case where gaps exist between cells belonging to the same packet.

The model for the departure process at the output buffer can be seen in Figure 6.8, with an off period of \( 1/a \), and an average packet size \( 1/b \). The time to buffer the packet in the reassembly buffer is elongated using the parameter \( G \) to model additional cell inter-arrival times. It is assumed that the packet size transmitted into the output-buffer is independent of packet reassembly. This is shown to be a reasonable assumption in (Widjaja and Elwalid, 1999). In order to solve the queue with \( N \) aggregated departure processes a new Markov model is defined using a tuple \((i, j)\) where \( i \) represents the number of chains that are in state 0 and \( j \) represents the number of chains in state 1 (Widjaja and Elwalid, 1999). The number of chains in state 2 is
obviously $N - i - j$. The solution for the output buffer is determined by using results for the D-BMAP/D/1 (Neuts, 1989). The case where there are no gaps between cells belonging to the same packet ($G = 0$) is described in (Widjaja and Elwalid, 1999). We extend this approach for the realistic case where gaps exist between cells within a packet ($G > 0$) by recalculating transition probability $p_{(i,j)}^{(i',j')}$ from state $(i', j')$ to $(i, j)$ as shown in Equation 6.21. Calculation of the steady state probabilities then follows the procedure in (Widjaja and Elwalid, 1999). Total buffer usage is obtained by convolving the reassembly buffer and output buffer steady state vectors. A more detailed description of the output buffer analysis follows:

### 6.5.1.3 Model Inputs

This section describes the inputs to the model.

\[
\text{Utilisation} = \rho \quad (6.14)
\]

\[
\text{Average Packet Size} = \frac{1}{b} \quad (6.15)
\]

\[
\text{Number of Switch Ports} = N \quad (6.16)
\]

\[
\text{Cell Separation} = G \quad (6.17)
\]

\[
a = \frac{\rho}{b(N - \rho(G + 1)) - \rho} \quad (6.18)
\]

The state transition matrix $D$ of the aggregate model is determined as:

\[
D = [p_{(i,j)}^{(i',j')}] \quad (6.19)
\]

where:

\[
p_{(i,j)}^{(i',j')} = \left( \frac{i}{N - (i' + j)} \right) a^{N-(i'+j)} \\
\cdot \left( 1 - a \right)^{i-(N-(i'+j))} \\
\cdot \left( \frac{j}{N - (i' + j')} \right) \left( \frac{b}{G + 1} \right)^{N-(i'+j')} \\
\cdot \left( 1 - \frac{b}{G + 1} \right)^{j-(N-(i'+j'))} \quad (6.20)
\]

\[
\text{(6.21)}
\]
for $N - k - i \leq i' \leq N - j$ and $N - i' - j \leq j' \leq N - i'$

The transition matrix that results in $k$ cell arrivals is $D_k$ where:

$$D = \sum_{k=0}^{\infty} D_k$$

In order to calculate $D_k$ we need to know the probability of $k$ cell arrivals when $r$ sources are in state 2. This is given by:

$$P\{A = k\} = \binom{k - 1}{r - 1} b^r (1 - b)^{k-r}$$

The procedure to calculate the steady state is now calculated in three steps.

Step 1: Determine the matrix $G$. Starting with $G(0) = 0$ we iterate:

$$G(n) = \sum_{k=0}^{K} D_k [G(n - 1)]^k$$

until

$$\max_{j,j'} |G_{jj'}(n) - \sum_{k=0}^{K} D_k [G(n)]^k_{jj'}| \leq \epsilon$$
Where $\epsilon = 10^{-10}$ is a small number.

Step 2: Determine $x(0)$

\[
\bar{D}_i = \sum_{k=0}^{K-i} D_i + kG^k \quad (6.26)
\]

\[
D_i^* = \sum_{k=1}^{K} kD_k \quad (6.27)
\]

\[
Z = D_0 + \bar{D}_1(I - \bar{D}_1)^{-1}D_0 \quad (6.28)
\]

\[
H = \bar{D}_1G \quad (6.29)
\]

so

\[
x(0) = \frac{1}{d}z \quad (6.31)
\]

where

\[
zZ = z, \, ze = 1 \quad (6.32)
\]

and

\[
d = 1 + \frac{1}{1 - \rho}z[I + (D - D_0 - H)(I - D - e\pi)^{-1}]D_1^*e \quad (6.33)
\]

Step 3: Calculate $x(i)$ using Ramaswami’s Recurrence:

\[
x(i) = [x(0)\bar{D}_i + \sum_{k=1}^{\min(i-1,i-k+1)} x(k)\bar{D}_{i-k+1}](I - \bar{D}_1)^{-1} \quad (6.34)
\]

In order to calculate total buffer usage it is assumed that the output and reassembly buffers are independent (shown to be a reasonable assumption in (Widjaja and Elwalid, 1999)). The steady state output buffer $x(i)$ is then convolved with our reassembly buffer (for $G \geq 0$) steady state distribution which is described in Section 6.5.1.

### 6.5.2 Simulation

Discrete event simulation techniques were used to compare aggregated and nonaggregated packet forwarding. The aim of the simulation comparison is to investigate
the average buffer usage for a given packet loss probability.

