Investigation of Advanced Optical Packet-Routed Network Architectures

Michael Düser

A thesis submitted to the University of London for the degree of Doctor of Philosophy (Ph.D.)

Department of Electronic and Electrical Engineering
University College London
January 2003
To my parents
Abstract

This thesis investigates the scalability of future dynamic optical network architectures with respect to their adaptability to rapid changes in traffic patterns and the expected growth in traffic volume, whilst minimizing the required network resources. Static wavelength-routed optical networks (WRON), optical packet switched (OPS) and optical burst switched (OBS) networks are discussed as potential candidates for dynamic future architectures. Due to the limitations of both electronic and optical technologies for switching, processing and buffering, however, neither of the three fully satisfies the requirements for future network architectures. A novel dynamic network architecture, the wavelength-routed optical burst-switched (WR-OBS) network, has been, therefore, proposed, and is analysed in depth in the course of the thesis. The WR-OBS architecture allows access to optical bandwidth in fractions of the optical line rate, hence improving resource utilization. In the WR-OBS architecture, all processing and buffering are concentrated at the edge of the network, and bursts are routed over an optical transport core using dynamic wavelength assignment. The queueing of packets in OPS networks and burst aggregation processes in OBS networks are investigated as a function of input traffic statistics, key for the efficient operation of networks which must meet the stringent demands for low packet loss and low latencies of network applications. An analytical network performance model is devised for the WR-OBS network architecture, and novel parameters are introduced to quantify the benefits of dynamic over static wavelength allocation. Efficient scheduling of wavelength requests in the control plane of dynamic networks is essential to meet hard deadlines of time-critical applications. The modified rate monotonic (RM) and earliest deadline first (EDF) algorithms are successfully applied for the scheduling of dynamic network resources. The experimental investigation of fast switching, operation under 10 Gb/s modulation, and fast wavelength stabilisation show the merits, but also the physical limitations of fast tuneable lasers, ultimately limiting the minimum packet size in OPS networks. The results of this work can be applied to optimize the design rules of future optical packet networks, and quantify the operation regimes which best make use of the static or dynamic network architectures.
Acknowledgements

My thesis would not have been possible for me to complete without the help of several people who generously supported me during the last four years.

My foremost gratitude belongs to my supervisor Prof Polina Bayvel whose constant support, encouragement and guidance were crucial for the progress of my work. Her ability of being omnipresent for discussion, and her inspiration were key to the success of the project, and I would finally thank her for a thorough introduction into scientific writing.

I would also like to thank Prof John E. Midwinter for his comments on optical switching and sharing his critical views on future optical networks which are reflected in the operational principle of the WR-OBS network.

I am deeply grateful to Prof Steve Ferguson, and to Marconi plc, for not only providing a generous studentship but also for his invaluable contributions in conceiving the WR-OBS network architecture, and providing valuable inputs throughout the course of the project. I would also like to thank Madeleine Glick and Martin Ward (both formerly Marconi plc) for their support.

I would like to acknowledge Dr Damon Wischik from the Statistical Laboratory, Cambridge University, for his significant contributions to the appendix on traffic statistics, burst aggregation and queueing theory.

The experimental results on edge router design would not have been possible without Drs. Dave Robbins and Peter Williams (Bookham Technology, formerly Marconi) whose help was indispensable in obtaining access to state-of-the-art tuneable lasers.

I was glad to learn during my project that the links to Bell Laboratories, Lucent Technologies could weather the difficulties faced by the telecommunication industries. I am grateful to Dr. Martin Zirngibl, Crawford Corner Labs, for valuable collaboration and opportunities for in-depth discussion of my work with him and his colleagues, and to Bill and Pat Reed for providing me with a second home in New Jersey.

I would like to thank Claudio Silva and Prof. Alwyn Seeds, Ultrafast Photonics Group, UCL, for their collaboration on the optical injection phase lock loop scheme for fast wavelength locking.
For many years, Dr Chris Park was our main contact at Agilent Technologies. Without him and the generous loan of equipment many of our experiments had not been feasible, and I particularly would like to thank him for the transceivers used in the multimode fibre experiments.

My knowledge on optical fibres and multimode fibres in particular was fostered during my stay at Bell Laboratories prior to UCL. The multimode fibre provided by Yangtze Optical Fiber (China), with the help of John Edwards, proved superior to many conventional designs and was key for the successful completion of high-speed transmission experiments showing the potential for its use in 10 Gigabit Ethernet.

Drs. Dirk Breuer and Andreas Gladisch, T-Systems Nova, Germany, provided me with the opportunity to extend the view from 10 to 100 Gigabit Ethernet, requiring some radical thinking on the design of future metropolitan area networks.

My time in London would have been much less enjoyable without my friends and colleagues of the Optical Networks Group and the Ultrafast Photonics Group, in room 808. I am convinced that our team work is part of the success story of the Optical Networks Group, and I am grateful for everybody to tolerate my German Ordungssinn. It also was a novel and nice experience to watch former project students turning into successful scientists. I hope that one day my dream will come true to visit every single one in their native country! I would like to explicitly acknowledge the contributions of my project students, David Orders and Jamie Abella in 1999-2000, as well as Tom Lovell (2001-03) and Michael Laws (2002-03) to the project.

Finally I would like to thank IEEE LEOS for the award of the 2000 Postgraduate Student Fellowship and an unforgettable time in Puerto Rico.
Contents

Abstract .................................................................................................................................................2
Acknowledgements .............................................................................................................................3
Contents .................................................................................................................................................5
List of Tables ......................................................................................................................................10
List of Figures .................................................................................................................................11
List of Acronyms ................................................................................................................................18

Chapter 1  Introduction ..................................................................................................................21
  1.1 Traffic growth and traffic diversity driving new network architectures .......................21
  1.2 Review of optical network architectures ......................................................................27
    1.2.1 Static wavelength-routed optical network (WRON) architecture ......................27
    1.2.2 Optical packet-switched (OPS) network architecture ......................................28
    1.2.3 Optical burst-switched (OBS) network architecture ..........................................32
    1.2.4 Operation of OBS networks ..............................................................................33
    1.2.5 Timing in dynamic optical networks .................................................................38
  1.3 The wavelength-routed optical burst switched (WR-OBS) network architecture ......48
  1.4 Thesis structure ...............................................................................................................53
  1.5 References .......................................................................................................................55

Chapter 2  Scalability of processing and switching ..................................................................61
  2.1 Optical processing and optical buffering .......................................................................63
  2.2 Electronic switching and processing ..............................................................................65
    2.2.1 Fundamentals .........................................................................................................65
    2.2.2 Electronic switch fabrics with high throughput (Tb/s) ........................................65
    2.2.3 Moore's Law and electronic scaling ......................................................................66
    2.2.4 Heat dissipation ....................................................................................................69
    2.2.5 Electronic memory access times .........................................................................70
  2.3 Optical switching ..............................................................................................................70
    2.3.1 Optical switch architectures analysed .................................................................72
    2.3.2 Non-blocking characteristics ................................................................................75
    2.3.3 Number of required switching elements .............................................................76
    2.3.4 Control complexity ...............................................................................................78
    2.3.5 Insertion loss .........................................................................................................79
    2.3.6 Crosstalk accumulation – signal-to-noise ratio (SNR) .......................................81
    2.3.7 Discussion and comparison of optical switching architectures ..........................84
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 7 Summary and conclusions</td>
<td>198</td>
</tr>
<tr>
<td>Chapter 8 List of publications</td>
<td>202</td>
</tr>
<tr>
<td>Appendix A Queueing theory</td>
<td>205</td>
</tr>
<tr>
<td>A.1 Single station queueing systems</td>
<td>205</td>
</tr>
<tr>
<td>A.2 Kendall’s notation and different types of queueing systems</td>
<td>206</td>
</tr>
<tr>
<td>A.2.1 M/M/1 system</td>
<td>207</td>
</tr>
<tr>
<td>A.2.2 M/M/8 system</td>
<td>208</td>
</tr>
<tr>
<td>A.2.3 M/M/m system</td>
<td>209</td>
</tr>
<tr>
<td>A.2.4 M/M/1/K finite capacity system</td>
<td>209</td>
</tr>
<tr>
<td>A.2.5 M/G/1 system</td>
<td>210</td>
</tr>
<tr>
<td>A.2.6 GI/M/1 system</td>
<td>211</td>
</tr>
<tr>
<td>A.2.7 GI/G/1 system</td>
<td>212</td>
</tr>
<tr>
<td>A.3 References</td>
<td>214</td>
</tr>
<tr>
<td>Appendix B Traffic statistics</td>
<td>215</td>
</tr>
<tr>
<td>B.1 Discrete PDF</td>
<td>215</td>
</tr>
<tr>
<td>B.2 Continuous PDF</td>
<td>216</td>
</tr>
<tr>
<td>Appendix C Application of queueing theory for heavy-tailed distributions and finite buffer size</td>
<td>218</td>
</tr>
<tr>
<td>C.1 Heavy-tailed distribution</td>
<td>218</td>
</tr>
<tr>
<td>C.2 Short- and long-range dependence</td>
<td>219</td>
</tr>
<tr>
<td>C.3 Self-similarity</td>
<td>219</td>
</tr>
<tr>
<td>C.4 Classification of queueing theory</td>
<td>219</td>
</tr>
<tr>
<td>C.5 Buffer scaling for heavy- and light-tailed traffic inputs in short and large buffers</td>
<td>220</td>
</tr>
<tr>
<td>C.5.1 Queueing systems with small buffers – insensitivity to traffic characteristic</td>
<td>221</td>
</tr>
<tr>
<td>C.5.2 Queueing systems with large buffers</td>
<td>222</td>
</tr>
<tr>
<td>C.6 References</td>
<td>224</td>
</tr>
<tr>
<td>Appendix D Power spectral density for discrete sources</td>
<td>225</td>
</tr>
<tr>
<td>D.1 Power spectral density of discrete ON-OFF source</td>
<td>225</td>
</tr>
<tr>
<td>D.2 References</td>
<td>228</td>
</tr>
<tr>
<td>Appendix E Packet loss in queueing systems with many aggregate flows</td>
<td>229</td>
</tr>
<tr>
<td>E.1 Introduction</td>
<td>229</td>
</tr>
<tr>
<td>E.2 Likhanov/Mazumdar estimate</td>
<td>230</td>
</tr>
<tr>
<td>E.2.1 Assumptions</td>
<td>230</td>
</tr>
<tr>
<td>E.2.2 Buffer overflow probability</td>
<td>231</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Traffic growth on US Internet backbones in the period 1990-2000. The traffic quoted is the total traffic transported during December of that year, and was converted into Gb/s average bit rate. The ranges given for 1997-2000 reflect the min-max values due measurement uncertainty [Odl02] ................................................................. 23

1.2 Comparison of different OBS schemes with one-way and two-way reservation (1-13), and datagram formats in standard protocols (14-16) for comparison ................................................................. 36

2.1 Summary of results for optical switches of 100 x 100 ports. Bold indicates failure of system requirements (insertion loss < 30 dB, OSNR > 15 dB), italic is used for those which missed the target criteria by less than 10% .......................................................................................................... 86

2.2 Summary of results for optical switches of 1000 x 1000 ports. Bold indicates failure of system requirements (insertion loss < 30 dB, OSNR < 15 dB), italic is used for those switches which miss the target criteria by less than 10% ..................................................................... 86

3.1 Summary of the loss curve l(b) scaling as a function of the buffer size and the ON and OFF traffic characteristics [Boo02]. The sublinear regime refers to functions of the form ln(b) or b^x with 0 < x < 1 ...................................................................................... 100

4.1 Definition of timing parameters used for the description of the WR-OBS architecture ............ 127

5.1 Network parameters used to quantify t_{R,W,A,max} in Figure 5.4 for seven given network topologies [Bar98] ....................................................................................................................... 153

A.1 Widely used symbols to describe the arrival and service time processes in queueing systems [Bol98] ....................................................................................................................... 206

A.2 Abbreviation and descriptions of widely used queueing disciplines [Bol98] ......................... 206

A.3 Upper and lower bounds (UB and LB, respectively) for the GI/G/1 queueing system [Bol98] ....................................................................................................................... 212

B.1 Discrete PDFs and their parameters, mean, variance, and coefficient of variation ............... 215

B.2 Continuous PDFs and their parameters, mean, variance, and coefficient of variation ........ 217

H.1 Physical interfaces defined for use with 10 Gigabit Ethernet (10GbE) as defined by the standard IEEE 802.3ae ....................................................................................................................... 240

H.2 Profile parameter \( \alpha \) and distortion \( \chi^2 \) measured for all eight GI-MMF under test .......... 247

H.3 Summary of packet transmission results on the 1.1-km and 4.4-km multimode fibre. No errors for transmission over 12.7-km SSMF recorded ....................................................................................................................... 262
List of Figures

1.1 Traffic growth on US Internet backbones in the period 1990-2000, data as shown in Tab. 1.1 [Odl02]. Reliable data was obtainable only until 1996 ................................................................. 24

1.2 Aggregate voice and IP traffic volume growth in the US for 2000, extrapolated to 2010 with growth factors of 10% for voice and 100% and 30% for data ................................................................... 25

1.3 Static wavelength-routed optical network (WRON) architecture, where lightpaths are provided as end-to-end connections. The inset shows the occupation of wavelengths, with packets on one wavelength each heading towards the same destination ........................................ 28

1.4 Optical packet-switched (OPS) architecture in a 4-node network, with insets of packets on wavelengths, and an optical packet switch where control information is extracted, and delay lines are required: at the input for delaying the packet until the control information was processed, and at the output as buffer for contention resolution ......................................................... 30

1.5 Optical burst-switching (OBS) network architecture with burst aggregation at the edge, separated control channel and burst control packets offset from the payload, and an OBS node with full wavelength conversion ......................................................................................... 33

1.6 Principle of one-way reservation of switching resources for the OBS Just-Enough-Time (JET) approach. At each hop the preset offset time, $t_{offset}$, is decreased by the processing time, $t_{proc}$, hence limiting the time the burst can spend in the network and the number of hops it can undergo ......................................................................................................................... 35

1.7 Timing chart for optical packet-switching (OPS). The packet is buffered at each node for header processing and switch reconfiguration .................................................................................. 40

1.8 Timing chart for OBS Just-Enough-Time (JET), without acknowledgement of successful transmission of datagrams. The control packet (header) preceding the burst configures the switch such that the following payload can pass without being buffered ........................................ 41

1.9 Timing chart for OBS Just-in-Time (JIT) providing an acknowledgement of a successfully received BCP at the destination edge router. When sending the burst it is unclear, however, whether the required resources will be available ......................................................................................... 42

1.10 Burst loss probability for an OBS JET core node versus relative load and for 4-128 wavelengths, using the Erlang loss formula ........................................................................................................ 43

1.11 Timing chart for the WR-OBS network architecture where bursts are released subject to a transmission acknowledgement by the network control. The edge delay $\tau_{edge}$ is counted from the first packet entering the empty queue. The release of resources depends on the burst aggregation scheme used; the limited burst size (LBS [Mig01]) was used here .................................................. 48

1.12 Principle of the wavelength-routed optical burst switched (WR-OBS) network architecture. Bursts are aggregated at the network edge, and wavelength request packets are sent to the control node for reservation of resources ..................................................................................... 50
1.13 Wavelength request processing in the central control node of the WR-OBS architecture. Requests are fed through a 2-stage queueing system before an acknowledgement is returned to the edge router, and control information sent to core routers for configuration .................................................. 52

1.14 After receipt of the acknowledgement, a tuneable laser in the sending edge router is tuned to the assigned wavelength and the burst transmitted via an end-to-end lightpath across the network ............................................................................................................................................................ 53

2.1 Moore's Law for the number of transistors integrated onto a single microchip with linear fit (on log scale). The number doubles approximately every two years .......................................................................................................................... 67

2.2 Prediction based on Moore's Law as it was shown in Fig. 2.1: Clock frequency (left axis, [Roa01]) and number of transistors per chip (Moore's Law, right axis) forecast for the period 2000 – 2016 ........................................................................................................................................ 67

2.3 Number of transistor operations per second for the period 2000 – 2016, prediction based on Moore's Law with an increase of approximately 63% p.a. ........................................................................................................................................ 68

2.4 Number of instructions carried out per unit time in microprocessors during 1986-1997 [Odl02]; the growth rate was approximately 60% p.a. ................................................................................................................................. 69

2.5 N-stage planar switch architecture [Spa84] ........................................................................................................ 72

2.6 Standard crossbar switch architecture [Non84] .................................................................................................... 73

2.7 Double-crossbar architecture [Kon85]; it offers a higher contrast ratio and therefore lower crosstalk than the standard crossbar .............................................................................................................................................. 73

2.8 Benes-switch architecture requires the lowest number of elements for its realisation [Ben85]...................... 73

2.9 Three-stage Clos switch which interconnects several small subnetworks [Clo53] .............................................. 74