A block diagram of the simulation is shown Figure 6.9. A single core label switch is modelled with N input ports and a single output port. An output buffered switch was modelled since most optical switching designs simulate an output buffered switch (Hunter et al., 1998a). Traffic arriving at each input port enters the Cell Interval Modifier (CIM) block that varies the cell inter-arrival time. Reassembly buffers are placed on the output side of the switch fabric. Cells are then placed in the switch output buffer.

The simulation is fed by packet level traffic traces as well as synthetic traffic. The synthetic traffic consists of packets with exponentially distributed sizes and inter-arrival times. The traffic traces used were obtained from the National Laboratory for Applied Networks Research (NLANR) and from a major Australian telecommunication provider’s backbone. Traffic for individual input ports is obtained by dividing the traffic into N equal segments where N is the number of input ports. The packets are divided into ATM cells with the cell inter-arrival time determined by the CIM block. The CIM block inserts a cell inter-arrival time between cells within each IP packet. This cell inter-arrival time is used to simulate the switch servicing other VC’s, other priority traffic, or traffic smoothing mechanisms as described in Section 6.4.

The reassembly buffers are used to facilitate the packet reassembly required by aggregated protocols (such as MPLS) to merge AAL5 streams without interleaving cells.
belonging to different packets. We implement the output buffers and reassembly buffers as separate simulation elements. It is assumed that in a physical implementation the output buffers will be simulated by electronic control of the optical switching mechanism as described in (Hunter et al., 1998b). We also assume that the reassembly buffers will be simulated in a similar fashion utilising the same shared feedback buffer used to simulate the output buffer.

Non aggregated approaches are modelled using separate VCs for each source and destination pair. We ignore connection setup delays in order to concentrate on the effect of varying packet threshold. This also enables us to group MPOA and IP switching. Aggregated approaches are modelled requiring packet reassembly for merging streams.

We are interested in the number of cells required for reassembly buffers to forward cells with a low packet loss probability. In order to do this we measure the number of cells in the reassembly buffer as each packet is forwarded. The simulation provides an average as well as cumulative distribution of reassembly buffer size. The cumulative buffer size distribution is used to determine buffer size requirements for a given packet loss.

6.5.3 Verification of Simulation and Analysis

This section compares simulation, analytical results and published results (Widjaja and Elwalid, 1999) (in the case where G=0) the next section describes the significance of the results. Figure 6.10 shows the reassembly, output and total buffer requirements for a given overflow probability with $G = 0$, $\rho = 0.2$, and $N = 16$ for both the simulation and analytical results. These results agree well for the reassembly buffer, output buffer and combined buffers. This result also compares favourably to the same validation in (Widjaja and Elwalid, 1999). Figure 6.11 shows a comparison of simulation and analytical results for reassembly buffers with $G = 0, 5, 10, 15, 20$. Figure 6.12 shows the total buffer usage, which includes output buffer and reassembly buffer, for the same case. In each case the simulation and analytical results are equal within small tolerances.
Packet Forwarding in Optically Switched Networks

Figure 6.10 Validation of case where G=0

Figure 6.11 Reassembly buffer validation for G=0,5,10,15,20
In this section we examine the performance of packet over cell protocols in an optical environment where minimising buffer usage is critical. We examine several different scenarios. Firstly we examine buffer requirements when switch utilisation is kept low to minimise the use of output buffers. It is known that reassembly buffers are insignificant at high utilisations (Widjaja and Elwalid, 1999), our work intends to determine the range of utilisation where reassembly buffers are significant. The inter-cell gap of cells within an AAL5 packet is then examined. A large range of cell gaps is examined. Small cell gaps may be introduced by queueing mechanisms providing differentiated services, high cell gaps may be introduced by traffic smoothing mechanisms introduced to reduce output buffer usage. In particular we examine the case where \( G = N \) as per the traffic smoothing mechanism described at the end of Section 6.4.
Packet Forwarding in Optically Switched Networks

6.6.1 Effect of Low Switch Utilisation

Figure 6.13 shows the relationship between buffer usage and switch utilisation for VC merge and non-VC merge switch architectures for both the traffic model described in Section 6.5.1 as well as the NLANR traffic trace. For both cases at low utilisation VC merge uses significantly more buffer cells. At higher utilisations the difference is significantly reduced.

The percent increase in buffering required by VC merge is shown in Figure 6.14. Although the absolute buffer values shown in Figure 6.13 are different for artificial and real traffic the percentage of additional cells of buffer required for VC merge matches closely. We arbitrarily deem an additional buffer requirement of over 50% to be significant. With "back to back" cells (cell gap = 0) use of VC merge requires significant additional buffering for utilisations below 23% (approximated from Figure 6.14). With a cell gap of one VC merge requires significant additional buffering at utilisations below 43%. It is clear that if switches operate at low utilisations then
VC merge adds significantly to buffer usage, especially if the cells are not "back to back". The effect of cell gap will be examined in more detail in the next section.

### 6.6.2 Effect of Cell Gap on Buffer Requirements

In the previous section we show that the introduction of gaps between cells has a major impact on the penalty of VC merge at low switch utilisations. This section will examine a range of cell gaps. In particular we are interested in the case where $N = G$ which should result in significantly smoother traffic, and lower output buffer requirements even at high utilisation.

Figure 6.15 examines the total buffer usage (including both output and reassembly buffers) for VC merge with the inter cell gap varied from 0 to 20 cell service times, and with switch utilisation varied from 10% to 80%. For utilisations at and below 60% increasing the cell gap has a significant impact on the total buffer usage. However at utilisations over 60% increasing cell gap causes a slight reduction (at 70%) and a dramatic improvement at 80%. This dramatic improvement was shown in
Packet Forwarding in Optically Switched Networks

Cell inter-arrival time $G$ (multiple of cell service time)

Figure 6.15 Total buffer requirement for VC merge versus utilisation

(Widjaja and Elwalid, 1999) and was used to dismiss the problem of inter-cell gaps particularly at high utilisations. The cause for the improvement at high utilisation was stated in (Widjaja and Elwalid, 1999) to be smoothing of the traffic. We believe that this is likely to have an equal or possibly greater effect on non-VC merge traffic. The effect of cell gap on non-VC merge traffic was not examined in (Widjaja and Elwalid, 1999). This issue is examined later in this chapter.

The VC merge reassembly buffer requirement is shown in Figure 6.16 for varying switch utilisation and cell gap. The effect of an increase in cell gap is an increase in the reassembly buffer requirement. At 10% utilisation the reassembly buffer requirement increases from 95 to 175 cells with an increase in cell gap of 0 to 16 cell service times, this represents an increase of 84%. At higher utilisations an increase in cell gap has a more significant effect. At 80% utilisation the buffer requirement increases from 140 to 340 cells which represents a 143% increase.

Figure 6.17 shows the VC merge output buffer requirements. Overall the output buffer requirements reduce as cell gap increases. Again the effect is more significant
Figure 6.16 Reassembly buffer requirement for VC merge versus utilisation

at high utilisations. Increasing the cell gap from 0 to 16 cell service intervals reduces buffer requirement by 1% at 10% utilisation and 57% at 80% utilisation.

Comparing the total, reassembly, and output buffer requirements for the VC merge approach shows that at low cell gaps the output buffer is the most significant contributor to total buffer needs. At high cell gaps the reassembly buffer is the most significant contributer to total buffer requirements.

The non-VC merge case is shown in Figure 6.18. Increasing cell inter arrival times reduces buffer usage dramatically. This is due to a smoothing of the traffic as in the VC-merge case, however, the reduction in buffer usage is more significant particularly at high cell gaps. The superior performance of the non VC merge case is a result of maintaining cell separation. The VC merge case reassembles packets before the output buffer. A cell gap equal to the number of input ports \( G = N = 16 \) represents the case where flows are smoothed to reduce output buffer usage as in (Moser and Melliar-Smith, 1996). This reduces the buffer requirement to less than \( N \) as described in Section 6.4.3.
Figure 6.17 Output buffer requirement for VC merge versus utilisation

Figure 6.18 Total buffer requirement for non-VC merge versus utilisation
The ratio of VC merge and non-VC merge buffer usage is shown in Figure 6.19. With no cell-gaps within packets VC merge introduces no significant penalty except for low utilisations (at 10% utilisation ratio of VC merge to non-VC merge buffers is 2). As discussed above, this is due to reassembly buffer requirements outweighing output buffer requirements. However when the cell gap $G = N = 16$ the ratio of VC merge to non-VC merge buffer sizes is large for all switch utilisations. This is due to differences in traffic burstiness. The tests were repeated with the NLANR traffic trace. The ratio of VC merge to non-VC merge buffer requirements for the trace driven simulation can be seen in Figure 6.20. The results show a similar trend, however, the penalty for the use of VC merge is less at $G = N = 16$ where the ratio is 20. However this is still significant. Other results generated from the trace driven simulation also show a similar trend to those generated using the synthetic traffic. These results are presented in Appendix B.
6.7 Network Layer Packet Forwarding

The previous section compares the performance of VC-merge and non-VC merge cut-through forwarding. In the cases examined, non-VC merge requires significantly less optical buffering for cut-through VCs. However not all packets are forwarded on cut-through VCs. Non aggregated protocols such as IP Switching forward a percentage of total traffic at the network layer. This is necessary to reduce the number of VCs and to account for setup delays. We examine two mechanisms to allow network-layer packet forwarding in cell switched optical networks. The first method uses an adjunct electrical router to perform reassembly and network layer forwarding when required. In this case no additional optical buffering is required, however, a small percentage of packets will be forwarded by the adjunct electrical router. The second method performs reassembly in the optical switch fabric to forward packets on a merged flow until a non-merged cut-through is created.
6.7.1 Adjunct Router

A block diagram of the first approach can be seen in Figure 6.21. Packets that require reassembly are switched through the optical switch to an adjunct electrical router. Within this router packets are converted to the electrical domain, reassembled, the IP forwarding decision is performed, they are segmented into cells, and finally an optical header to route the cell to the correct output port is added. The cells are then re-routed through the switching fabric to the correct output. AAL5 cell sequence integrity is maintained by sequential passage of routed packets thorough the electrical router to one input port of the cell switch.

The main design requirement of this approach is to minimise the use of the electrical router and switch a high percentage of cells in the optical switch. Results for the percentage of packets switched by the non-aggregated IP Switching protocol are presented in (Newman et al., 1998). This indicates that the utilisation of the adjunct router will be between 10% and 20% with a packet threshold of 10.
A significant advantage of the use of the adjunct router approach is that any underlying optical buffer can be used, as no reassembly is required in the optical domain.