2.10 Spanke switch type 1 [Spa86] .......................................................................................................................... 74

2.11 Spanke switch type 2 [Spa86] ........................................................................................................................................ 74

2.12 Jajszczyk switch type consisting 1:k splitters and combiners connecting to kxk subnetworks [Tho91] ....... 75

2.13 Extended Baseline switch type. The term I stands for input, the term D for the subnetwork, and the term O for output switches [Jaj93] .......................................................................................................................... 75

2.14 Levels of blocking and the determination of non-blocking characteristics for the switches investigated in this chapter. The last level, redundant strictly non-blocking has not been reached yet .............................................................................................................................. 76

2.15 Number of required 2x2 switching elements. The number grows O(N²) except for the Benes architecture with O(N-logN). ............................................................................................................................................. 78

2.16 Control complexity of optical switching architectures, i.e. the number of required drivers/controllers to operate the switching matrix ............................................................................................................................................ 79

2.17 System insertion loss for various switching architectures ............................................................................... 81

2.18 Crosstalk accumulation as function of number of switching elements, expressed in optical signal-to-noise ratio (OSNR) assuming 20 dB crosstalk per 2x2 switching element ............................................. 84
3.1 ON-OFF source for traffic generation of variable length packet and variable length gaps. When active, the source operates with a constant output bit rate $b$ ................................................................. 97

3.2 Simulation results for burst size and PLR as a function of $t_{\text{edge}}$ and a mean input bit-rate $b_{\text{in}} = 10 \text{ Gb/s}$. The burst size increase is plotted for an infinitely large buffer size, but PLR for a finite buffer size $B = 400 \text{ Mbit (47.7 MB)}$. Bars indicate 95% confidence level for Pareto distributed packets and interarrival times, with minimum packet size of 5 kB ........................................ 109

3.3 Deviations of the PLR for Pareto packet, Pareto inter-arrival time statistics ($\alpha = 1.5$) for $B = 400 \text{ Mbit (47.7 MB)}$ and minimum packet size of 5 kB. For clarity, frequency values > 1000 were omitted. The results show that the heavy-tailed input traffic statistics result in a significant distribution of the PLR. In Fig. 3.2 only the mean PLR was plotted. Selected traces are re-plotted in Fig. 3.4 .............................................................................................................. 110

3.4 Probability $P$ for PLR $> X$ (cumulative distribution) for Pareto packet size ($\alpha = 1.5$), Pareto inter-arrival time distribution ($\alpha = 1.5$), min. packet size of 5 kB .............................................................................................................................. 111

3.5 Simulation results for the PLR as a function of the edge delay for $B = 400 \text{ Mbit (47.7 MB)}$ and a mean input bit-rate $b_{\text{in}} = 10 \text{ Gb/s}$, for different levels of packet size, 5 kB (dash), 0.5 kB (dot) and 50 bytes (dash-dot). The PLR calculated from burst size distribution (assumed normal) is shown by solid line ........................................................................................................... 112

3.6 Feedback control mechanism for burst assembly based on measurement of the traffic mean and variance ............................................................................................................................................. 114

3.7 Aggregation delay and burst size for overflow probability of $10^{-4}$ for a buffer size of 4 Mbit. The mean bit rate was increased in a step like manner from 20 to 27, and 27 to 32 Gb/s after 300,000 and 600,000 simulations. Inset: fit of the burst size distribution with a normal distribution .............................................................................................................................................. 116

4.1 Timing diagram showing key timing parameters in the burst aggregation and burst transmission processes (limited burst size scheme, LBS). The timing constants are defined in more detail in Table 4.1 below ........................................................................................................... 127

4.2 Diagram for the calculation of the results on burst size distribution during the edge aggregation process, obtained by simulation or analytical model as specified in chapter 3 (grey shaded), are used to derive network performance parameters and parameter deviations in this chapter ........................................................................................................................................... 131

4.3 Wavelength holding time, $t_{\text{wht}}$, as a function of the edge delay, for $b_{\text{core}} = 100 \text{ Gb/s}$, $b_{\text{in}} = 10 \text{ Gb/s}$ and $t_{\text{RTT}} = 2, 5, 10 \text{ ms (a)}$, and $t_{\text{RTT}} = 5 \text{ ms and } b_{\text{core}} = 20, 40, 100 \text{ Gb/s (b)}$. Bars show 95% confidence level ........................................................................................................................................... 132

4.4 Bandwidth-per-wavelength, $B_{\text{pack}}$, for $b_{\text{core}} = 100 \text{ Gb/s}$, $b_{\text{in}} = 10 \text{ Gb/s}$ and $t_{\text{RTT}} = 2, 5, 10 \text{ ms (a)}$, and for $t_{\text{RTT}} = 5 \text{ ms and } b_{\text{core}} = 20, 40, 100 \text{ Gb/s (b)}$. Bars show 95% confidence level ........................................................................................................................................... 133

4.5 Mean bandwidth utilization, $U$, as function of the edge delay $t_{\text{edge}}$ and bit rate ratio $A$, for $t_{\text{RTT}} = 2 \text{ ms (a)}$ and $t_{\text{RTT}} = 10 \text{ ms (b)}$ ............................................................................................................................................. 134

4.6 Mean wavelength reuse factor, RUF, as function of edge delay $t_{\text{edge}}$ and bit rate ratio $A$, for $t_{\text{RTT}} = 2 \text{ ms (a)}$ and $t_{\text{RTT}} = 10 \text{ ms (b)}$ ............................................................................................................................................. 135
4.7 Network performance parameter, $P$, showing the trade-off between utilization $U$ and wavelength reuse $RUF$ for constant $t_{\text{edge}}$, $t_{\text{RTT}} = 2$ ms (a) and $t_{\text{RTT}} = 10$ ms (b). Optimum values $P_{\text{opt}}$ are shown by dashed line .............................................................. 136

4.8 Wavelength reuse factor, $RUF$, as a function of the round-trip time ($t_{\text{RTT}}$) for $t_{\text{edge}} = 10, 20, 50$ ms, and bit rate ratio $A = 10$. Shaded region: network requires more wavelengths than in a static WRON ......................................................................................................................................... 137

4.9 Lightpath utilization $U$, as a function of the round-trip time ($t_{\text{RTT}}$) for $t_{\text{edge}} = 10, 20, 50$ ms and bit rate ratio $A = 10$ ........................................................................................................................................ 138

5.1 Number of edge routers as a function of processing time $C$ and for 3 CoS using the RM and the EDF algorithms with $T_1 = 5$, $T_2 = 15$ and $T_3 = 45$ ms, as well as for 1 CoS (lowest latency $T_1 = 5$ ms only), $t_{\text{RTT}} = 5$ ms ........................................................................................................................................ 147

5.2 NSFNet architecture with node number 6 designated as control node; link length given in kilometers (a). Distribution of wavelength request periods $T_k$ in 5 ms intervals (b) ..................................................................................................................... 148

5.3 Periods $T$ for the NSFNet architecture for the case of 3 CoS with edge delays of 40, 50, and 80 ms. Periods of CoS 1 and 2 overlap due to the asymmetric distribution of round-trip times within the network .................................................................................................................................... 149

5.4 Upper bound for the processing time of the RWA algorithm for the case of a single processor (solid line) with the number of wavelengths $W$ as parameter. The multiprocessor case is independent of $W$ (dashed line). Values for real network architectures for both cases are given as well (open symbol – single processor, full symbol – multi-processor). Linear fit to values for real networks, single processor case indicates scaling $O(N^{1.7})$ .................................................................................................................. 152

6.1 Principle of a four-section fast tuneable semiconductor laser .................................................................................................................. 162

6.2 SG-DBR laser #2: Wavelength versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA. The intra-modal boundaries were chosen to 0.4 nm (50 GHz) .............................................................................................................................. 164

6.3 SG-DBR laser #2: Optical output power versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA .............................................................. 165

6.4 SG-DBR laser #2: Side-mode suppression ratio (SMSR) versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA .............................................................. 165

6.5 Wavelength versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear) .......................................................... 166

6.6 Output power versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear) .......................................................... 167

6.7 Side-mode suppression ratio (SMSR) versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear) .......................................................................................... 167

6.8 $dP/dI$ for GCSR laser (a) with power dip due to mode hopping. Linear fit to the ramp (b) to determine the $dP/dI$ for gain currents of 15-30 mA .................................................................................................................. 168
6.9 Selection of wavelength for the SG-DBR laser #1 according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 40 dB (a) and -3dBm ≤ output power ≤ 0 dBm (b) ..... 169
6.10 Selection of wavelength for the SG-DBR laser #2 according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 40 dB (a) and -3dBm ≤ output power ≤ 0 dBm (b) ..... 169
6.11 Selection of wavelength for the GCSR laser according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 30 dB (a) and -6 dBm ≤ output power ≤ -3 dBm (b) ..... 170
6.12 Set-up for tuneable laser switching experiments to analyze temperature dependent wavelength drift ................................................................................................................. 171
6.13 Switching of SG-DBR laser #2 to wavelength 1540.7 nm, using a burst duration of 90 ms. The electrical control signal (blue) is shown for reference ................................................................. 173
6.14 Wavelength drift and 90%-10% signal fall time as a function of total current change for SG-DBR laser #2 ........................................................................................................................................... 174
6.15 Transmitter eye diagram using a MZ modulator at 1540.7 nm with 1.2 dBm output power; the measured extinction ratio was 10.2 dB ....................................................................................................................... 176
6.16 Experimental set-up for burst BER analysis in back-to-back configuration.............................................................................................................................. 177
6.17 BER as a function of receiver input power for continuous transmission and burst lengths of 2 μs – 100 ms at 1547.8 nm ........................................................................................................................................... 178
6.18 Eye diagrams for burst generation for the continuous case (a), a maximum burst length in the experiment of 100 ms (b), and a minimum burst length of 2 μs (c) at 1547.8 nm .......................................................................................................................... 179
6.19 BER as a function of receiver input power for continuous transmission and burst lengths of 2 μs at 1554.2 nm ........................................................................................................................................... 180
6.20 Eye diagrams for burst generation for the continuous case (a), a maximum burst length in the experiment of 100 ms (b), and a minimum burst length of 2 μs (c) at 1554.2 nm .......................................................................................................................... 181
6.21 Receiver input power as function of burst length from 100 – 900 μs using the SG-DBR laser #2 with PRBS 2\(^{-1}\) and extinction ratio of approx. 10 dB. The dashed line indicates performance for continuous transmission of signal .......................................................................................................................... 182
6.22 Optimum Q-factor versus burst length for 100 – 900 μs using SG-DBR laser #2 at 1540.8 nm and a 2\(^{-1}\) PRBS ........................................................................................................................................... 183
6.23 Experimental set-up for fast wavelength-locking experiments using an SG-DBR laser; locking is obtained by using an optical injection phase-locked loop (OIPPL, 2\(^{nd}\) order - loop path in blue), which uses an optical frequency comb generator (OFCG) as a reference .......................................................................................................................... 186
6.24 Non-inverting integrator [Tit93] used as PLL loop filter in the OIPPL experiment, for provisioning of a feedback signal into the modulator input of the current controller of the gain section of the laser .......................................................................................................................... 187
6.25 Electrical control signal fed into the modulation input (\(V_{\text{gain}}\)), also converted in frequency deviation based on a chirp of 88 MHz/mV. The signal was averaged 16 times. The trigger signal (13 Hz) is shown for reference .......................................................................................................................... 188
6.26 Wavelength drift of the free-running tuneable laser versus the case when wavelength locking with OIPPL is operational. Wavelength measured with 2 pm resolution ........................................................................................................................................... 190
6.27 Wavelength locking for the same wavelength as in Figure 6.26 (a), and corresponding power reading and trigger signal (b), indicating that the switching and wavelength stabilization process are carried out in < 20 μs.................................................................190

6.28 Wavelength locking (a), and corresponding power reading and trigger signal (b) for a switching speed of 1 kHz........................................................................................................................191

6.29 Wavelength locking (a), and corresponding power reading and trigger signal (b) for a switching speed of 10 kHz......................................................................................................................192

A.1 Service station with single queue and m servers.................................................................................205

A.2 Comparison of the mean number of tasks for different queueing systems as a function of utilization ρ (M/D/1, M/M/1 and M/G/1 for c_b = 0.5, 1, 2, 3).............................................................................................211

C.1 I(b) ∝ v(b) plotted for Pareto distributed ON times with parameters A = 2 and α = {1.1, 1.5, 1.9}, showing sublinear increase for large values of b for large buffer size. Absolute values are not relevant since the function v(b) requires appropriate scaling................................................223

H.1 The bandwidth-distance product as a function of profile parameter α for 850 nm (a) and 1300 nm (b) for three optical sources with different rms linewidth: DFB laser (<0.1 nm), VCSEL (1 nm) and LED (15 nm) .........................................................................................................245

H.2 Typical refractive index profile for GI-MMF used in the experiment, data taken by a refractive-index measurement technique ..................................................................................................................246

H.3 Profiles of four selected GI-MMF used in the experiments. Distortions on fibre 6, right slope, are assumed to be local....................................................................................................................247

H.4 MMF transmission measurements set-up using a controlled singlemode launch .........................249

H.5 Bit error rate (BER) vs received optical power for back-to-back (btb) (filled square), and for four fibres showing the best case (open square), the worst case (circle), and two typical fibres (diamond, triangle) .............................................................................................................249

H.6 Transmission penalty (for 10^(-9) BER) as a function of the input launch position (for same fibres as in Figure H.5) .......................................................................................................................250

H.7 Transmission penalty at BER 10^{-10} for centre launch (O) and the variance of the profile distortions $\chi^2 (\cdot )$ .........................................................................................................................251

H.8 Experimental set-up for testing of the RouterTester equipped with 1310 nm, OC-192 transceiver units ...............................................................................................................................................253

H.9 Eye diagram of the optical transmitter containing a directly modulated DFB laser with -2.7 dBm output power and 7.4 dB extinction ratio .................................................................254

H.10 Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -6.88$ dBm........................................257

H.11 Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -7.2$ dBm, jitter peak-to-peak (pp) 33.4 ps ..............................................................................................................................257

H.12 Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -7.3$ dBm, jitter (pp) 37.2 ps ........257

H.13 Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm........................................259
List of Figures

H.14 Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm, for the same measurement as shown in Figure H.13 .................................................................259

H.15 Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm ........................................260

H.16 Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm, for the same measurement as shown in Figure H.15. The eye was recorded for a 60 minute period. The eye shows significant distortions due to modal noise and jitter .........................................................260

H.17 Principle of bursty traffic generation in Agilent RouterTester ..........................................................261

H.18 Arrival statistics of SDH/SONET BIP-1 errors as a function of the elapsed measurement time, showing the burstiness of errors observed in the experiment .........................................................264
**List of Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3R</td>
<td>Reamplification, retiming, reshaping</td>
</tr>
<tr>
<td>ACF</td>
<td>Autocorrelation function</td>
</tr>
<tr>
<td>ACK</td>
<td>Acknowledgement</td>
</tr>
<tr>
<td>ASON</td>
<td>Automatically switched optical network</td>
</tr>
<tr>
<td>ATM</td>
<td>Asynchronous Transfer Mode</td>
</tr>
<tr>
<td>AWG</td>
<td>Arrayed waveguide grating</td>
</tr>
<tr>
<td>BCP</td>
<td>Burst control packet</td>
</tr>
<tr>
<td>BER</td>
<td>Bit-error rate</td>
</tr>
<tr>
<td>BIP-1/-2/-3</td>
<td>Parity check bits in SDH/SONET</td>
</tr>
<tr>
<td>BIPS</td>
<td>Billion instructions per second</td>
</tr>
<tr>
<td>BPF</td>
<td>Bandpass filter</td>
</tr>
<tr>
<td>CBR</td>
<td>Continuous bit rate</td>
</tr>
<tr>
<td>CLT</td>
<td>Central-limit theorem</td>
</tr>
<tr>
<td>CO</td>
<td>Central office</td>
</tr>
<tr>
<td>CoS</td>
<td>Class of service</td>
</tr>
<tr>
<td>DAC</td>
<td>Digital-to-analogue converter</td>
</tr>
<tr>
<td>DBR</td>
<td>Distributed Bragg-reflector</td>
</tr>
<tr>
<td>DFB</td>
<td>Distributed feedback</td>
</tr>
<tr>
<td>DG</td>
<td>Delay generator</td>
</tr>
<tr>
<td>DMD</td>
<td>Differential mode delay</td>
</tr>
<tr>
<td>DRWA</td>
<td>Dynamic routing and wavelength assignment</td>
</tr>
<tr>
<td>DWDM</td>
<td>Dense wavelength division multiplexing</td>
</tr>
<tr>
<td>EDF</td>
<td>Earliest deadline first</td>
</tr>
<tr>
<td>EDFA</td>
<td>Erbium-doped fibre amplifier</td>
</tr>
<tr>
<td>EMI</td>
<td>Electromagnetic interference</td>
</tr>
<tr>
<td>EON</td>
<td>European Optical Network</td>
</tr>
<tr>
<td>FBAT</td>
<td>Fixed burst aggregation time</td>
</tr>
<tr>
<td>FBG</td>
<td>Fibre Bragg grating</td>
</tr>
<tr>
<td>FBM</td>
<td>Fractional Brownian motion</td>
</tr>
<tr>
<td>FDL</td>
<td>Fibre delay line</td>
</tr>
<tr>
<td>FIFO</td>
<td>First in first out</td>
</tr>
<tr>
<td>FPI</td>
<td>Fabry Perot Interferometer</td>
</tr>
<tr>
<td>FSR</td>
<td>Free spectral range</td>
</tr>
<tr>
<td>GbE</td>
<td>Gigabit Ethernet as defined by IEEE 802.3z</td>
</tr>
<tr>
<td>GCSR</td>
<td>Grating-assisted codirectional coupler with rear sampled-grating reflector</td>
</tr>
<tr>
<td>GFP</td>
<td>Generic Framing Procedure</td>
</tr>
<tr>
<td>GI</td>
<td>Graded-index</td>
</tr>
<tr>
<td>GMPLS</td>
<td>Generalized Multiprotocol Label Switching</td>
</tr>
<tr>
<td>H</td>
<td>Hurst parameter</td>
</tr>
<tr>
<td>IBT</td>
<td>In-band-terminator</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Identical and independently distributed</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>IPv6</td>
<td>Internet protocol version 6</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunications Union</td>
</tr>
<tr>
<td>JET</td>
<td>Just-Enough-Time</td>
</tr>
<tr>
<td>JIT</td>
<td>Just-In-Time</td>
</tr>
<tr>
<td>LAN</td>
<td>Local area network</td>
</tr>
<tr>
<td>LAUC-VF</td>
<td>Latest available unused channel with void filling</td>
</tr>
<tr>
<td>LBS</td>
<td>Limited burst size</td>
</tr>
<tr>
<td>LDC</td>
<td>Laser diode controller</td>
</tr>
<tr>
<td>LDP</td>
<td>Large deviations principle</td>
</tr>
<tr>
<td>LiNb0_3</td>
<td>Lithiumniobate</td>
</tr>
<tr>
<td>LRD</td>
<td>Long-range dependence</td>
</tr>
<tr>
<td>MAN</td>
<td>Metropolitan area network</td>
</tr>
<tr>
<td>MDP</td>
<td>Moderate deviations principle</td>
</tr>
<tr>
<td>MEMS</td>
<td>Micro electro mechanical systems</td>
</tr>
<tr>
<td>MMF</td>
<td>Multimode fibre</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-protocol label switching</td>
</tr>
<tr>
<td>MZI</td>
<td>Mach-Zehnder interferometer</td>
</tr>
<tr>
<td>NBF</td>
<td>Narrow bandpass filter</td>
</tr>
<tr>
<td>NOLM</td>
<td>Nonlinear optical loop mirror</td>
</tr>
<tr>
<td>OBS</td>
<td>Optical burst switching</td>
</tr>
<tr>
<td>OCDMA</td>
<td>Optical code division multiple access</td>
</tr>
<tr>
<td>OFCG</td>
<td>Optical frequency comb generator</td>
</tr>
<tr>
<td>OFL</td>
<td>Overfilled launch</td>
</tr>
<tr>
<td>OIPLL</td>
<td>Optical injection phase lock loop</td>
</tr>
<tr>
<td>OSA</td>
<td>Optical spectrum analyzer</td>
</tr>
<tr>
<td>OSNR</td>
<td>Optical signal-to-noise ratio</td>
</tr>
<tr>
<td>PC</td>
<td>Polarization controller</td>
</tr>
<tr>
<td>PD</td>
<td>Photodiode</td>
</tr>
<tr>
<td>pdf</td>
<td>Probability density function</td>
</tr>
<tr>
<td>PHY</td>
<td>Physical interface</td>
</tr>
<tr>
<td>PLL</td>
<td>Phase-locked loop</td>
</tr>
<tr>
<td>PLR</td>
<td>Packet loss rate</td>
</tr>
<tr>
<td>PPG</td>
<td>Pulse pattern generator</td>
</tr>
<tr>
<td>PRBS</td>
<td>Pseudo-random bit sequence</td>
</tr>
<tr>
<td>PSTN</td>
<td>Publicly-switched telephone networks</td>
</tr>
<tr>
<td>Q-factor</td>
<td>Quality factor</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of service</td>
</tr>
<tr>
<td>RAM</td>
<td>Random access memory</td>
</tr>
<tr>
<td>RBN</td>
<td>Rearrangeably nonblocking</td>
</tr>
<tr>
<td>RFD</td>
<td>Reserved-Fixed-Duration</td>
</tr>
<tr>
<td>RM</td>
<td>Rate monotonic</td>
</tr>
<tr>
<td>RTT</td>
<td>Round-trip time</td>
</tr>
<tr>
<td>RUF</td>
<td>Reuse factor</td>
</tr>
<tr>
<td>RWA</td>
<td>Routing and wavelength assignment</td>
</tr>
<tr>
<td>SCM</td>
<td>Subcarrier multiplexing</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>SDH</td>
<td>Synchronous Digital Hierarchy</td>
</tr>
<tr>
<td>SG-DBR</td>
<td>Sampled-grating distributed Bragg-reflectors</td>
</tr>
<tr>
<td>SGN</td>
<td>Strictly nonblocking</td>
</tr>
<tr>
<td>SMSR</td>
<td>Side mode suppression ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SOA</td>
<td>Semiconductor optical amplifier</td>
</tr>
<tr>
<td>SONET</td>
<td>Synchronous Optical Network</td>
</tr>
<tr>
<td>SRAM</td>
<td>Static random access memory</td>
</tr>
<tr>
<td>SRD</td>
<td>Short-range dependence</td>
</tr>
<tr>
<td>TAG</td>
<td>Tell-and-go</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission control protocol</td>
</tr>
<tr>
<td>TDM</td>
<td>Time division multiplexing</td>
</tr>
<tr>
<td>TOAD</td>
<td>Terahertz optical asynchronous demultiplexer</td>
</tr>
<tr>
<td>U</td>
<td>Lightpath utilization</td>
</tr>
<tr>
<td>UBS</td>
<td>Unlimited burst size</td>
</tr>
<tr>
<td>UDP</td>
<td>User datagram protocol</td>
</tr>
<tr>
<td>VCSEL</td>
<td>Vertical cavity surface emitting laser</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide-area network</td>
</tr>
<tr>
<td>WDM</td>
<td>Wavelength division multiplexing</td>
</tr>
<tr>
<td>WHT</td>
<td>Wavelength-holding time</td>
</tr>
<tr>
<td>WR-OBS</td>
<td>Wavelength-routed optical burst switching</td>
</tr>
<tr>
<td>WRON</td>
<td>Wavelength-routed optical network</td>
</tr>
<tr>
<td>WSNB</td>
<td>Wide-sense nonblocking</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

1.1 Traffic growth and traffic diversity driving new network architectures

Recent years have seen a significant increase in the capacity provided by optical transmission systems with transmission capacities well in excess of 10 Tbit/s over a single fibre [Big01, Var02]. This development was fuelled by the often quoted (but incorrect, as shown below) assumption that data traffic would overtake traditional telephony traffic by doubling in volume every 3-4 month – the equivalent to a growth factor of 8-16 per year.

The traffic growth was enabled by the improvements in optical transmission systems operating at ever increasing bit rates per fibre: From the 1980s, single-channel optical transmission systems using single mode fibre (SMF) were used in wide-area networks (WAN) [Kog00]. Capacity was increased by advances in the time-division multiplexing (TDM) technology, which made it possible to aggregate a number of lower bit rate channels into a single wavelength at higher bit rate [Kei99].

Since these networks were mainly used for transmission of voice channels, electronic multiplexing technologies such as Asynchronous Transfer Mode (ATM) [Kes97] and the Synchronous Digital Hierarchy (SDH, referred to as Synchronous Optical Network (SONET) in the US) [Cav02, Bal02/1] were developed for reliable transmission and access to the transmission bandwidth. The advent of the Erbium-doped fibre amplifiers (EDFA) [Des02] was the key enabler to use multiple wavelengths on a single fibre, leading to wavelength-division multiplexing (WDM) transmission systems [Kec00]. In WDM closely spaced wavelengths carrying differently modulated channels could be transmitted on a single fibre, increasing point-to-point transmission capacity.
The elimination of electronic repeaters was a cost-effective way to increase in bandwidth removing the need for installation of new fibres. The development of the EDFA in the 1990s led to an increase in the maximum capacity of data carried per fibre. Data from [Kog00] shows that for single-channel optical transmission systems (laboratory experiment) the annual increase in capacity was approximately 63% for the period 1982-93, and then showed a sharp rise, increasing to approximately 147% p.a. for the period 1994-2000. The capacity of commercially deployed WDM transmission systems followed the same trend but delayed for about 5-8 years, so WDM systems started to have an impact on operators' networks from the mid-1990s. Despite the surge in capacity enabled by WDM systems, however, the capacity-distance product (measured in Gb/s-km) has shown a steady growth rate of approximately 100% p.a. for the period 1975-2000 [Kog00]. Thus WDM technology provided the means to maintain a steady growth rate for optical bandwidth when the single-channel transmission systems faced new challenges, for example through the impact of dispersion [Agr92].

In today's backbone networks, WDM transmission is used in combination with SDH/SONET technology for provisioning of point-to-point connections. Extensions to the SDH/SONET format, such as the Generic Framing Procedure (GFP) [Gor02], are aiming at providing a higher efficiency in carrying data traffic, usually variable in length, over TDM systems, which are typically operated in a synchronous manner (e.g. SDH/SONET 125 μs) with fixed slot sizes.

However, the steadily increasing optical transmission capacity leads to scalability problems of the electronic routers faced with throughput requirements of several Tb/s. Despite progress in high-capacity router design [Cha02/1, Sin02] it would be conceptually easier to completely avoid the opto-electronic conversion in core nodes, and to extend the optical reach by providing end-to-end lightpaths, as in wavelength-routed optical networks (WRON, see for instance [Bar97]), where wavelengths are used for routing. Termed transparent optical networks they are relatively easy to design, but with the assumption that the traffic demand is known a-priori, these networks may be slow to respond to changes in traffic volume or distribution.

Dynamically reconfigurable optical networks, such as optical packet-switched (OPS) networks, could potentially provide even greater flexibility in responding to changes in the network load. This would mean to substitute circuit-switched by all-optical packet-switched network architectures, including optical processing and buffering functionalities for packet forwarding, which were proposed (e.g. [Blu00, Cha98]). OPS
networks have been proposed either as a hybrid solution, where the processing is carried out electronically (e.g. WASPNET [Hun98], HORNET [Shr00]), or as true optical networks where all functionalities are implemented in optics, e.g. the processing and forwarding of optical labels [Blu00]. Labels are short headers attached to the original packet for fast processing in the optical network, and are removed on exit from the optical network.

As already mentioned, the main driver for the design of various future network architectures was the expected increase in data traffic volume compared to voice traffic which dominated the network design to date. However, obtaining precise raw data on the growth rate of networks is nearly impossible as those are regarded as strategically important to the success of a business and are, therefore, never released for public use. For analysis in this chapter, data collected on different US networks for the period 1990-2000 was used [Odl02]. Despite its degree of imprecision for the period 1997-2000, it is the most comprehensive collection of data and network traffic analysis to date. [Odl02] argues that the assumption on traffic growth in US backbones with doubling rates of 3-4 months was correct only for a brief period in 1995-1996 as shown in Table 1.1, and plotted in Figure 1.1, respectively.

<table>
<thead>
<tr>
<th>Year</th>
<th>Data transfer TB/month</th>
<th>Average bit rate Gb/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>1.0</td>
<td>0.003</td>
</tr>
<tr>
<td>1991</td>
<td>2.0</td>
<td>0.007</td>
</tr>
<tr>
<td>1992</td>
<td>4.4</td>
<td>0.015</td>
</tr>
<tr>
<td>1993</td>
<td>8.3</td>
<td>0.028</td>
</tr>
<tr>
<td>1994</td>
<td>16.3</td>
<td>0.055</td>
</tr>
<tr>
<td>1995</td>
<td>155.9</td>
<td>0.529</td>
</tr>
<tr>
<td>1996</td>
<td>1,500</td>
<td>5.09</td>
</tr>
<tr>
<td>1997</td>
<td>2,500 - 4,000</td>
<td>8.5 - 13.6</td>
</tr>
<tr>
<td>1998</td>
<td>5,000 - 8,000</td>
<td>17.0 - 27.1</td>
</tr>
<tr>
<td>1999</td>
<td>10,000 - 16,000</td>
<td>33.9 - 54.3</td>
</tr>
<tr>
<td>2000</td>
<td>20,000 - 35,000</td>
<td>67.9 - 118.7</td>
</tr>
</tbody>
</table>

Table 1.1: Traffic growth on US Internet backbones in the period 1990-2000. The traffic quoted is the total traffic transported during December of that year, and was converted into Gb/s average bit rate. The ranges given for 1997-2000 reflect the min-max values due to measurement uncertainty [Odl02]

1 TB/month is equivalent to approximately 3.4 Mbit/s average bit rate
2 data for 1995 interpolated from 1994-1996 period
It can be seen that Internet traffic grew by approximately 100% in the years from 1990-1994, then showed a sharp increase by 860% p.a. in 1995-1996, and finally dropped to approximately the previous growth rate of 106% for 1996-2000. Interestingly, the traffic growth rate observed at 100% p.a. for most of the last decade nearly coincides with the growth rate for the capacity-distance product of optical transmission systems, which is 100% p.a [Kog00].

This analysis of traffic growth and the prediction of future traffic levels is essential for the design of future network architectures since it allows to predict the required capacity and is key to all future network design. In addition it is essential to forecast the composition of traffic according to different applications. The service requirements of different traffic applications (e.g. IP or telephony) may vary widely with respect to the required bandwidth, and the tolerable levels of latency and packet loss per application.

Figure 1.1 shows the increase in traffic volume in US backbones for the period 1990-2000, and the growth rate is then used to predict the increase in data and voice traffic for the period 2000-2010 (Fig. 1.2).

In 2000, the average bit rate in US backbones was < 120 Gb/s in total, or the equivalent of 12 x 10 Gb/s channels. Projecting the results into the future and assuming that traffic growth maintains 100% growth per year, this would result in a maximum aggregate
Internet traffic in the US backbone of approximately reaching 100 Tb/s by the year 2010. Taking into account the low average utilization of data networks, the respective installed optical transmission and routing capacity would be 1 Pb/s, assuming a utilization factor of 10%. The low utilization is based on figures from a European operator for the year 2000 [Ten02]: The average bit rate of IP traffic transported was 5.07 Gb/s on an installed capacity allowing for a maximum of 80 Gb/s, resulting in an average utilization of 6.3%.

Figure 1.2 shows a comparison of the forecasts for the volume increase in voice and IP backbone traffic in the US for the period of 2000-2010, based on the minimum value for the year 2000 (70 Gb/s for data and 179 Gb/s for voice [Odl02]). The extrapolation considers a 10% increase in voice traffic as well as 30% and 100% growth in Internet traffic. The numbers for the growth of IP traffic are derived from the average growth rate in IP traffic for the last decade to 100% p.a. – Figure 1.1 showed that the growth rate of approximately 100% was true for most of 1990-2000. A growth rate of 30% p.a. would reflect a more pessimistic outlook on the growth of data traffic, as before of fibre optic communication systems started to impact available capacity in the 1970s [Kec00]. Figure 1.2 shows that for growth rates considered (30% and 100%) data traffic would overtake the volume of voice traffic in the next couple of years, justifying the shift from a voice-centric to a data-centric network architecture.

![Figure 1.2: Aggregate voice and IP traffic volume growth in the US for 2000, extrapolated to 2010 with growth factors of 10% for voice and 100% and 30% for data](image)
The time for this cross-over, however, depends on the actual growth rate of data traffic. For a 100% increase as observed in the last decade, voice and data traffic should equalize by the end of 2002 (with a volume of approximately 200 Gb/s each), whereas as slowdown to 30% p.a. would mean that a cross-over would not happen before 2006. However, the requirements of data traffic in terms of packet loss rate and delays must be analyzed as these parameters will determine the design of new networks. Based on observations for the period 1994-2000, [Odl02] also proposes that there is reason to be cautious about future traffic requirements since even Internet traffic is composed of a variety of different types. With web (browsing) accounting for the largest portion of traffic (40-50%), e-mail and ftp have a share of 5% and are declining in importance, but unclassified traffic types are rising (30-40%). Amongst these is traffic from file sharing applications such as Napster, the (partly illegal) sharing of music files on the Internet. In conclusion, future optical network architectures must be able to mainly cope with new types of data traffic, but still need to support quality of service (QoS) for growing voice traffic, not least since voice traffic currently provides most of operators’ revenues. This is highlighted by figures from the year 2000, when operators’ revenues for voice were $300 billion compared to $25 billion from Internet services [Odl02].