6.7.2 Optical Reassembly

The second option we consider is performing reassembly within the optical switch fabric. We assume a fast IP lookup (Degermark et al., 1997) for the small percentage of cells while they are being buffered, or alternately the cells may be switched on an MPLS style routing table linked label switched path (which requires VC merge) until the non-aggregated path is set-up. After the non-aggregated label switched path is created the remaining packets in the flow do not require reassembly. We envisage that reassembly will be done in a similar manner to optical VC merge as we discuss in Section 6.4.1.

If reassembly is performed in the optical domain then the most important metrics to measure are related to buffer usage. It is important to reduce the size of buffers, and to reduce the time that cells spend in buffers.

6.7.3 Simulation of Network Layer Forwarding Options

The simulation model described previously was used to examine the reassembly buffer requirements of the two packet forwarding alternatives. Unless otherwise stated $\rho = 0.3$, $G = 5$ and a buffer overflow probability of $10^{-5}$ is used. We use actual backbone traffic traces to drive the simulation. All the results in the remainder of this section were generated with Australian trace 3 (See Table 5.1 for details). Results from other traces were similar and are not shown. In the adjunct router case we show the percentage of packets that must bypass the optical switch. The optical reassembly results show that aggregated approaches require twice the buffering of non-aggregated protocols for the same packet loss probability. This result is due to a smaller number of packets that require reassembly, as well as a significant difference in the size of routed and switched packets for the non-aggregated approach.
6.7.4 Results

6.7.4.1 Adjunct Router Results

Use of the adjunct router approach does not require additional cell buffering in the switch core. However it requires processing of a proportion of cells in the adjunct electrical router. It is important to reduce the percentage of cells processed by this router.

The percentage of packets and cells switched with a varying packet threshold is shown in Figure 6.22. The packet threshold value also has a significant effect on the number of VCs which is shown in Figure 6.23. The packet threshold value 10 is commonly chosen (Newman et al., 1998) as a compromise between VC usage and percentage of packets switched. At a packet threshold of 10 the percentage of packets forwarded by the adjunct router for source destination pair flows is 13.5% which corresponds to 9% of cells switched.

In Chapter 4 we examine flows of a higher aggregation. In particular we examined
the use of source/destination subnet pairs for use with our Destination Site Label Switching proposal when VC merge is unavailable (DSLS). In this case 7.4% of packets (4.6% of cells) are switched at a packet threshold of 10. However in Chapter 4 we also show that a lower packet threshold is suited to more highly aggregated flows. For a packet threshold of 5 the percentage of packets switched is 4.5% which represents 2.5% of cells.

6.7.4.2 Reassembly of Packets Requiring Network Layer Forwarding

This section examines the performance of non-aggregated label switching when the first few packets in a flow are reassembled to be forwarded on a merged circuit. The main parameter controlling the performance is the packet threshold. Varying this parameter controls the percentage of packets forwarded on the merged circuit and will have an effect on the reassembly buffer requirements.

In Figure 6.24 we show the relationship between packet threshold and buffer usage for an overflow probability of $10^{-5}$. The buffer usage for the aggregated approach
Packet Forwarding in Optically Switched Networks

Figure 6.24 Buffer size versus packet threshold for $10^{-5}$ packet loss probability

(where all packets require reassembly at the switch) of 105 cells is shown for comparative purposes, this is not dependent on packet threshold. Source pair flows use on average 45 cells of buffer for a packet threshold of 1 this increases to 79 cells for a packet threshold of 80. For source-pairs with a packet threshold of 10 (commonly used for source-pair flows) 62 cells are required. The subnet-pair approach requires slightly smaller buffers. For subnet-pair flows 55 cells are required for a packet threshold of 10 and 48 cells are required for a packet threshold of 5.

We examined the percentage of packets forwarded by IP in Figure 6.22. It can be seen that the percentage of cells forwarded at the packet level is significantly less than the percentage of packets. This result indicates a difference in average packet sizes for switched and routed packets. The average packet sizes for switched and routed packets versus the packet threshold, before a cut-through is created, can be seen in Figure 6.25. The choice of packet threshold has a significant effect on the average packet sizes of routed packets. With a packet threshold of 5 the average size of routed packets is 170 bytes this is 41% less than the average size of switched packets. At a packet threshold of 80 the average size of routed packets increases
to 240 bytes. A reduction in the size of routed packets will result in a smaller re-assembly buffer requirement for non-aggregated forwarding as seen in Figure 6.24. It is interesting to note the similarity between the average buffer size curve for non-aggregated approaches in Figure 6.24 and the average packet size for routed packets in Figure 6.25.

Figure 6.26 shows the relationship between buffer usage versus switch utilisation. At 30% utilisation the VC merge case requires 106 cells for an overflow probability of $10^{-5}$. The source-pair approach requires 62 cells of buffer which is 41% less than the VC merge case. The subnet pair approach requires 53 cells of buffer which is 50% of the VC merge buffer requirement. At 80% utilisation the difference is larger. The VC merge case requires 185 cells for an overflow probability of $10^{-5}$. Source-pair flows requires 52% less and the subnet-pair flows require 60% less buffer cells. Clearly the non-aggregated approach is significantly less sensitive to switch utilisation.

Even though the non-aggregated approach uses significantly less buffering, the required buffer sizes are still large for optical buffers. The adjunct router approach
with a low packet threshold in conjunction with traffic smoothing appears to be the best alternative.

### 6.8 Conclusions

This chapter in conjunction with the additional results in Appendix B examines the feasibility of using contemporary packet forwarding techniques within optical cell switches. We classified protocols into aggregated and non-aggregated depending upon reassembly buffer usage. The classifications were then compared by simulation to establish buffer size requirements. Initially we examined the penalty imposed by VC merge when output buffers were minimised. Alternative mechanisms were then examined to implement non-aggregated protocols in optical switches.

We investigated two cases where output buffers were minimised. The first instance examines the case where the switch was operated at low utilisation to minimise buffer usage. We found that at utilisations below 23%, VC merge approaches required sig-
significantly more buffering (greater than 50% additional buffering required). When a delay of one cell service time was added between each cell within an AAL5 encoded packet the performance of VC merge protocols was significantly worse. In this case VC merge approaches required significant additional buffering at utilizations below 43%. The next case examined was where traffic burstiness is altered by a traffic smoothing mechanism described in (Moser and Melliar-Smith, 1996). Each downstream switch was constrained, and could transmit only one cell every $N$ cell service time, where $N$ is the number of input ports. In the case where VC merge is not required buffer requirements was less than $N$ (in this case 16) cells per port. Use of VC merge removes cell gaps before adding traffic to output buffers. This increases the burstiness of the traffic arriving at the output buffer, and requires large reassembly buffers. The total buffers required for VC merge in this case was found to be 185 cells at 10% switch utilisation and 405 at 80% utilisation. This is significantly higher than the requirement for non-VC merge flows which in this case required less than 16 cells. The results were repeated with independent traffic traces from NLANR, and similar trends were observed.

The remainder of the chapter examined non-aggregated protocol implementation issues. Two cases were considered: use of adjunct electrical router, and optical reassembly of the packets that require network layer forwarding. The adjunct router approach forwards a small percentage of packets (4.5%) and cells (2.5%) in an adjunct electrical router if subnet-pair flows are used. The optical reassembly approach assumes that the packets that require network layer forwarding are reassembled in the optical switch. A simulation comparison shows that in this case aggregated forwarding requires over twice the number of cells of optical buffering than non-aggregated forwarding. The large difference was shown to be a result of a significant reduction in the average size of reassembled packets in the non-aggregated case. This was found to be an artifact of the packet threshold mechanism. Non-aggregated forwarding was also shown to be less sensitive to network utilisation. The buffer sizes required per port by each approach, for this trace, were determined to be 105 cells for aggregated and 48 cells for the non-aggregated packet forwarding approach with a packet threshold of 5 and subnet-pair flows. However, even for the non-aggregated case the buffer requirements are high when considering likely optical buffer sizes.
The adjunct router with a low packet threshold may be the most sensible choice.

This study shows that non-aggregated protocols such as MPOA and IP switching are more suited to the optical switching environment than aggregated approaches such as MPLS and Tag switching due to optical switch buffer constraints. We recommend using a non-aggregated label switching protocol with subnet-pair flows and traffic smoothing techniques in conjunction with an adjunct router. This approach uses the minimum buffer and only requires 2.5% of cells to be forwarded by the adjunct router.

The main contribution of this chapter is an examination of the use of VC merge in an optical switching environment where buffers are severely limited. Literature (Widjaja and Elwalid, 1999) indicates that VC merge does not add significantly to buffer requirements. We show that this is not the case with limited buffer sizes and gaps between cells that belong to the same segmented packet. An examination of different mechanisms to avoid VC merge and packet based optical label switching is left for future work.
Chapter 7

Conclusions

7.1 Overview

Label switching forwarding protocols such as Multi-Protocol Label Switching are likely to be incorporated into core Internet switches. This thesis examined the implementation of several label switching protocols in the future Internet. We concentrated on the scalability of best effort forwarding. Of particular interest was minimising the percentage of packets that require network layer forwarding in core Internet routers. A comprehensive review of label switching protocol and performance evaluation literature was presented first. We then provided a comprehensive classification of these protocols. This classification led onto three areas of work examining the performance of label switching protocols in future network scenarios. These future network scenarios were: a hierarchically routed IP version 6 Internet, usage of congestion sensitive multipath dynamic routing protocols, and operation over optical cell switches. This section presents the major conclusions of these studies.

7.2 Classification

In literature label switching protocols are generally classified as traffic driven or topology driven (White, 1998). However, there are many different protocols including: Multi-Protocol Over ATM (Fredette, 1997) (MPOA), Multi-Protocol Label Switching (MPLS) (Callon et al., 1997; Ahmed et al., 1997), Cell Switch Relay
(Katsube et al., 1997), IP Switching (Newman et al., 1998), and a hybrid approach proposed by the author called Destination Site Label Switching (DSLS) (Boustead et al., 1998a). We presented a comprehensive classification which serves as a framework that maps the similarities and differences of the large number of proposals. This framework was then used to highlight the advantages and disadvantages of each approach and motivated the subsequent chapters of this thesis.