Both the traffic growth rate (Fig. 1.2) and the growth rate of the capacity-distance product of optical transmission systems [Kog00] have been approximately 100% p.a. for the past decade. It would be, therefore, fair to assume that the traffic volume reaching each existing router in the core network through its fibre optic links grows at the same rate – requiring to double the throughput per router each year. According to this growth rate, current commercially available electronic routers with a maximum non-blocking throughput of 320, 480 and 640-Gb/s (Cisco 12416, Marconi BXR-48000 and Juniper T-640, respectively) would be expected to be upgraded to Tb/s throughput within 1-2 years from now. However, taking into account the numbers for traffic growth in the US (Fig. 1.2, with 100% increase in data traffic p.a.), then it is not before 2004 that the average aggregate bit rate for the US backbone will reach 1 Tb/s, so routers with Tb/s throughput may only be required in a few nodes, but not across the entire network.

Any future optical network architecture will thus need to reflect the increase in total traffic volume, as well as the change in traffic volume caused by individual applications. A number of optical network architectures which have been proposed to address these aspects are described and analysed in the next section.
1.2 Review of optical network architectures

In the following, three types of optical network architectures are described in more detail, which have been proposed as potentially capable of meeting the requirements as set out in the previous section to meet the demand of future network applications with respect to traffic volumes and patterns. To date, three main architectures have been proposed, namely the wavelength-routed optical network (WRON), the optical packet-switched (OPS) and the optical burst-switched (OBS) network architectures, and are described in the following sections.

1.2.1 Static wavelength-routed optical network (WRON) architecture

The wavelength-routed optical network (WRON) architecture (for example [Bar98/1] and references therein) provides end-to-end lightpaths across the whole network, as shown in Figure 1.3. The principle of operation is depicted for a 4-node network for simplicity. For connection between nodes, transparent lightpaths are established between nodes, e.g. between node 1 and node 3 using wavelength $\lambda_1$, and between node 4 and node 2 using $\lambda_2$. Each wavelength only carries datagrams between two nodes, intermediate nodes (such as node 1 for $\lambda_2$) do not have access to this bandwidth (except acting as transit nodes) even if it is not fully used. Additional wavelengths (denoted here by $\lambda_n$) are available, but only used where necessary – so $\lambda_n$ is shown to be idle on the link between nodes 1 and 2 since it is only used to connect nodes 4 and 3.

The functionality of core nodes is relatively simple, since only demultiplexers and optical space switches are required with the optical channel determining only the routing direction. Hence, relatively slow (millisecond operation) space switches would probably be sufficient for this purpose, such as micro electro mechanical system (MEMS) switches, scalable to port counts of 1296x1296 with switching times of 5 ms [Ryf01].

Although relatively simple to design, dimension and operate, WRONs are less flexible in adapting to rapidly changing loads since the wavelength allocation problem must be solved and optimized globally [Bar97] as it is assumed that static demands are known a-priori. A second problem is the coarse bandwidth granularity since connections can access fibre bandwidth only in the form of entire wavelengths, e.g. 10 Gb/s. As not all applications will entirely consume this bandwidth, the WRON approach may result in an inefficient use of bandwidth, and more flexible architectures such as OPS and OBS,
as described in the following, might be preferred. In terms of performance, WRONs have no delay or blocking as it is assumed that enough wavelengths would be allocated to satisfy the traffic demand. In designing these networks, the resources (number of wavelengths) used must be minimized and the investigation of dynamic optical network architectures must address whether this is possible.

**Figure 1.3: Static wavelength-routed optical network (WRON) architecture, where lightpaths are provided as end-to-end connections. The inset shows the occupation of wavelengths, with packets on one wavelength each heading towards the same destination**

### 1.2.2 Optical packet-switched (OPS) network architecture

Dynamic networks such as OPS and OBS network architectures are fuelled by the perception that they provide a finer granularity and react faster to traffic changes than static WRON where the wavelength allocation is based on the assumption that traffic demand is known a-priori.

In OPS networks, the functionalities of electronic packet networks, such as forwarding in IP [Kes97], are mapped into the optical domain, where packets are forwarded on a packet-by-packet basis. Because of the high-speed operation (≥ 40 Gb/s) it is likely that optical packets will contain several electronic packets to minimize overhead required for header and receiver clock synchronization. This would be case, for instance, if the optical packet length was fixed. The OPS schemes described in [Gri01], [Rub02], and [Tan00] would serve as examples, with payload of the optical packet was assumed to be
fixed at 32 kbit, 2.5...25 kbit, 3 kbit, respectively, significantly larger than the minimum IP packet size of 320 bits. These sizes would, however, be compatible with larger electronic packets and match the average packet length of 389.5 bytes (3.12 kbit) of measured traffic quoted in [Xio00].

The key advantage of OPS networks as compared to electronic packet switching would be transparency. In electronic packet switching, core nodes operate at a specified bit rate, e.g. 2.5 or 10 Gb/s. Upgrading to a higher bit rate means in the electronic domain that the router architecture (line cards) must be changed to allow termination of transmission signals at the higher bit rate. In an OPS network, however, the control information and payload can be coded at different bit rates. If only the header information was extracted and processed at each core node, but the payload switched optically, then a wide variety of payload formats and bit rates could be supported. For OPS the time allowed for header processing and reconfiguration of the switch are critical parameters. For instance, a throughput of 1 Tb/s is equivalent to a forwarding (and packet processing) rate of 321 million packets/second, assuming an average packet size of 389.5 bytes [Xio00]. With the same figures the average packet length becomes 311 ns at 10 Gb/s line rate, requiring fast reconfiguration of the optical switch fabric.

The aspects of header processing and switch reconfiguration times are investigated in more detail in chapter 2 of this thesis.

The processing of header information can be carried out either in the electronic domain, as for example in the WASPNET [Chi01] and HORNET [Shr00] architectures which use a sub-carrier multiplexing (SCM) header, or also carried out in the optical domain, as for optical label processing [Blu00]. All these are examples for the principle operation of OPS networks. An example of a generic OPS network architecture is shown schematically in Figure 1.4: as in electronic packet networks, wavelengths are used for transmission only. There are no restrictions as to which wavelengths can be used for transport, as the inset to Figure 1.4 shows, where packets heading to the same destination are transmitted on different wavelengths, e.g. towards node 2; wavelengths are not used for routing as in WRONs.
Figure 1.4: Optical packet-switched (OPS) architecture in a 4-node network, with insets of packets on wavelengths, and an optical packet switch where control information is extracted, and delay lines are required: at the input for delaying the packet until the control information was processed, and at the output as buffer for contention resolution

The design of optical packet switches is technically more complex than the switch required in WRONS since optical components cannot be integrated monolithically as well as electronic circuits. The principle of an optical packet switch is shown as inset to Figure 1.4: at the ingress of the optical packet switch the header information is extracted from each channel, and the entire packet delayed in fibre delay line buffers for the time required for header processing. As mentioned above, processing of control information could be carried out either electronically or optically, configuring the optical switch (could be a space switch as in Fig. 1.3). The optical packets are forwarded to the respective output ports. For the output-buffered switch configuration shown here, fibre delay lines (FDL) or other delay elements such as recirculating loop buffer [Tuc98] are provided for contention resolution, and optional wavelength conversion can reduce packet loss further [Dan98, Hun99] – but requiring a wavelength converter per output port and hence increasing the component count. OPS, however, faces a number of technical challenges, discussed in detail in section 2.1:

1. FDLs and other elements such as recirculating loop buffer are difficult to integrate and inherently less flexible than electronic buffer
2. Full wavelength conversion (from any wavelength to any other wavelength) across the C-band and with acceptable SNR is not yet available. Also the cascadability of SOAs is limited [Wol00].
3. Estimation of required switch sizes: a concept is assumed as shown in Fig. 1.4, where the optical WDM signal from each fibre is demultiplexed into individual wavelengths when entering the switch. Assuming further a nodal degree [Bar98/1], i.e. number of links (fibres) attached to each node, between 2 and 7, and a high capacity transmission system with 128 wavelengths per fibre, demonstrated with bit rates of 10 Gb/s [Gro01] or 40 Gb/s [Gro02], the port count per switch must be in the range 256...896. An additional challenge are the fast header processing times mentioned above, and the fast reconfiguration time required in the ns regime for packet-by-packet forwarding. Integrated optical switches such as optical MEMS switches have been demonstrated for a port count of 1296x1296 [Ryf01], but are slow to reconfigure (millisecond regime, e.g. 5 ms in [Ryf01]) and, therefore, insufficient for OPS.

Optical packet switch fabrics using passive arrayed waveguide gratings (AWG) in combination with wavelength agile sources (e.g. tuneable lasers) for addressing a particular output port have been proposed and experimentally demonstrated for switch capacities of 1.2 Tb/s [Gri01]. Packet switches incorporating AWGs have also been shown to be cascadable for up to 25 hops without optical or electrical generation at 2.5 Gb/s (penalty-free for pseudo-random bit sequence [PRBS] length of $2^{7}$-1) [Tza00], an important feature for the implementation of transparent optical networks.

Optical packet-switches with throughputs of several Tb/s and with active switching elements (other than MEMS, e.g. SOAs) have also been proposed - see, for example [Wan02] and references therein, and [Chi00] for an optical switch with 10 Tb/s throughput. Experimental test beds for optical packet switches are, however, typically limited to switches with smaller dimensions, e.g. to 4x4 [Gui00/2].

Optical packet switches are analogue in nature, so unlike in digital switches the optical signal is impaired by crosstalk and insertion loss as it passes through the switch. These impairments depend on the technology used to implement the switch fabric. The previously mentioned 1296x1296 optical MEMS switch [Ryf01] showed 5.1±1.1 dB loss and an average 38 dB crosstalk, and a switching speed of 5 ms. Optical packet switches with SOAs as switching gates are inherently faster (sub-nanosecond), but their impairments limit the achievable cascadability. A study for SOAs operating at 10 Gb/s showed that the number of SOAs which can be cascaded depends on the length of the device, the gain and output power [Wol00]. For an output power of -10 dBm it was possible to cascade 15 SOAs, which decreased to a maximum of 6 when the output power was increased to 0 dBm (both for a penalty of 1 dB at 10 Gb/s [Wol00]), severely
Introduction

limiting the cascadability. This leads to a complicated switch design process, e.g. with respect to the minimum input power and extinction ratio, which depends on the physical parameters of both the SOAs (loss, output power, crosstalk) and the optical input signal (e.g. extinction ratio). See, for example [Chi00, Wan02]. In [Cha02/2] target value for the crosstalk of SOAs in optical switch fabrics was shown to be 45 dB for a power penalty of less than 1 dB.

The switching speed of optical packet switches with fast tuneable lasers and AWGs for routing is determined by the tuning speed of the lasers, measured to be 50 ns maximum by several research groups [ODo01, Gri01, Klo02]. Experimental 4x40 Gb/s transmission through an AWG with 3.1 dB insertion loss was shown to be feasible with a penalty of 4 dB...4.5 dB per channel [Gri02]. In a recirculating loop experiment carried out at 2.5 Gb/s bit rate (27-1 PRBS), penalty-free transmission was achieved when cascading 25 AWGs [Tza99]. Extending this experiment to include packet-switching functionality [Gui00/1] allowed to cascade 14 nodes incorporating the AWGs for routing, for a bit error rate in excess of 10^9.

The scalability of fast optical switches by cascading individual, fast 2x2 switching elements to this port count is studied for different switch architectures and with respect to crosstalk and insertion loss in chapter 2. The experimental evaluation of fast tuneable lasers used as sources in future dynamically reconfigurable optical networks (chapter 6) provides a lower bound to the reconfiguration times that will be required in such networks.

Although OPS networks provide the finest bandwidth granularity, they have been shown difficult to implement: the main disadvantage of all-optical processing over electronics is the lack of large-scale integration, resulting in bulky set-ups. A particular problem presents the implementation of FDL buffer, where buffer depths equivalent to several thousand packets were shown to be needed when bursty traffic was present at the input [Hun98]. Fast scalable switches (with reconfiguration times in the nanosecond regime) are difficult to scale, as shown in chapter 2 of this thesis. Relaxing these requirements as in OBS (described next) may provide technically a more feasible solution.

1.2.3 Optical burst-switched (OBS) network architecture

OBS has been proposed [Qia00/1, Tur99] in an attempt to reduce the processing in network nodes needed for packet forwarding, as well as allowing access to bandwidth in
fractions of a wavelength for higher granularity and thus potentially increasing the use of resources achievable in WRONs. Despite the lack of a formal definition, the term *burst* in literature refers to an entity variable in length, comprising at least two, but usually several packets destined to the same network egress. In contrast to optical packet-switching, the control information (burst header) is offset from the payload in time, and transmitted on a separate control channel (hence offset in both time and wavelength), as shown in Figure 1.5.

It is assumed that an (electronic) edge router is attached to each (optical) core node to carry out the burst aggregation. The core is assumed to be optical and bufferless. The principle of OBS operation is discussed in more detail in the next section of this chapter.

![Figure 1.5: Optical burst-switching (OBS) network architecture with burst aggregation at the edge, separated control channel and burst control packets offset from the payload, and an OBS node with full wavelength conversion](image)

### 1.2.4 Operation of OBS networks

The OBS approach reflects some of the initial ideas from electrical burst-switching [Ams89], which, however, was never implemented in a real network environment. At the time, operating on the much lower bit rates of individual voice circuits, bursts were proposed as a means to transport data fitting into the silence periods of telephony traffic, increasing the bandwidth utilization and allowing a variable length data container. The self-routing feature in electrical burst-switching resembles the labeled OBS scheme proposed in [Qia00/2]. It was also already acknowledged that the electrical burst
switching was competing with fast circuit-switching and packet-switching, and might contain selected features of each, such as processing of header information per burst, not per packet, and thus reducing the required amount of electronic processing. The size of bursts, however, would be smaller than the equivalent of a network round-trip time, so breaking with the principle of circuit-switching.

Although, as mentioned, there is not a universal definition of optical burst switching, [Dol01/1] lists a number of characteristics, which are inherent to most of the schemes. These are:

- burst size granularity which lies between OPS and circuit-switched networks such as WRONs.
- separation of control information (header) and data (payload), a one-way reservation scheme (for most cases)
- variable burst length
- integrity: bursts are not disassembled during transmission in intermediate nodes. This rule would be violated, however, by schemes which drop parts of contending bursts in a core router [Vok02], requiring complex switching matrices with explicit access to individual packets within a burst.
- no optical buffering (but use of full optical wavelength conversion in core nodes to reduce burst loss). However, it was recently proven that some form of buffering might be necessary to implement QoS when using one-way reservation, see for example [Yoo00, Dol02]

All OBS schemes proposed to date assume the use of separate burst header (control) and payload (data) channels, where headers are sent into a bufferless switch network with an appropriately chosen offset time, $t_{\text{offset}}$, from the data to reserve switch resources for routing the associated data appropriately along the selected path. The principle of the one-way reservation scheme is shown in Figure 1.6 for the OBS Just-Enough-Time (JET) scheme [Qiao00/1], described further in section 1.2.5.2. The payload is preceded by a control packet which contains information about the burst destination, the length of the burst, and the offset time. After a preset offset time, $t_{\text{offset}}$, the payload is released into the network. At each intermediate core router information is extracted from the control packet and switch resources reserved if available.
If the reservation was successful, a new control packet would be generated and forwarded to the next router, with the offset time decreased by the amount necessary for processing, $t_{\text{proc}}$. Figure 1.6 presents a case where $t_{\text{offset}} = 3 \cdot t_{\text{proc}}$, i.e. the burst can undergo a maximum of three hops. The mechanism of decreasing the offset time in each node provides the same functionality as the time-to-live (TTL) field in IP packets [Kes97] since the burst can only undergo a limited number of hops before being discarded from the network, hence preventing congestion in the OBS network.

Over time, modifications to the initially proposed OBS schemes such as JET [Qiao00/1] and HORIZON [Tur99] appeared, but the principle of one-way reservation was maintained. Due to the lack of acknowledgements (except for schemes number 1 and 7 in Table 1.2, explained in more detail in section 1.2.5), OBS schemes with one-way reservation are also referred to as “send-and-forget.”

Table 1.2 summarizes the properties of OBS schemes proposed to date and their key features, such as admission schemes, buffering in core nodes, typical burst lengths and the respective burst duration assuming 40 Gb/s line rate in a future optical core network, and the consideration of QoS, where this aspect received some recent attention after it was shown that sole one-way reservation even with varying offsets for QoS would result in unacceptable burst loss when implemented in the network context [Dol01/2]. The burst sizes considered vary, but for most one-way reservation schemes are on the same order as individual IP packets or Ethernet frames. One exception is the proposed segmented OBS scheme [Vok02] which considers bursts with burst duration of 100 ms at 10 Gb/s.
### Table 1.2: Comparison of different OBS schemes with one-way and two-way reservation (1-13), and datagram formats in standard protocols (14-16) for comparison

<table>
<thead>
<tr>
<th>No</th>
<th>Scheme</th>
<th>Admission scheme</th>
<th>Buffer in core nodes</th>
<th>Typical burst length at 40 Gb/s</th>
<th>Burst duration at 40 Gb/s</th>
<th>QoS</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WR-OBS</td>
<td>Limited</td>
<td>No</td>
<td>25 MB</td>
<td>5.24 ms</td>
<td>Yes</td>
<td>[Due02]</td>
</tr>
<tr>
<td>2</td>
<td>Horizon</td>
<td>Full</td>
<td>Yes</td>
<td>1 kB</td>
<td>0.2 µs</td>
<td>No</td>
<td>[Tur99]</td>
</tr>
<tr>
<td>3</td>
<td>JET</td>
<td>Full</td>
<td>Yes</td>
<td>15 kB</td>
<td>3.1 µs</td>
<td>Yes</td>
<td>[Yoo00]</td>
</tr>
<tr>
<td>4</td>
<td>RFD</td>
<td>Full</td>
<td>No</td>
<td>MB^</td>
<td>ms</td>
<td>Yes</td>
<td>[Qia00/2]</td>
</tr>
<tr>
<td>5</td>
<td>TAG</td>
<td>Full</td>
<td>No</td>
<td>15 kB^</td>
<td>3.1 µs</td>
<td>No</td>
<td>[Qia00/1]</td>
</tr>
<tr>
<td>6</td>
<td>IBT</td>
<td>as TAG</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>JIT</td>
<td>Full</td>
<td>No</td>
<td>12.2 kB; 1.1 TB*</td>
<td>2.5 µs; 250 s</td>
<td>No</td>
<td>[Wei99]</td>
</tr>
<tr>
<td>8</td>
<td>Comparison 2-7</td>
<td>Full</td>
<td>No</td>
<td>12.5 kB</td>
<td>2.56 µs</td>
<td>Yes</td>
<td>[Dol01/1]</td>
</tr>
<tr>
<td>9</td>
<td>LAUC-VF</td>
<td>Full</td>
<td>Yes</td>
<td>0.39 kB</td>
<td>0.8 µs</td>
<td>No</td>
<td>[Xio00]</td>
</tr>
<tr>
<td>10</td>
<td>JumpStart</td>
<td>Full/Ltd.*</td>
<td>Yes</td>
<td>variable*</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>MPLS-based</td>
<td>Limited</td>
<td>No</td>
<td>0.059-64 kB^</td>
<td>12 ns-13.1 µs</td>
<td>Yes</td>
<td>[Ver00]</td>
</tr>
<tr>
<td>12</td>
<td>Assured Horizon</td>
<td>Limited</td>
<td>Yes</td>
<td>10 kB</td>
<td>2 µs</td>
<td>Yes</td>
<td>[Dol02]</td>
</tr>
<tr>
<td>13</td>
<td>Segmented OBS</td>
<td>Full</td>
<td>No</td>
<td>1 GB</td>
<td>25 ms</td>
<td>Yes</td>
<td>[Vok02]</td>
</tr>
<tr>
<td>14</td>
<td>IPv6</td>
<td>–</td>
<td>–</td>
<td>40 byte - 64 kB</td>
<td>12 ns-13.1 µs</td>
<td>–</td>
<td>RFC 1683</td>
</tr>
<tr>
<td>15</td>
<td>Ethernet</td>
<td>–</td>
<td>–</td>
<td>64 - 1536 byte</td>
<td>12.8 ns-0.3 µs</td>
<td>–</td>
<td>IEEE 802.3</td>
</tr>
<tr>
<td>16</td>
<td>SONET/SDH</td>
<td>–</td>
<td>–</td>
<td>610 kB</td>
<td>125 µs</td>
<td>–</td>
<td>ITU-T</td>
</tr>
</tbody>
</table>

**List of acronyms:**

- **WR-OBS**: Wavelength-routed optical burst-switching
- **JET**: Just-enough-time
- **RFD**: Reserved-fixed-duration
- **TAG**: Tell-and-go
- **IBT**: In-band-terminator
- **JIT**: Just-in-time
- **LAUC-VF**: Latest available unused channel with void filling
- **MPLS**: Multi-protocol label switching
- **IPv6**: Internet protocol version 6
- **GbE**: Gigabit Ethernet as defined by IEEE 802.3
- **SONET**: Synchronous optical networks hierarchy
- **SDH**: Synchronous digital hierarchy

| ^ values unspecified  
| ^ values as quoted in [Yoo00]  
| * two values investigated  
| * assumes full wavelength conversion in core nodes  
| * two-way reservation for QoS, one-way reservation for best-effort; burst lengths vary accordingly: best-effort type with < network diameter, connection-oriented ≥ network diameter  
| * burst length identical with IP packet length; min. IPv6 packet length 60 byte, min. GbE packet length 64 byte  

---
OBS JET was defined as a one-way reservation scheme without any acknowledgement of successful (or unsuccessful) burst transmission [Qia00/1]. A two-way reservation scheme has also been proposed (no. 7, [Wei99]), but there it was assumed that bursts are also sent prior to the receipt of an acknowledgement, so at the time of sending it is unclear whether the burst will arrive at its destination. This scheme, termed 'Just-in-Time (JIT)' is explained along with OBS JET in more detail in section 1.2.5.2. Recognizing this as a limitation, schemes [Dol02, Qia00/2, Yoo00] have been proposed to provide class-of-service differentiation by offset times, i.e. to assign larger offsets for higher priority traffic. This, however would have the effect of reducing the burst loss for high priority traffic, at the expense of an increase for lower priority bursts, especially for dynamically varying traffic loads. The result is reduced network capacity for acceptable burst loss rates [Dol01/1], jeopardizing the initial rationale of OBS to increase utilization of resources. In OBS networks with one-way reservation, lower priority bursts experience loss as a penalty. However, given that each burst may contain a large number of TCP/IP (transmission control protocol, internet protocol, [Kes97]) acknowledgements representing up to 50% of traffic [Xio00], each lost burst would affect a number of higher layer connections. So in OBS networks care should be taken to also minimize the loss of lower priority bursts to prevent this. Keeping to predefined latencies is also essential since some protocols such as TCP/IP operate with timeouts and trigger re-sending of packets after a pre-set time-out, even if packets were not lost but held up in a queue within the network. Even deflection routing (see for instance [Mye01]) would bring little benefit since out-of-order burst arrivals will require large buffers in the receiving edge routers for re-ordering. Although buffers for higher layer protocols such as TCP are designed to hold (and re-order) traffic for a full network round-trip time, burst loss would aggravate the problem of re-ordering since a number of TCP sources would potentially re-start transmission, requiring larger buffer sizes than already used. Finally, in all the proposed schemes, wavelengths are assigned on a link-by-link basis, requiring full wavelength conversion at every node, as end-to-end lightpath reservation is difficult because of short offset times and short packets. Hence wavelengths are not used for routing, as in WRONs (section 1.2.1), but simply to increase available transport channel capacity.

The JumpStart approach [Bal02/2], recently proposed, combines the one- and two-way reservation OBS schemes using Generalized Multiprotocol Label Switching (GMPLS, [Yat02]) signaling in the control plane. A recent addition to the family of OBS schemes, it envisages to use a one-way reservation scheme for best-effort type traffic and short
bursts – where the notation of ‘short’ refers to the burst duration is less than a network round-trip time. For traffic volumes exceeding the short burst size, an explicit two-way connection would be established. However, there is no data available on the performance of this scheme in comparison with the previously discussed OBS schemes. The performance analysis of OBS networks with one-way reservation has mostly been limited to investigation of the packet or burst loss probabilities in a single node, e.g. [Tur99, Yoo00, Xio00, Qiao00/1]. OBS network performance analysis to date considering the operation in a network context was carried out for the OBS-JET scheme only [Dol01/2, Mye01], and shortcomings of the one-way reservation scheme are discussed in section 1.2.5.2. Burst loss due to contention in core switches was identified as the main source of performance degradation in OBS networks with one-way reservation [Yoo00, Mye01, Dol01/2] since the burst aggregation delays are small (typically microsecond regime) compared to the propagation delay (typically milliseconds). It was shown, for instance that for the implementation of quality of service (QoS) [Dol01/2] longer bursts in the same class of service (CoS) experienced higher burst loss than shorter bursts. Also, the burst loss probability was load dependent for low priority bursts and could reach values of up 0.1 when operated in a network [Dol01/2].

1.2.5 Timing in dynamic optical networks

Dynamic network architectures (WRONs, OPS and OBS networks) are distinguished by the different timing of packet transmission, which mainly determines network performance in terms of packet delay between the network ingress and egress. The timing of resource provisioning also provides an indication of the speed with which a network architecture can adapt to changes in the input traffic pattern. In combination with the packet loss, the latency plays a critical role for the provisioning of higher-layer applications such as TCP, whose throughput depends on both delay (measured as round-trip time) and packet loss [Lak97].

For the WRON it is assumed that static lightpath are assigned according to traffic demands which are known a-priori. Lightpaths are set up irrespective of the lightpath utilization (which might be low), but provide low delay (propagation delay only) and no packet loss during transmission.
OPS networks offer the finest granularity of all dynamic network architectures, and, therefore, potentially the highest bandwidth utilization. Unlike WRONs, packets in OPS networks do not follow a pre-assigned route through the network, but routing decisions are made on the spot at each intermediate router. So packets between the same source-destination pair may take different routes within the core network, leading to potential packet re-ordering problems at the destination. Packets are buffered at each node to allow for processing of control information and for contention resolution, contributing to the delay. Packets may also be dropped at any point along the path if no sufficient switch resources are available, so OPS networks are prone to both loss and delay.

The main source of performance degradation in OBS networks with one-way reservation, such as OBS JET, is burst loss due to collisions due to insufficient switch resources and the absence of buffering in core nodes [Mye01, Dol01/1]. Bursts are released into the core network after a pre-defined offset time (typically in the microsecond regime), so burst aggregation delays are short, see, for instance, [Xio00]. In the WR-OBS network the burst aggregation delay will have the largest impact on performance since packets are queued considerably longer than in OBS networks with one-way reservation. Delays are expected to be in the millisecond regime, and the trade-off between resource utilization and aggregation delay is investigated in chapter 4. Buffers at the network edge would be designed such that buffer overflow is avoided or does not exceed a predefined value irrespective of the input traffic statistics by rapid adaptation of the burst aggregation time – detailed in chapter 3.

In the following sections, the timing of packet and burst transmission in three networks is discussed: an OPS network, OBS JET as the main type of OBS schemes with one-way reservation, and OBS JIT representing the OBS schemes with one-way reservation and acknowledgement on path set-up. The wavelength-routed optical burst-switching (WR-OBS) scheme with two-way reservation will be discussed in section 1.3.
1.2.5.1 Timing in OPS

Figure 1.7 shows the timing diagram for an OPS network as discussed in section 1.2.2, where the control information (header) is directly followed by the datagram (payload). The length of the packet could be either fixed (as shown here) or variable. The fixed size packet is referred to as a cell (e.g. in ATM). At each intermediate node buffering is required whilst the header is inspected and information extracted to determine the output port for the next hop, and to configure the switch accordingly. More buffering is required for contention resolution as shown in Figure 1.4 (assuming an output-queued switch). Fast optical label extraction and packet forwarding has been demonstrated at a line rate of 100 Gb/s (proof principle for one element, [Tol98]).

![Figure 1.7: Timing chart for optical packet-switching (OPS). The packet is buffered at each node for header processing and switch reconfiguration](image)
1.2.5.2 Timing in OBS JET and OBS JIT

As already mentioned in the context of Figure 1.6 and now detailed in Figure 1.8, the offset time ($t_{\text{offset}}$) between the burst control packet (BCP) and the payload distinguishes optical burst switching from optical packet switching. $t_{\text{offset}}$ is a multiple of the control processing time in each router, setting the switching matrix such that the following burst is transmitted without further buffering – a prerequisite in optical packet switching. Once a BCP enters the router, switching resources from the input to the desired output are reserved ahead of the arrival of the burst, and even when currently a burst is leaving the router on the same output port and wavelength. Wavelength conversion must be included to resolve contention. If even despite wavelength conversion the contention problem cannot be resolved, one of the contending bursts is dropped without further acknowledgement.

![Figure 1.8: Timing chart for OBS Just-Enough-Time (JET), without acknowledgement of successful transmission of datagrams. The control packet (header) preceding the burst configures the switch such that the following payload can pass without being buffered](image-url)
The operation of the Just-in-Time (JIT) scheme as shown in Figure 1.9 is similar to OBS JET. The burst is again sent after a predefined period $t_{offset}$ after sending of the BCP has elapsed. As in the previous case, switching resources are reserved ahead of time and potential contention is resolved by wavelength conversion. The difference to OBS JET is the provisioning of an acknowledgement once the BCP successfully reached the destination edge router. Since the BCP only arrived because switching resources were successfully reserved along the path, the burst is expected to arrive at the destination edge router – but it is the BCP which is acknowledged, not the actual burst. The acknowledgement message travels back to the sending edge router which after approximately one network round-trip time obtains information on successful transmission.

*Figure 1.9: Timing chart for OBS Just-in-Time (JIT) providing an acknowledgement of a successfully received BCP at the destination edge router. When sending the burst it is unclear, however, whether the required resources will be available*
The performance analysis of OBS schemes with one-way reservation was restricted in most cases to only one node, with limited analysis of OBS in a network context as discussed later in this paragraph.

The burst loss in an OBS JET network can be calculated analytically under the following assumptions: Poisson arrival process and negative exponentially distributed burst length, and full wavelength conversion for contention resolution. In this case the burst loss $B$ can be calculated using the Erlang loss formula, based on an $M/M/k/k$ queueing model (an extension to the $M/M/1/k$ model presented in appendix A.2.3):

$$B(\rho, n) = \sum_{i=0}^{n} \frac{(\rho \cdot n)^i}{i!}$$

where $\rho$ is the relative load per channel with $0 \leq \rho \leq 1$, and $n$ is the number of ports and is identical to the number of wavelengths. Figure 1.10 shows the result for the burst loss probability $B$ as a function of relative load and for 4...128 channels. Even in this simplified case, burst loss depends strongly on both relative load and the number of wavelengths used — so burst loss would vary with the switch size and the corresponding number of wavelengths in real networks.

Figure 1.10: Burst loss probability for an OBS JET core node versus relative load and for 4-128 wavelengths, using the Erlang loss formula
As the burst loss probability depends on both the load and the number of wavelengths used within the OBS switch, it would be very difficult to implement applications with strict packet loss requirements. Quality of service (QoS) differentiation was, therefore proposed [Yoo99]: different classes of service (CoS) would be distinguished by their offset time $t_{\text{offset}}$, where higher CoS with lower burst loss requirements were granted larger offsets than lower CoS. Initial results [Yoo99, Yoo00] showed that indeed burst loss could be reduced by orders of magnitude, e.g. by a factor $10^6$ for a system with 4 CoS at load 0.5 (Fig. 2 [Yoo99]). A closer inspection of the results showed [Dol01/2], however, that the CoS differentiation process was load dependent, i.e. the ratio between burst loss for high CoS and low CoS decreased with increasing network load. In addition, the scheme was proven to provide unfairness within a single CoS: for bursts within the same low CoS, longer bursts experienced a higher burst loss than shorter bursts [Dol02/1].

Results for the OBS JIT network have not been investigated in the same depth as those for OBS JET [Wei99, Dol01], but indicate no fundamental difference: [Dol01/1] showed in a comparative study that under the same assumptions for one node OBS JIT performed worse than OBS JET. High burst loss (approximately 0.1 for a relative load of 0.5) was also observed in [Wei99] when simulating the OBS JIT performance over a 5x5 torus network architecture with a Poisson burst arrival process.

The problems of QoS differentiation observed with one-way reservation OBS schemes aggravated when the analysis was extended from one node to cover networks. [Dol01/2] showed that in an OBS JET type network with 2 CoS, the burst loss probability for the higher CoS increased with the number of hops, and was limited to $10^{-4}$ at the last hop, whilst the lower CoS always experienced a burst loss probability of 0.1. The analysis of 3 different network architectures in [Mye01] was carried out under the restrictive assumptions of 1 CoS, fixed burst length, and a burst Poisson arrival process, necessary to limit the computational complexity of the problem. The results showed high burst loss probabilities also, with burst loss for all source-destination pairs $>10^{-3}$ for traffic loads $\geq 0.4$. The introduction of deflection routing could in fact reduce the burst loss probability by more than two orders or magnitude (for a load of 0.4), but simultaneously resulted in the necessity of re-ordering bursts at the receiving end, where the degree of re-ordering was approximately $10^{-3}$, again for a traffic load of 0.4.
[Mye01] is also the only study to date which took into account the number of wavelengths required to support the OBS JET architecture. The comparison with the results for a static WRON showed that OBS JET would provide operational benefits only for traffic loads $\leq 0.3$ – otherwise the OBS JET network required more wavelengths than a static WRON, where no loss and only propagation delay arises between source-destination pairs.