7.3 Hierarchical IP Version 6 Networks

This study examined the use of label switching protocols in an IP version 6 network. The main aspect of IP version 6 that we examined was the hierarchical provider based address structure that will allow significant aggregation of routing table entries in core Internet routers. We showed that control driven label switching approaches such as Tag switching will not perform well in this environment. A large number of packets will require network layer forwarding at hierarchical boundaries. We proposed a new approach, called Destination Site Label Switching (DSLS), for use in this environment. This protocol takes advantage of the strict provider based nature of IPv6 addressing by using the address structure to classify flows. DSLS uses a flow based label assignment, that aggregates traffic using the IPv6 destination-site address, on the Internet wide AS path. DSLS also takes advantage of table-linked label switching within ASs where routing table aggregation is less of a concern.

We developed a methodology to examine the performance of DSLS using IP version 4 traffic traces and information from Internet address registries, route servers and route arbiter databases. In addition to examining the performance of DSLS we used this methodology to examine the performance of aggregated label switching protocols in general. In particular we examined the relationship between aggregation and the packet threshold parameter and show that a packet threshold lower than 10 is suitable for aggregated flows. We also showed that DSLS will forward less than 0.15% of packets at the network layer in autonomous system gateway routers and requires no network layer forwarding in the internal switches of an autonomous system. This represents a significant improvement over IP switching which required
approximately 15% network layer forwarding in all label switch routers. DSLS also required significantly less network layer forwarding than other highly aggregated flow based label switching approaches, such as FATDLM which aggregates flows based on routing information (instead of IPv6 address information as used by DSLS) and requires network layer forwarding for approximately 1% of packets in all label switches.

### 7.3.1 Summary of Major Findings

- The increased route aggregation likely with the introduction of IPv6 will impact on the performance of table-linked routing protocols by increasing the number of packets requiring network layer forwarding at boundary routers.

- We developed a label switching protocol for use with IPv6 called Destination Site Label Switching. This protocol uses the IPv6 provider based address structure to aggregate label switched paths between gateway routers independent of routing table aggregation. Within autonomous systems DSLS takes advantage of table linked forwarding.

- We showed that we can reduce the network layer forwarding requirement at gateway routers to less than 0.15% regardless of IP routing table aggregation using DSLS, this is significantly less than the 15% required by IP switching and 1% required by FATDLM.

### 7.4 Congestion Sensitive Routing Protocols

Flow based forwarding makes the routing decision at the start of a flow then creates a fixed label switched connection for the remainder of the packets within the flow. This study examined the use of flow-based forwarding in conjunction with congestion sensitive dynamic routing protocols. In this scenario routing information is likely to change frequently. We are interested in how sensitive flow based forwarding is to changes in underlying routing protocol information. This is particularly important to examine in light of our DSLS proposal that uses aggregated flow based forwarding.
We found that flow-based cut-through forwarding reacts significantly slower than expected, by examination of published flow length distributions (Claffy et al., 1995), to changes in routing tables. This problem was seen to get progressively worse between June 1995 and November 1997 indicating a possible trend in traffic characteristics. The reason for this unexpectedly slow reaction to route changes was shown to be a small percentage (between 3% and 4%) of long flows that contribute significantly to the total packet rate (18-50%). We also showed that the response of destination-site flows is unsuited to congestion sensitive multipath routing.

We found that adjustment of the packet threshold and flow-timeout parameters does little to improve the performance of flow based forwarding. The introduction of a maximum flow length will improve the sensitivity, however it will also increase the level of network layer forwarding required. Selecting a maximum flow length of 200 seconds for source-pair (IP Switching and MPLS) increased the number of packets requiring network layer forwarding by 2% in the worst case examined while dramatically improving the response to routing table changes. For source pair flows the percentage of packets forwarded on the new route 90 seconds after the route change (90 seconds is the update interval for IGRP) was 95%. Destination-site flows also showed significant improvement with the addition of an MFL of 200. Approximately 95% of packets follow the new route 90 seconds after the route change with a small increase in the network layer forwarding requirement.

7.4.1 Summary of Major Findings

- Flow based forwarding reacts to underlying routing table changes slower than predicted by published flow length distributions.

- DSLS destination-site label switched paths were found to perform poorly when used in conjunction with congestion sensitive routing protocols.

- Variation of flow timeout parameters does not significantly improve performance.

- The introduction of a maximum flow length significantly improves sensitivity to underlying routing table changes for aggregated (including DSLS destina-
tion site flows) and non-aggregated label switching. The penalty of this improved sensitivity is an increased network layer forwarding requirement. We found that this penalty was small. The additional network layer forwarding requirement, in terms of the percentage of total packets, was 2% for IP Switching flows and 0.13% for DSLS flows in the worst cases examined.

7.5 Optical Cell Switches

The use of optical cell switches will significantly increase Internet switching performance. In Chapter 6 we studied the feasibility of using contemporary label switching forwarding techniques in conjunction with optical cell switches. This study focused on the implementation of VC merge in this environment as well as buffer minimisation techniques. Previous studies (Widjaja and Elwalid, 1999) of VC merge show that the switch output buffer sizes greatly outweigh VC merge reassembly buffer requirements. However, optical buffer sizes are severely limited and output buffer minimisation techniques may be required. In the first part of this study we found that when buffer minimisation techniques are used to reduce output buffer sizes the VC merge reassembly buffer requirements became a significant proportion of total buffer needs. In fact, the use of traffic smoothing techniques can introduce additional gaps between cells that increases the VC merge reassembly buffer requirements.