The physical layer of optical networks serves higher-layer applications such as TCP/IP and, therefore, determines their performance in terms of throughput (bandwidth), packet loss, delay and delay variation (jitter), and reliability. The importance of these parameters with respect to each other as well as their values will vary considerably between applications: so may audio applications require low bit rates ($\geq 128$ kbit/s) but low delays (30 ms), whereas data-centric applications such as database access might require much higher bit rates ($\geq 1$ Gb/s), but has more relaxed delay requirements ($< 100$ ms) [Fos99]. The achievable throughput of widely used TCP connections, for instance, has been shown to be sensitive to both the bandwidth-delay product and the packet loss [Lak97]. Hence all three parameters (loss, delay and throughput, plus their interaction) will be considered in the analysis of future dynamic network architectures in chapters 3-5 of this thesis.

Results on the impact of the OBS physical transport layer on the performance of higher layer applications are limited to the investigation of the TCP performance over OBS JET type networks, which was investigated, for example, in [Cha02/3] and [Det02]. [Cha02/3] considered a network with 3 core nodes and 5 edge routers and an end-to-end round-trip time of 1 ms. Burst loss and TCP throughput, however, were analyzed only for the output of one core router output interface with 3 wavelengths operating at OC-12 line rate (622 Mb/s) each, i.e. the maximum traffic volume considered was less than 10 Gb/s. Over the 3 wavelengths 10 label-switched paths (LSP) were established, which contained 4 TCP-Reno and 1 UDP (User datagram protocol [Kes97]) each. The results when using an OBS type scheme showed that the burst loss probability was always higher than $10^{-2}$. A traffic shaping mechanism based on random scrambling of the burst sending times was proposed to reduce this loss rate [Cha02/3] by preventing the synchronization of the sending rate of different TCP sources. With the help of this traffic shaping mechanism it has been shown that burst loss $< 10^{-4}$ could be achieved at a single router interface.
[Det02] investigated a single TCP connection (TCP Reno [Kes97]) running across an OBS network. The impact of burst loss was modeled analytically, and three different source types with fast, medium, and slow sending rates were used for the generation of TCP segments. Burst loss was considered only on the way to the destination (by a fixed value, not varying with traffic intensity), the reverse path (for acknowledgements) was assumed to be lossless – an assumption which will not hold in real networks. In [Det02] it was shown that the TCP sending rate was affected for burst loss probabilities > 10⁻⁴, and that the TCP connection was nearly starved for a burst loss probability of 0.1, down from 850 kb/s to 50 kb/s. The burst loss rate of 0.1 found for low-priority bursts in [Dol01/2] would according to the results in [Det02] be insufficient to support TCP connections, which in current networks account for the majority of the best-effort type traffic.

Taking into account the preliminary findings of [Cha02/3] and [Det02], then a value of 10⁻⁴ appears justified as target for the maximum burst loss probability in OBS networks to support high throughput of TCP connections.

But [Det02] and [Cha02/3] also show the limits of modelling complex communication systems, i.e. higher-applications over a physical network, by the means of Monte Carlo simulation [Jer92]. The networks in both cases were much smaller and carried less traffic than would be expected in real networks (see, for example, Table 5.1 later for 7 real networks). The limitations of simulation networks as shown above, therefore, motivate the use of analytical models. An analytical approach was, for instance, used to calculate performance parameters for the WR-OBS network architecture, and allowed to achieve upper bounds on the maximum resource utilization to be expected.

The wavelength-routed optical burst-switched network as discussed in the following section provides QoS by a tight admission control of bursts entering the network. A burst will only be admitted once the network management ensured that sufficient resources (wavelength) are available for successful completion of the data transfer. The concept of admission control has already been introduced in the context of ATM networks to distinguish different CoS [Kes97]. Resource reservation in the WR-OBS network resembles the constant bit rate (CBR) scheme of ATM networks, where a fixed amount of bandwidth (wavelength) is reserved, but for duration on the same timescale as the network round-trip time (milliseconds) only.
Measurement-based admission control schemes have also been proposed [Bre00, Bar98/2, Bjo96], but exclusively for operation in the electronic domain. The schemes are based on the principle that a new connection (usually referred to as flow) will only be admitted subject to current measurement of network performance in terms of bandwidth and buffer resources, as those need to be reserved for each hop. They originate from the fact that the best-effort approach of transmission in the internet is insufficient for QoS guarantees, and reservation protocols such as RSVP (resource reservation protocol, [Kes97, Whi98]) and DRP (dynamic reservation protocol, [Whi98]) were developed to remove these shortcomings. The underlying measurements and required bandwidth estimations are based on either large deviations theory (for example [Bre00, Bar98]) or the principle of effective bandwidths.

The aim in designing the WR-OBS network and the adaptive burst assembler (details in chapter 3) was to explicitly address the need for QoS reservations. In contrast to approaches such as RSVP or DRP, the WR-OBS network design is driven by the need to implement hard performance guarantees, such as latencies and packet loss, to support a number of different CoS.
1.3 The wavelength-routed optical burst switched (WR-OBS) network architecture

To overcome the shortcomings of both the OPS and the OBS schemes with one-way reservation discussed in the previous section and to investigate whether savings in the wavelength utilization could be achieved in dynamic networks in comparison to WRONS, the WR-OBS architecture was proposed as part of this work, and forms the focus of this thesis. The analysis of the timing of the WR-OBS network architecture is shown in Figure 1.11 to compare this architecture to the previously mentioned OPS, OBS-JET and OBS-JIT architectures (Figs. 1.8 and 1.9).

This architecture requires a two-way reservation between the sending and the network control node (assuming centralized control for simplification), so that the burst is held at the network edge until an explicit acknowledgement from the network control and management is received. The delay for packets during aggregation in the edge router is quantified through the edge delay, $t_{edge}$.

![Figure 1.11: Timing chart for the WR-OBS network architecture where bursts are released subject to a transmission acknowledgement by the network control. The edge delay $t_{edge}$ is counted from the first packet entering the empty queue. The release of resources depends on the burst aggregation scheme used; the limited burst size (LBS [Mig01]) was used here.](image-url)
The edge delay is defined as the elapsed time between the arrival of the first bit of the first packet to the buffer until the entire burst is released into the network. It is envisaged that the bit rate ($b_{in}$) with which the buffer is filled from applications in the access layer will be significantly lower than the core bit rate used for transmission ($b_{core}$). The buffers located in the edge router, therefore, adapt the bit rate of the access layer to the bit rate of the core network. The ratio of $b_{core}/b_{in}$ was identified to be essential for the analysis of network performance parameters in chapter 4. Once a predefined timeout or a buffer filling threshold is reached, a wavelength request is sent to the control node. The related propagation delay is equivalent to half the round-trip time to the control node. If a solution to the routing and wavelength assignment (RWA) problem was found, a control message is prepared carrying the acknowledgement back to the sending edge router, whilst other control messages are broadcasted to the core routers containing information on switch configuration. The release of switching resources depends on the burst aggregation scheme used. Three schemes have been proposed: fixed burst aggregation time (FBAT), limited burst size (LBS), and unlimited burst size (UBS) [Mig01]. The LBS scheme was considered in this work because the burst size is fixed in the acknowledgement, and therefore resources can be released after the period for transmission advertised in the control message is elapsed. The timing is explained in more detail in chapter 4.2.3. The FBAT and LBS schemes provide a high granularity, but require signalling for each individual burst and a burst aggregation delay. These overheads could be avoided when using the UBS scheme, where a connection would exist until an edge buffer is completely depleted ('streaming' of data). The disadvantage related to the UBS scheme, however, is the fact that additional signalling is required to tear down an existing lightpath before it could be reused.

The main research work described in this thesis was concerned with the design and the performance of the WR-OBS network architecture. The term WR-OBS was first introduced in [Due01]. The scheme was specifically designed to address two of the main shortcomings observed with one-way reservation OBS schemes such as OBS JET (as described above), namely to provide QoS guarantees under varying load and input traffic statistics, and to make use of the WDM domain for wavelength-routing. Its analysis aims to explore whether the use of dynamic optical networking allows to achieve improved resource (and hence wavelength) utilization compared to WRONs. WR-OBS combines elements of the previously proposed conventional burst-switching and circuit-switching (WRON) schemes to create a scalable architecture capable of
supporting multiple CoS over the same physical infrastructure, and takes advantage of wavelength-routing. The WR-OBS architecture, the signalling for lightpath establishment, and burst transmission are shown as a sequence in Figs. 1.12-1.14.

It is assumed that electronic packets generated in the access layer arrive at the network edge where they are processed by one of the edge routers optically connected to the core network (Fig. 1.12). Packets are separated into individual queues according to their destination and CoS such that each queue only contains packets heading to the same destination and with the same latency and PLR requirements. The time from the arrival of the first packet into an empty queue until the burst is sent is defined as edge delay, $t_{edge}$, as it was already shown in Figure 1.11.

To establish whether such an architecture would be practical, analysis of different parts of this architecture would be necessary:

It is assumed that the packet processing and buffering would be carried out entirely in the electrical domain due to the limitations of large-scale optical processing and the shortcomings of optical buffering as described in chapter 2, which also deals with the scalability of optical switching matrices as envisaged to be used for the core nodes in the WR-OBS architecture. Bursts would then be mapped to wavelengths, however, direct access to wavelengths in case of high traffic loads from the access layer would also be possible.

Figure 1.12: Principle of the wavelength-routed optical burst switched (WR-OBS) network architecture. Bursts are aggregated at the network edge, and wavelength request packets are sent to the control node for reservation of resources.
Since the traffic statistics in data-centric networks can vary considerably from those in telephone networks [Pax95], particularly the self-similar characteristics of traffic caused concerns with respect to appropriate dimensioning of buffers. This was the rationale for detailed analysis of the burst aggregation and queueing processes under various traffic statistics, as carried out in chapter 3 of this thesis.

For finite buffer lengths as used in most practical systems the packet loss performance is severely impaired by the *burstiness* of the incoming traffic streams. Chapter 3 and related appendices focus on explaining the effects on light- and heavy-tailed traffic statistics on traffic performance, and results are applied to the WR-OBS network architecture. Two different theories, the central limit theorem and the large deviations principle are introduced for mathematical modelling of the effects, together with a relative recent addition, the moderate deviations principle aiming at combining the two. Moderate deviations theory provides an estimate for the buffer overflow probability depending only on the mean and the variance of the incoming traffic, leading to the design of an adaptive burst assembler. It is shown that in the context of WR-OBS networks the adaptive burst assembler provides pre-defined buffer overflow probabilities and even in the presence of self-similar input traffic generates wavelength requests at approximately deterministic rate.

A key aspect in the commissioning of the WR-OBS architecture was the possibility of resource utilisation. Where in static WRONs entire lightpaths are set up independent of the actual demand [Bar97], the WR-OBS architecture envisages the reuse of wavelengths on a ‘as-needed’ basis. Although the actual saving depends also on the actual topology and the dynamic RWA algorithm used, the analytical model developed in chapter 4 shows significant potential savings expressed by the dimensionless metric of the wavelength reuse factor (RUF). Other parameters investigated comprise the wavelength holding time, lightpath utilization and bandwidth-per-wavelength. For the study of performance parameters, the round-trip time ($t_{RTT}$) was defined as the propagation delay of the acknowledgement from the network control node back to the edge router (assuming that a wavelength was reserved), plus the propagation delay of the burst across the network. $t_{RTT}$ was identified as the critical parameter in the network design in this thesis. The analysis of network performance also includes a comparison of the number of wavelengths required for a static WRON with the resource savings achieved when applying WR-OBS.
Once the request arrives at the network control node, it is processed to determine the route and wavelength available for transport within a particular period, as indicated in Figure 1.13, and detailed in chapter 5. The wavelength request packets are first scheduled using a rate monotonic (RM) or earliest deadline first (EDF) algorithm for provisioning of latency guarantees. Subsequently they are forwarded to the DRWA algorithm for lightpath calculation. The network scalability in terms of the numbers of supported edge routers and CoS depends on the processing time $t_c$ per request, and is investigated in detail in chapter 5. On return of the acknowledgement packet to the edge router, the control information is extracted, and a rapidly tuneable laser switched to the appropriate wavelength, as shown in Figure 1.14. The burst is then transmitted on the established end-to-end lightpaths.

*Figure 1.13: Wavelength request processing in the central control node of the WR-OBS architecture. Requests are fed through a 2-stage queueing system before an acknowledgement is returned to the edge router, and control information sent to core routers for configuration.*
The practical implementation of the WR-OBS architecture depends on the feasibility of rapidly changing wavelengths between consecutive bursts. Tuneable lasers, therefore, provide the desired functionality. The results discussed in chapter 6 were obtained in experiments with fast tuneable lasers as light sources, operating across the entire C-band. The experimental part of the investigation contains data on the characterization of the lasers (wavelength, output power and power uniformity, side mode suppression ratio), fast switching and wavelength stabilisation using an optical injection locking technique, as well as BER/Q-factor measurements to determine the expected penalties in dynamic wavelength provisioning over static link design.

1.4 Thesis structure

The remainder of the thesis is organised as follows:

Chapter 2 analyses the difficulties in scaling both purely electronic and purely optical switching architectures to Tbit/s throughputs (optically to 1000x1000 ports) as envisaged for future networks, justifying the WR-OBS architecture.

Chapter 3 presents the analysis of the impact of long-range dependent (LRD) and short-range dependent (SRD) traffic statistics on the queueing performance in OPS and OBS networks. Two different models are used to describe the queueing in both networks. The queueing process in OPS networks can be described by queueing model with continuous depletion of buffer contents, and existing queueing theory can be applied. A novel approach is required, however for the burst aggregation process in OBS networks, where packets are aggregated at the edge and periodically released into the network.
Since data is mainly lost due to buffer overflow at the edge of the OBS network, the related queueing processes are investigated as a function of traffic statistics. The delay which packets experience during aggregation is represented by the edge delay, $t_{\text{edge}}$. The aim of the buffer analysis is to design the queue such that there is no or just pre-defined overflow, and packets are delayed for aggregation ($t_{\text{edge}}$) only. An adaptive burst assembler is, therefore, proposed which is capable of automatically adjusting the edge delay to changes in the traffic statistics whilst guaranteeing a pre-defined overflow probability. The results of this chapter have an impact on both the performance analysis (chapter 4) and network scalability (chapter 5).

Chapter 4 describes the performance analysis of the WR-OBS network architecture and introduces the new metrics of wavelength holding time, lightpath utilization and wavelength reuse factor to quantify the benefits of dynamic network operation.

Chapter 5 covers the analysis and results on the scalability of the WR-OBS network architecture. Efficient scheduling of wavelength requests in the control node of the network is key if hard deadlines of applications are to be met. The modified rate monotonic (RM) and earliest deadline first (EDF) algorithms are applied.

The experimental work on the investigation of fast tuneable lasers for dynamic optical networks is described in Chapter 6, and falls in three parts. The first section is concerned with the experimental implementation of the WR-OBS network architecture. Extensive characterisation of fast tuneable lasers, namely the sampled-grating DBR (SG-DBR) and grating-assisted codirectional coupler with rear sampled-grating reflector (GCSR) lasers was carried out. Of particular interest in the network context are the transient effects occurring during the switching process. For switching times down to 2 $\mu$s these impairments were quantified as a function of burst length in terms of BER, Q-factor and system penalty in the second part of this chapter. The third section of this chapter concerns the use of optical injection locking for fast laser wavelength stabilisation.

Chapter 7 draws the main conclusions of the work, and it is shown how the results presented in this work can be applied to optimize future optical network architectures, and how to quantify the operation regimes which best make use of static or dynamic network architectures. Future directions of research in both theory and experiment are also discussed.
1.5 References


Introduction


G. Tenzer, Deutsche Telekom, “Netze und Netzzugang,” presentation for analysts, Frankfurt, 19th April 2000


Chapter 2 Scalability of processing and switching

In the previous chapter, network growth was investigated and three optical network architectures (WRON, OPS and OBS) were described which were proposed to provide both the capacity and the flexibility to scale with future traffic demands. In section 1.1 it was shown that both volume and the optical capacity-distance product have increased by approximately 100% p.a. for the last decade, leading to the same growth rate of required throughput for routers and switches. The three types of network architectures are mainly distinguished by the switching technology they use in the core nodes. Figs. 1.3-1.5 showed schematically the switch architectures required for the WRON, the OPS and the OBS architecture. The key differences between the switches in these architectures are the switching speed (reconfiguration time) and the processing capacity required.