Two buffer minimisation techniques were examined: maintaining low switch utilisation, and using traffic smoothing. In order to examine the first approach we varied switch utilisation to determine when VC merge reassembly buffers became a significant proportion of total buffer requirements (greater than 50%). Initial experiments were performed with no gaps between cells belonging to the same packet and it was found that at utilisations below 23% the VC merge buffers were a significant proportion of the total buffer requirement. When a cell gap of one cell was used the point at which VC merge reassembly buffers became significant increased to 43%. Our results indicate that VC merge reassembly buffers requirements are sensitive to an increase in cell gap particularly at low utilisation. The next case examined was the use of traffic smoothing techniques. Analysis and simulation techniques were used
to examine the output buffer and VC merge reassembly buffer requirements when traffic burstiness is altered by a traffic smoothing technique designed for optical cell switches (Moser and Melliar-Smith, 1996). Using this technique without VC merge the total buffer requirement was reduced from approximately 550 cell to 10 cells at 80% utilisation. If the VC merge mechanism is used then the traffic smoothing mechanism reduces the total buffer requirements (output + reassembly) from 550 cells to 250 cells. The traffic smoothing is not as effective in the VC merge case because the smoothing mechanism introduces gaps between cells belonging to the same packet and increases the VC merge reassembly buffer requirement.

Two cases were examined for the implementation of nonaggregated label switching protocols over optical cell switches: Use of an adjunct electronic router; and optical reassembly of the packets that require network layer forwarding. The adjunct router approach uses a standard optical cell switch with an electronic router connected in parallel to be used for network layer forwarding. Using IP Switching source - destination pair flows 15% of packets and 9% of cells required processing by the electronic router for the traces examined. Using IPv6 source site - destination site address pairs with a packet threshold of 5 the network layer forwarding requirement is reduced to 4% of packets and 2.5% of cells. The optical reassembly approach assumes that the packets that require network layer forwarding are reassembled within the optical switch fabric. A trace driven simulation comparison showed that the aggregated forwarding approach required 115 cells of optical buffer which was over twice the nonaggregated buffer requirement (50 cells).

This study shows that nonaggregated forwarding is more suitable for use with optical cell switches because VC merge reassembly buffers are not required. In addition nonaggregated approaches can be implemented using both the adjunct router and optical reassembly approach. The label switching implementation that has the smallest optical buffer requirement and minimises packet layer forwarding requirements is the use of the adjunct router approach with source site - destination site flows.
7.5.1 Summary of Major Findings

- We find that the use of VC merge, in conjunction with output buffer minimisation mechanisms, can significantly increase the buffer sizes required.

- Non aggregated label switching protocols that do not use VC merge are more suited to use in an optically switched environment.

- In order to minimise buffer requirements we recommend using non-aggregated label switching with site-pair flows in conjunction with traffic smoothing techniques and an adjunct router for the 2.5% of cells that require network layer forwarding.

7.6 Future Work

The work presented in this thesis provides a clear path for further work in several areas:

Packet Switching

This thesis examines packet over cell label switching protocols. A logical and important extension of this work is the examination of packet based label switching protocols.

Hierarchical Networks and Destination Site Label Switching

We highlighted a problem with scalable MPLS forwarding when used in a highly hierarchical IPv6 environment. We argue that highly aggregated routing tables will lead to a high level of network layer forwarding at major backbone BGP routers. An additional area of work is estimating the increase in route aggregation in future networks after the introduction of IP version 6.

Using an estimation of route aggregation in an IPv6 network (as discussed in the previous paragraph) in conjunction with an appropriate hierarchical network model it will be possible to quantify the MPLS network layer forwarding penalty. This will
enable complete quantitative comparison between the performance of MPLS and the proposed Destination Site Label Switching (DSLS) protocol.

Extending the work presented in Chapter 4 to packet based switching opens up an interesting area of work related to hierarchical networks. Currently an open area for work is an investigation into the use of label stacks to overcome the forwarding requirements at gateway routers in a highly hierarchical IPv6 network. It is important to develop and examine different alternatives of using label stacks in hierarchical networks. In particular it is necessary to examine the scalability of the state information required to be kept in switches to maintain label switched paths between routers at each level of the hierarchy.

The focus of this thesis has been on the use of label switching and IPv6 address hierarchies to reduce the number of packets forwarded at layer 3. This simplifies forwarding lookups from a longest prefix match algorithm to a direct table lookup. Current generation routers are able to maintain wire-speed forwarding up to OC-192 speeds by using Application Specific Integrated Circuits (ASICs). It is an interesting extension to examine the difference in complexity and scalability (in order to keep up with transmission increases due to DWDM) of hardware longest-prefix match algorithms versus hardware table lookups. This comparison is particularly interesting in light IPv6 which will increase the possible address space (from 32 bits to 128 bits) and QoS routing which will dramatically increase the number of routing table entries. In addition there is some concern that the use of ASICs for forwarding will reduce the ability to keep up with changes in IP standards due to the long time to market for ASICs (up to two years) (Partridge et al., 1998). An examination of the use of label switching taking into account these future issues is an interesting area for future work.