The OPS switch imposes the strictest requirements since the switch is reconfigured on a packet-by-packet basis – for an average packet length of 389.5 bytes [Xio00] this requires reconfiguration in the nanosecond regime for line rates ≥ 10 Gb/s. As discussed in section 1.2.2, however, optical packets might be fixed in length and, therefore, require, aggregation of short packets for efficient use of the optical bandwidth. The reconfiguration is driven by the processing of packet headers within a period equivalent to one packet length, putting strain on the required processing speed. For instance, 1 Tb/s of traffic across all input ports of the router would be equivalent to an average 320.9 million packets/s when assuming pure packet switching (with an average packet size of 389.5 bytes, [Xio00]), showing the dimension of the problem. The scheduling of switch resources has also been shown to provide an upper bound on the necessary switching speed [Sad00]: due to the constraints in the scheduling of an optical switch with Tb/s throughput, optical switching times faster than 100 ns provided negligible benefit for the switch performance – in [Sad00] represented by the queuing delay. Since due to the technical constraints of optical processing (as detailed below) it is assumed that optical processing and optical buffering will not be able to substitute electronics for the foreseeable future, the first task of this chapter will be to review the scalability of electronic processing.

Providing entire lightpaths in a WRON is conceptually easier than packet-forwarding in OPS, and WRONs require large-scale optical switches with a port count of several hundred when assuming that in future networks several fibres with up to 128 wavelengths each [Gab01, Gab02] are terminated at each core node (see also section 1.2
for details). The reconfiguration times in the millisecond regime may be sufficient when assuming that in *Automatically Switched Optical Networks* (ASON, [ITU]), for instance with a control plane implemented in GMPLS, the provisioning time for a lightpath can take up to 100 milliseconds [Yat02].

The only optical switching technology offering a fully integrated switch fabric to date is the MEMS switch. The scalability of optical MEMS switches was shown for port counts of up to 4000x4000 [Ryf02], and experimentally demonstrated for 1296x1296 [Ryf01]. Despite significant progress in manufacturing of MEMS to these port sizes, rigorous testing will have to be carried out with respect to their reliability before these switches will be allowed into commercial applications. These values would suit the need of WRONs, but MEMS are not sufficiently fast enough for the OPS operation range described above. In this chapter, therefore, the limits on the scalability of fast optical switches were explored, which through technology limitations are not monolithically integrated, but composed of individual 2x2 switching elements [Hun98, Tuc98].

As optical switches are analogue, not digital (as electronic switches), they introduce impairments, in addition to those from transmission. The main concern in optical switching fabrics is the integrity of the optical signal and the accumulation of impairments such as loss and crosstalk. Optical switches, unlike their electrical digital counterparts, are analogue in nature, i.e. they do not provide 3R-regeneration (retiming, reshaping and reamplification of the signal). The second part of this chapter, therefore, analyses the design of large optical switching matrices and investigates scalability limits on the number of required switching elements, the insertion loss, the control complexity in terms of drivers, and the crosstalk for 8 published switch architectures.
2.1 Optical processing and optical buffering

A number of authors have proposed the replacement of electronic control and buffering in optical switches with optical processing and buffering, so eliminating the need for electronics in optical backbone networks completely, see for example [Cha98, Cot98, Tol98]. Experimental demonstration is limited to prove the principle, for instance in [Blu00] the label extraction and replacement for optical packets was carried out using a single nonlinear optical loop mirror (NOLM) to implement a logical exclusive-OR (XOR) operation for header (2.5 Gb/s) erasure and replacement, and routing between two wavelengths for payloads of 40 Gb/s. Optical processing, however, faces the inherent disadvantage of competing with highly-integrated electronic chips.

Optical processing of packet header and label information has been proposed and experimentally demonstrated to operate for line rates up to 100 Gb/s [Tol98] and 40 Gb/s [Blu00, Rau02], respectively. However, the functional blocks such as NOLMs and Terahertz optical asymmetric demultiplexers (TOAD, [Tol98]) consist of individual elements such as semiconductor optical amplifiers (SOA), splitter/combiners and fibre patchcords which cannot be fully integrated to date, so the required surface space per gate is larger than in highly integrated electronic chips. In addition, all active optical components, such as SOAs and modulators, still require electronic driver circuitry.

Fibre Bragg gratings (FBG) have been proposed and experimentally demonstrated for routing in optical networks by selecting packets according to their wavelength in Add/Drop multiplexers and switches, for instance [Riz02, Che98]. Super-structured FBGs are also a key component for the implementation of optical code division multiple access (OCDMA) networks, e.g. [Kit02, Mur02, Teh02], where the FBG holds the code information used as optical address label. A routing decision will be made based on a matching of a given code with the grating. However, all label-based techniques do not allow for easy re-configuration, which is required in the context of packet routing, for example to frequently update address lookup tables.

Optical buffering experiences similar problems as processing with respect to the practical implementation. Fibre delay line (FDL) [Zho98] and recirculating loop buffers have been proposed [Lan96], but both suffer from integration and control limitations:
Scalability of processing and switching

- For Poisson traffic arrival processes and low switch dimensions (4x4 and 8x8) it has been shown that a considerable amount of FDLs (output buffered switch) was required for low cell (fixed packet length) loss rates [Dan98]: under the given conditions, for a cell loss probability \(10^{-6}\), 18 and 48 FDLs are required for traffic loads of 0.4 and 0.8, respectively. The same reference shows that for constant buffer size the admissible load is drastically reduced with increasing burstiness. These results were confirmed by [Hun98], where a switch with large optical buffers (SLOB) was proposed with a buffer depth of several thousand packets.

- Without 3R regeneration the optical signal in a recirculating loop buffer will accumulate noise, limiting the number of recirculations, varying largely with the input power into the loop and the initial extinction ratio of the optical signal [Lan96]. Propagation delays are temperature dependent and therefore introduce jitter when the temperature in the environment changes [Hun98].

- In contrast to electronic buffer, optical memory can only be read out after a pre-defined time given by the propagation delay of the fibre. Nor is it possible to keep packets in the buffer longer than the pre-defined period if this should be necessary. Electronic buffers provide much greater flexibility as they can be flushed at any point in time and also easily accommodate variable packet sizes.

The conclusion from the findings on optical processing and buffering in this section is that both technologies appear not to be suitable to compete with the electronic chips. Even if planar lightwave circuits (PLC) [Oka02] could potentially be used to implement delay lines, they would still require larger size with respect to their functionality than a highly integrated electronic chip. And unlike electronic chips (e.g. field-programmable gate arrays, FPGA), PLCs cannot be reprogrammed or altered. The question whether electronic chips will keep up with the growth in network capacity and how this impacts network design is addressed in the following section.
2.2 Electronic switching and processing

The conclusion of the previous section is that optical processing and optical buffering with available technology are difficult, if not impossible, to integrate in the same manner as electronic chips which due to their high density of logical gates are better suited for processing and buffering than their optical counterparts. The growth in traffic volume, observed at > 100% p.a. over the last decade, as discussed in chapter 1, raised questions with respect to the scalability of electronics itself, and in particular with respect to the mismatch between the growth in optical line rate for TDM systems, currently at 10 Gb/s, and the achievable increase in electronic processing speed governed by what is known as Moore's Law [Moo65]. In this section observations from Moore's Law are combined with analysis of the semiconductor industry [Roa01]. The result is an estimate of the achievable scalability of processing operation speed as required in any large routers, and, most importantly, its impact on future network design.

2.2.1 Fundamentals

The success of electronics in general is based on the possibility of large scale integration of a huge number of gates onto a single chip. The increase of integration has been predicted by Moore's Law as early as 1965, i.e. the increase in the number of electronic gates per die as a function of time [Moo65]. This scaling process was evaluated, and in this section compared to the growth of optical network capacity, since for a number of years it had been assumed that the growth of optical capacity would outpace the scaling of electronics. This would create an "electronic bottleneck," especially in the backplane of large-scale routers required to interconnect line cards with the switching fabric [Aka01, Hav01, Nak01]. The problem was considered serious enough to introduce short-reach (metres) optical interconnects [Bux01] between the line cards and the switching fabric.

2.2.2 Electronic switch fabrics with high throughput (Tb/s)

An additional problem in router design not discussed here in-depth, but contributing to the complexity of building large routers with an electronic switching fabric, is the fact that the throughput of single electronic switching matrices is limited. Although the switch architectures vary in their architecture, they all consist of at least two chip types (for the crossbar and for the implementation of queues, for example) which are interconnected to achieve the desired throughput. For a comparison of different
Scalability of processing and switching

electronic switch architectures see, for example, [Cha02]. The maximum obtainable throughput of commercially electronic switch fabrics to date is 2.56 Tb/s (Agere PI40 [Age02]), whose interconnected crossbar switches have a throughput of 160 Gb/s each, hence a minimum of 16 of those crossbar switches are required to achieve 2.56 Tb/s. The throughput of a single-switch solution is typically much smaller (for example 40 Gb/s for the Agere switch PI-40SAX). Optical interconnection of several such chips has been proposed to create sufficiently large throughput due to the extensive amount of wiring required to exchange data between individual chips within the switch fabric [Aka01, Bux01, Nak01]. Commercially available switch chip sets use an optical interconnection technique termed SERDES (serializer/deserializer [Voi01]), which uses vertical cavity surface emitting lasers (VCSEL) for optical interconnection at 2.5 Gb/s bit rate. An additional problem is fast scheduling and configuration of such a complex switching architecture. A numerical example from [Bux01] shows that for a 64x64 switch with OC-768 port speed (approx. 40 Gb/s, leading to a aggregate capacity of approx. 2.5 Tb/s) a single central scheduler must process 4096 queues (64x64 queues) within 12.9 ns when the switching cell size remains at the currently used value of 64 bytes. A switching cell size of 64 bytes is used, for example in the switch fabric of Cisco routers (referred to as 'Cisco cell'), and it coincides also with the minimum frame size for Ethernet, and is used in Ethernet switches. This cell size leaves approximately 3 ps to deal with each queue, faster even than optical processing at 100 Gb/s. The switch control problem has a striking similarity with the design of a central network scheduler discussed in chapter 5 – the only difference being that the problem in this section would arise in each core node, not only in the central network control node. The difficulties encountered also make a case for using significantly large cell sizes in switches, and correspondingly larger burst sizes in networks to alleviate the processing.

2.2.3 Moore's Law and electronic scaling

Moore's Law [Moo65] predicts that the number of transistors which can be integrated onto a single microchip doubles approximately every 2 years. The drivers for the constant scaling are the increase in die size and the miniaturization of the gate sizes with each microchip generation. Moore's Law for the number of transistors per chip is shown in Figure 2.1 for the period 1970 - 2000[^1], with the linear fit showing a doubling every 2 years over the past 20 years.

The reduction in the gate size of each transistor in turn also enables to operate the individual gates at higher frequencies, thus increasing the clock rate at which the chip operates. The chip performance in terms of number of operations carried out per second increases further. The increase in clock speed in shown in Figure 2.2 for the period 2000 - 2016 [Roa01], combined with the prediction on the number of transistors.

Figure 2.2: Prediction based on Moore's Law as it was shown in Fig. 2.1: Clock frequency (left axis, [Roa01]) and number of transistors per chip (Moore's Law, right axis) forecast for the period 2000 – 2016.
Whereas the number of transistors is expected to grow with a constant rate, the growth of clock speeds slows down from 2006. With clock rates doubling every 2.5 years on average before 2006, it is expected to double only every 4.2 years thereafter. The main reason for the slow-down is the heat dissipation problem (see also 2.2.3).

The overall computational performance of a microchip is given by the number of logical operations carried out per unit time, given by the product of clock rate and number of transistors. To be independent of particular implementations, here the number of gate operations per second is used, as shown in Figure 2.3. For the period 2000 – 2016, the number of operations increases by a factor 2500, or 1.63 per year. For comparison, Figure 2.4 shows the increase in the number of instructions carried out per unit time obtained in real processors [Odl02] in the period 1986-1997. The annual increase factor of 1.60 is very close to the factor 1.63 observed in the number of operations, implying that the number of instructions carried out scales in the same way as the number of operations. Based on this heuristic, in 2010 a single processor would be expected to carry out 344 BIPS (billion instructions per second). In addition to the enhancement of physical performance, further improvements in these numbers could be expected from novel microprocessor designs and simplified instruction sets, resulting in higher processor throughput [Mos01, Pat01].

![Figure 2.3: Number of transistor operations per second for the period 2000 – 2016, prediction based on Moore’s Law with an increase of approximately 63% p.a.](image)
2.2.4 Heat dissipation

The problem of undesired heat dissipation from integrated electronic circuits is a persistent problem in the design of each new generation of microchips, resulting from the switching operation of each gate. This was already acknowledged in a previous section (2.2.3), leading to a slow-down of clock operation speed. The problem aggravates with each new generation as power is proportional to die-area times frequency. The scaling of the related effects for each new chip generation is as follows [Pol02]:

- Frequency increases by factor 2.0
- The supply voltage will scale by factor 0.8
- Active power consumption will scale by 0.9
- The active power density increases by factor 1.3-1.8

Currently the heat dissipation of the chip surface is 40 W/cm², exceeding that of a hot plate (e.g. electrical hot plate as used in laboratories) by a factor of 4. The resulting need for cooling of chipsets in routers, both for processing and routing, limits the density with which those can be stacked in the racks of central offices (CO), limiting the total installed processing power. The operating frequency of future microchips is, therefore, expected to double only every 4.2 years as shown in Figure 2.2.
2.2.5 Electronic memory access times

A third key application for electronic memory, apart from buffering for processing and contention resolution, is to hold the address tables which are required in IP networks for the route lookup and packet forwarding [Cha02]. The processing time of packet headers, therefore, depends on the time required for memory lookups. Assuming a link rate of 10 Gb/s and an average packet size of 389.5 bytes [Xio00], this results in the need to process approximately 3.2 million packets per second. For the average packet size, the total processing time is 312 ns, which is reduced to 32 ns for 40-byte TCP/IP acknowledgements.

Highly integrated electronic random access memory (RAM), typically designed as dynamic RAM (DRAM) with periodic refreshing of the memory contents, has to date memory access times > 50 ns [McK02] – which would be larger than the packet length of a 40-byte acknowledgement at 10 Gb/s. Static RAM (SRAM) provides faster memory access times (10 ns [Cha02], < 1 ns [Nam00]), but is strictly limited in size, e.g. to 1 Mbit for access times less than 1 ns [Nam00]. This restriction causes problems since it might not be possible to hold entire route lookup tables (whose size grew with the increase in traffic volume) in fast memory, and sophisticated strategies to compact information, or to combine SRAM and DRAM are required in future electronic router design [Cha02].

The fact that memory access times decrease at a slower rate than the operating frequency of microchips [McK02] leads to a design challenge in electronic routers since the lookup process slows down over time relative to the operating speed of microchips. Figures available for DRAM memory have shown that the access time for DRAMs decreases by approximately 6.7% p.a. [McK02], whilst in Figure 2.4 the number of operations per microchip increased by 60% p.a. This discrepancy in growth rates may result in a potential electronic bottleneck for address lookups, which could ultimately limit the number of packets which can be processed per unit time and, therefore, limit the router throughput.

2.3 Optical switching

An alternative to complex electronic switching fabrics is optical switching [Mid93], eliminating the need for OEO conversion and fast electronic gate operation. A relatively simple approach to implement an optical switch without active switching elements is the broadcast-and-select switches, e.g. [Bor98, Chi01, Hab98], where the optical signal from input port is equally split and reaches all output ports. A pre-configured tuneable
optical filter in the receiver would select the correct channel on which to receive the packet destined to that particular output port. The simplicity of the approach is offset, however, by the accumulation of loss and crosstalk when scaling the switch to large port counts. For a 100 output ports, for instance, the output power at the selected port drops by 20 dB compared to the input signal into the splitter. A more scalable approach to building optical switching matrices with passive cores would be the use of arrayed-waveguide gratings (AWG) in combination with tuneable sources at the input, as discussed in chapter 6.1.

This section investigates in detail the scalability of optical switching fabrics for the transparent switching of packets, bursts, and lightpaths in the OPS, OBS and WRON network (Figs. 1.3-1.5). As the fabrication of integrated optical switching matrices is complex, switching matrices apart from micromechanical switches are composed of individual 2x2 switching elements, which could, for instance, be based on Mach-Zehnder interferometers (MZI) SOA switches [Tuc99]. The analysis carried out in this chapter does not depend on the actual technology in which the switches are implemented. The only data required are data on the insertion loss, crosstalk and delay of individual 2x2 switching elements from which the values for larger switching matrices are extrapolated.