**Flow Based Forwarding and Dynamic Routing Protocols**

Chapter 5 examines the use of congestion based dynamic routing protocols in conjunction with flow based label switching protocols. We examined the reaction to a change in the underlying routing table. An interesting extension of this study would be to examine the implementation of a particular congestion based routing protocol.
Conclusions

We propose the implementation of a Maximum Flow Length to reduce the length of label switching flows and improve the response to routing table changes. We are only able to examine a limited range of values for MFL because of limited trace lengths. For MFL greater than 200 seconds (even lower for DSLS flows) the simulation did not reach a steady state. Use of longer traces would enable a more complete examination to determine the optimum value for MFL.

Optically Switched Networks

There is considerable scope for a continuation of the examination of the implementation of label switching protocols in optical networks, concentrating on optical packet switching. This thesis examines packet over cell label switching protocols (due to the cell switching focus of the entire thesis) and explicitly states the limitations of VC merge in this environment. It is also important to examine in more detail, as future work, alternate methods of avoiding the requirement for VC merge. An additional logical and important extension of this work is the examination of packet based label switching protocols including fixed size packet switching and unslotted variable size packet switching. Unslotted optical packet switches (with variable size packets) are starting to be considered the preferred option due to reduced complexity even though the chance of contention is higher (Yao and Dixit, 2000). The study of the impact of the use of label switching protocols over unslotted packet switches is an important area of future work.

The model used to examine the performance of VC merge in conjunction with optical cell switches used a fixed time interval between cells belonging to the same packets. A fixed gap between the cells was used to model round robin scheduling as well as the traffic smoothing technique proposed in (Moser and Melliar-Smith, 1996). A possible extension to this work is to introduce a random delay in addition to the fixed cell gap. This addition is likely to significantly increase the complexity of the model.

The traffic model used in Chapter 6 is based upon exponentially distributed packet sizes and packet inter-arrival times. These results predict the performance with actual traces within a reasonable degree. A more accurate traffic model, that takes into account self-similarity, may be examined as future work.
Our examination of optical switching concentrated on the use of optical delay lines and fiber loop buffers. This study can be extended to take into account other forms of contention resolution such as wavelength conversion as discussed in (Elmirghani and Mouftah, 2000a).

Another interesting extension of this optical switching work is to examine the interaction between optical label switching nodes and WDM wavelength routing nodes since they are likely to coexist (Danielsen et al., 1998b).
Bibliography


Appendix A

Hierarchical Trace Details

A.1 Subnet Size Distributions

This appendix examines the hierarchical traces used in Chapter 4. In particular we examine the hierarchy entries in the trace file that were used to examine the relationship between aggregation and label switching performance.

An example packet’s address hierarchy is shown in Figure A.3. Each level represents entries in the hierarchical database (described in Section 4.7 which match the packet. The lowest level hierarchy (level 1) has an anonymised address of 1286 and a bitmask of 20 bits. This represents $2^{32-20} = 4096$ addresses. The highest level mask is 8 bits which represents $2^{32-8} = 16777216$ addresses.

This section examines distributions that describe the hierarchical trace information for one of the Australian traces (Trace 3) which was taken on 16/11/1999. The duration of the trace is 1510 seconds and it contains 40,000,000 packets. Distributions for all other traces produced similar results.

Figure A.2 shows the distribution of hierarchy lengths (N) for all packets. This hierarchy length indicates the number of entries in the hierarchy database match individual packets. Less than 0.3% of packets only fit within one entry in the database. The percentage of packets that match 2 entries is 18%. Over 81% of packets match 3 or more entries in the database.
Hierarchical Trace Details

5055 391 A 1286\20 785\18 ..... 391\8

LevelN subnet\bitmask
(least specific match)

Level2 subnet\bitmask

Level1 subnet\bitmask (most specific match)

Pre CIDR class (A,B, or C)

Pre CIDR subnet address

Packet’s source or destination address

Figure A.1 Example address hierarchy

Figure A.2 Length distribution
Figure A.3  Subnet size distribution: level 1

The graphs in Figures A.3 to A.6 show the distributions of masks for each hierarchy level. The average bitmask size of level 1 is 20.9 bits. This increases to 12 bits at level 4.
**Figure A.4** Subnet size distribution: level 2

**Figure A.5** Subnet size distribution: level 3
Figure A.6 Subnet size distribution: level 4
Appendix B

Examination of Optical VC Merge Penalty Using Australian Traces

B.1 Introduction

This appendix presents the results from an examination of penalties introduced by the VC merge mechanisms in an optically switched network where buffer availability is severely limited. These results are placed in an appendix because they are similar to the analytical/simulation results presented in Chapter 6. Only the graph showing the additional buffer requirement is shown and discussed in Chapter 6.

B.2 Results

These results in figure B.1 to figure B.5 were generated using Australian trace number 3 as described in Table 5.1. This trace was obtained on 16/11/1999 has a duration of 1510 seconds and contains 40,000,000 packets.
Figure B.1 Total buffer requirement for VC merge versus utilisation

Figure B.2 Re-assembly buffer requirement for VC merge versus utilisation
**Figure B.3** Output buffer requirement for VC merge versus utilisation

**Figure B.4** Total buffer requirement for non-VC merge versus utilisation
Figure B.5 Trace : additional buffer requirement for VC merge