The node architecture could be significantly simplified by introducing transparent optical switching if only the header information could be extracted and processed electronically. The payload would remain in the optical domain. Despite intense research on principles of optical switching and of optical switch design, the design of large scale (> 100 ports) switching matrices with low loss and low crosstalk levels whilst providing fast switching speeds remains a key problem.

The limit on scalability of optical technology needs to be explored, and in particular that of switches. The analysis in this chapter was carried out taking into account the following parameters:

- Basic architectures
- Nonblocking characteristics
- Number of required switching elements
- Control complexity
- The maximum tolerable loss is defined by the maximum dynamic range for optical amplification, currently approximately 30 dB between two EDFAs.
- Accumulation of crosstalk between switching elements
Except of MEMS switches [Ryf02], optical switching fabrics cannot be monolithically integrated, but are implemented using different technologies, such as the electro-optic, thermal-optic or acousto-optic effects. The 2x2 units are then cascaded to compose the switch architecture. The switching architectures implemented out of 2x2 switches investigated in the course of this chapter are:

- Crossbar [Non84]
- N-Stage Planar [Spa84]
- Double Crossbar [Kon85]
- Benes [Ben85]
- Three-Stage Clos [Clo53]
- Spanke Type 1 and 2 [Spa86]
- Jajszczyk [Jaj93]
- Extended Baseline [Wu97]

2.3.1 Optical switch architectures analysed

The principal architecture of the different switches is shown in Figs. 2.5-13. For the calculations in the following paragraphs it is assumed that all switches are based on 2x2 switching elements (shaded). This is due to the fact that except from MEMS switches it is currently technologically infeasible to integrate a large number of optical switches on the same substrate.

![Figure 2.5: N-stage planar switch architecture [Spa84]](image-url)
Figure 2.6: Standard crossbar switch architecture [Non84]

Figure 2.7: Double-crossbar architecture [Kon85]; it offers a higher contrast ratio and therefore lower crosstalk than the standard crossbar

Figure 2.8: Benes-switch architecture requires the lowest number of elements for its realised [Ben85]
Figure 2.9: Three-stage Clos switch which interconnects several small subnetworks [Clo53]

Figure 2.10: Spanke switch type 1 [Spa86]

Figure 2.11: Spanke switch type 2 [Spa86]
2.3.2 Non-blocking characteristics

The non-blocking characteristic of a switch determines whether any input port can be connected to any outlet of the same switch. Depending on the architecture of the switch, six levels of non-blocking are distinguished, as displayed in Figure 2.14 [Mar94] [Mid93].
The non-blocking characteristic of a given switch reflects its architecture and the control algorithm used to build connections between the inlets and outlets. The switch architectures analysed in this chapter are either rearrangeably non-blocking, wide-sense non-blocking, or strictly non-blocking. The definitions are as follows:

- **Rearrangeably non-blocking**: all permutations are possible but some existing connections may need to be torn down and rearranged to allow a new connection to be added.
- **Wide-sense non-blocking**: an algorithm exists for setting up the paths in a way that guarantees that any future connection can always be made without additional rearrangement of the existing paths.
- **Strictly non-blocking**: every inlet can be connected to every unused outlet without rearrangement regardless of the connecting algorithm.

The non-blocking characteristic of an optical switch refers to its ability to easily build up new connections or rearrange the switching matrix when necessary. Improvements in the non-blocking characteristic simplify the rearrangement of the switch and allow a maximum number of connections with a minimum number of control operations. The number of switching elements, however, is only slightly affected by the non-blocking characteristic as defined in Figure 2.14. Except from the Benes switch, the number of 2x2 elements scales as \(O(N^2)\).

### 2.3.3 Number of required switching elements

Depending on the architecture of the switch, a different number of elements are used to build up the switch. The number of 2x2 elements required is determined as follows [Wu97]:

**Figure 2.14**: Levels of blocking and the determination of non-blocking characteristics for the switches investigated in this chapter. The last level, redundant strictly non-blocking has not been reached yet.
Crossbar: \( N^2 \) \hspace{1cm} (2.1-a)

N-Stage Planar: \( \frac{N \cdot (N - 1)}{2} \) \hspace{1cm} (2.1-b)

Double Crossbar: \( 2 \cdot N^2 \) \hspace{1cm} (2.1-c)

Benes: \( \frac{N}{2} \cdot 2 \cdot \log_2 (N - 1) \) \hspace{1cm} (2.1-d)

Three-Stage Clos: \( 2 \cdot r \cdot m \cdot n + m \cdot r^2 \) \hspace{1cm} (2.1-e)

Spanke Type 1: \( 2 \cdot N \cdot (N - 1) \) \hspace{1cm} (2.1-f)

Spanke Type 2: \( N \cdot (N - 1) \) \hspace{1cm} (2.1-g)

Jajszczyk: \( 2 \cdot N \cdot (N - 1) + \frac{N^2}{4} \) \hspace{1cm} (2.1-h)

Extended Baseline: \( \frac{3}{2} \cdot N^2 - \frac{5}{2} \cdot N \) \hspace{1cm} (2.1-i)

Where: \( N \) : number of input and output ports

\( m \) : number of subnetworks, \( m = 3 \)

\( n \) : number of inlets grouped in a subnetwork, \( n = 2 \)

\( r \) : dimension of subnetwork, \( r = N/2 \)

Except from the Benes network, which shows \( O(N \cdot \log(N)) \), all other networks scale with \( O(N^2) \), which is reflected in the slope of the graphs shown in Figure 2.15. The number of switching elements is an important parameter in terms of manufacturability because the likelihood of a failing switching element increases with the total number of 2x2 switching elements. Reliability is a particular concern in practical applications where failures must either be compensated by spare elements held in reserve, or must be avoided by using very reliable 2x2 switching elements. If, for example, the individual failure probability of a 2x2 element is \( 10^{-6} \), the probability of one of these elements failing in a 100x100 switching matrix increases to approximately \( 10^{-3} \). This indicates the need to integrate spare switching elements within the same matrix for redundancy, or to operate a second matrix of the same type to enable a rapid switch-over in the case of an element failure.
2.3.4 Control complexity

The control complexity determines the number of electrical/electronic drivers required to control the switching elements. From the following equations can be derived that the cross bar, N-stage planar, double crossbar, Benes, and 3-stage Clos architecture require one driver per 2 x 2 switching module, whereas the other architectures require a lower number of driver circuits. This is achieved by using the same driver for two switching elements, so reducing the complexity of the electronic control of the switch. The control complexity can be determined as follows [Wu97] and is plotted in Figure 2.16:

- **Crossbar:** \( N^2 \)  \hspace{2cm} (2.2-a)
- **N-Stage Planar:** \( \frac{N \cdot (N-1)}{2} \)  \hspace{2cm} (2.2-b)
- **Double Crossbar:** \( 2 \cdot N^2 \)  \hspace{2cm} (2.2-c)
- **Benes:** \( \frac{N}{2} \cdot 2 \cdot \log_2(N+1) \)  \hspace{2cm} (2.2-d)
- **Three-Stage Clos:** \( 2 \cdot r \cdot m \cdot n + m \cdot r^2 \)  \hspace{2cm} (2.2-e)
- **Spanke Type 1:** \( 2 \cdot N \cdot \log_2(N) \)  \hspace{2cm} (2.2-f)
- **Spanke Type 2:** \( N \cdot \log_2(N) \)  \hspace{2cm} (2.2-g)
Scalability of processing and switching

Jajszczyk: \[2 \cdot N \cdot \log_2(N)\] \hfill (2.2-h)

Extended Baseline: \[6 \cdot N \cdot \log_2(N) - 4 \cdot N\] \hfill (2.2-i)

Note that number of required drivers scales with either \(O(N^2)\) or \(O(N \cdot \log(N))\). The latter is an attractive feature since it reduces the component count, and an additional benefit of reduced complexity in the driver network.

![Figure 2.16: Control complexity of optical switching architectures, i.e. the number of required drivers/controllers to operate the switching matrix](image)

### 2.3.5 Insertion loss

The insertion loss depends on the material used to fabricate the waveguides in the switching element, e.g. LiNiO₃, and the number of switching elements \(N\) that the optical signal has to traverse during propagation through the switch. The overall insertion loss is composed of the insertion loss \(L\) of every switching element and an additional waveguide-to-fibre coupling loss \(W\). Typical values are \(L = 1\) dB and \(W = 1..2\) dB. To plot results, \(L = 1\) dB and \(W = 2\) dB are assumed to calculate worst case insertion loss.

The insertion loss is determined as follows [Wu97]:

- **Crossbar:** \((2 \cdot N - 1) \cdot L + 2 \cdot W\) \hfill (2.3-a)
- **N-Stage Planar:** \(N \cdot L + 2 \cdot W\) \hfill (2.3-b)
- **Double Crossbar:** \((N + 1) \cdot L + 2 \cdot W\) \hfill (2.3-c)
Scalability of processing and switching

Benes: \((2 \cdot \log_2(N) - 1) \cdot L + 2 \cdot W\) \hspace{1cm} (2.3-d)

Three-Stage Clos: \((2 \cdot n + 2 \cdot m + 2 \cdot r - 3) \cdot L + 6 \cdot W\) \hspace{1cm} (2.3-e)

Spanke Type 1: \((2 \cdot \log_2(N)) \cdot L + 4 \cdot W\) \hspace{1cm} (2.3-f)

Spanke Type 2: \(\log_2(N) \cdot (3 + L) + 2 \cdot W\) \hspace{1cm} (2.3-g)

Jajszczyk: \(6 \cdot (\log_2(N) - 1) + 2 \cdot L + 2 \cdot W\) \hspace{1cm} (2.3-h)

Extended Baseline: \((3 \cdot \log_2(N) - 2) \cdot L + 2 \cdot W\) \hspace{1cm} (2.3-i)

Results are shown in Figure 2.17, and show a large difference in the performance of the different switches. Lowest losses with scaling with N are achieved by switches based on logarithmic scaling, whereas the insertion loss increases dramatically for those switches with linear scaling of insertion loss (in dB).

From a system point of view, a maximum loss of 30 dB is acceptable before the signal must be re-amplified, depending on the signal amplitude and the receiver. This is done under the assumption of providing amplification both at the input and the output of the switch. The additional amplification adds to the costs of the whole switching fabric, but the output of the switch might be a good place for 3R-regeneration of the optical signal anyway. From Figure 2.17 can be derived, that only a few switches (Benes, Spanke and Extended Baseline) meet this requirement assuming a 100 x 100 switch matrix.

This result is important in the context of network design because it limits the ability of building large-scale optical switches which are the core building blocks in a totally transparent optical network. The need for optical amplification contributes to the noise impairments of the optical signal and hence either reduces the total distance over which the signal can be transmitted, or requires the additional use of optical 3R-regenerators.
2.3.6 Crosstalk accumulation – signal-to-noise ratio (SNR)

Crosstalk is a problem especially in large switching arrays, where noise is added at each switching element, degrading the OSNR. As it is not possible to completely isolate neighbouring 2x2 switching elements, crosstalk accumulates along the switching path in the switch. The crosstalk would manifest itself as noise imposed on the signal, reducing the BER as follows [Agr92]:

The optical SNR (OSNR) degraded by crosstalk is not strictly defined and depends on type of measurement [Dao02]. For our purposes it is sufficient to define OSNR as the ratio of signal power with respect to the background noise within the optical switching or filter bandwidth of each channel, for instance given by the channel bandwidth of an arrayed-waveguide grating (AWG).

As BER is a metric measured in the electronic domain, it is necessary to convert the OSNR to the related electrical quantities. The optimum Q-factor as a metric for signal quality is defined as

\[
Q = \frac{I_1 - I_0}{\sigma_1 + \sigma_0},
\]

(2.4)
where $I_1, I_0$ are the receiver currents for the 1 (mark) and 0 (space) levels, and $\sigma_1, \sigma_0$ are the respective standard deviations. The electrical SNR is defined by the mark levels only (defined for active signal only) as

$$\text{SNR} = \frac{I_1}{(\sigma_1)^2}. \quad (2.5)$$

Assuming Gaussian noise levels, the BER is related to the Q-factor by

$$\text{BER} = \frac{1}{2} \cdot \text{erfc} \left( \frac{Q}{\sqrt{2}} \right) = \frac{\exp \left( -\frac{Q^2}{2} \right)}{Q \cdot \sqrt{2 \cdot \pi}}, \quad (2.6)$$

where the approximation is valid for $Q > 3$.

The effects degrading BER are thermal noise in the receiver $\sigma_t$, shot noise of the photodiode $\sigma_s$, and the optical noise $\sigma_{\text{opt}}$, all defined by their root-mean square (rms) values:

$$\sigma_t = \text{thermal noise of the transimpedance and post amplifier} \quad (2.7)$$

$$\sigma_s = 2 \cdot \sqrt{q \cdot R \cdot P_{\text{avg}} \cdot \Delta f} \quad (2.8)$$

$$\sigma_{\text{opt}} = 2 \cdot r_{\text{opt}} \cdot R \cdot P_{\text{avg}} \quad (2.9)$$

where $q$ is the charge of the electron, $R$ the optical-to-electrical conversion factor, $\Delta f$ the electrical bandwidth, and $r_{\text{opt}}$ the inverse OSNR denoting the fluctuation of the optical power level as intensity noise:

$$r_{\text{opt}} = \sqrt{\left( \frac{\Delta P_{\text{opt}}^2}{P_{\text{opt}}} \right)^2} = \text{OSNR}^{-1} \quad (2.10)$$

The respective electrical current $I_1$ is

$$I_1 = R \cdot P_1 = 2 \cdot R \cdot P_{\text{avg}}. \quad (2.11)$$

Where $P_1$ is the power of the mark levels, and $P_{\text{avg}}$ the average received optical power, assuming that marks and spaces are equally distributed. Assuming further that $I_0 = 0$ and noise level $\sigma_0^2 = \sqrt{\sigma_t^2 + \sigma_s^2 + \sigma_{\text{opt}}^2}$ (optical noise on both mark and space), the Q-factor as in eqn. (2.1) is calculated as:

$$Q = \frac{2 \cdot R \cdot P_{\text{avg}}}{\sqrt{\sigma_t^2 + \sigma_s^2 + \sigma_{\text{opt}}^2 + \sqrt{\sigma_t^2 + \sigma_{\text{opt}}^2}}} \quad (2.12)$$

Assuming that the BER degradation is dominated by optical noise $\sigma_{\text{opt}}$ only with $\sigma_{\text{opt}} >> \sigma_t, \sigma_s$, the Q-factor becomes

$$Q_{\text{opt}} \approx \frac{2 \cdot R \cdot P_{\text{avg}}}{2 \cdot \sigma_{\text{opt}}} = \frac{1}{2 \cdot r_{\text{opt}}} = \frac{\text{OSNR}}{2}. \quad (2.13)$$

In this idealized situation, a BER of $10^{-9}$ requires $Q = 6$, so the OSNR in this case must be at least 12 (10.8 dB). Since in a real receiver further degradation by thermal and shot
noise is inevitable, it would be justified to assume a minimum value of 15 dB for the OSNR to maintain a BER of $10^{-9}$.

Assuming an initial extinction ratio of $X^\dagger$ [dB] for a given wavelength between the ‘0’ and the ‘1’ states in the system, the OSNR is calculated as follows [dB] [Wu97]:

- **Crossbar:** $X - 10 \cdot \log_{10} (N - 1)$ (2.14-a)
- **N-Stage Planar:** $X - 10 \cdot \log_{10} (N)$ (2.14-b)
- **Double Crossbar:** $2 \cdot X - 10 \cdot \log_{10} (N - 1)$ (2.14-c)
- **Benes:** $X - 10 \cdot \log_{10} (2 \cdot \log_2 (N) - 1)$ (2.14-d)
- **Three-Stage Clos:** $X - 10 \cdot \log_{10} (n + m + r)$ (2.14-e)
- **Spanke Type 1:** $2 \cdot X - 10 \cdot \log_{10} (\log_2 (N))$ (2.14-f)
- **Spanke Type 2:** $X - 10 \cdot \log_{10} (\log_2 (N))$ (2.14-g)
- **Jajszczyk:** $X - 10 \cdot \log_{10} \left( \frac{N}{2} \right)$ (2.14-h)
- **Extended Baseline:** $X - 10 \cdot \log_{10} (2 \cdot \log(N) - 1)$ (2.14-i)

The respective graphs are shown in Figure 2.18 for an initial crosstalk ratio of $X = 20$ dB. Optical crosstalk is obviously a main source of performance degradation, except from the Benes, the Spanke types and Extended Baseline architectures. This kind of signal distortion requires a 3R-regeneration of the optical signal at a certain OSNR level as described above where a minimum OSNR of 10.8 dB was required to obtain $10^{-9}$ BER. Assuming a relatively low level like 10.8 dB OSNR, this shows that the switch dimension of some switches is tightly restricted, unless internal 3R-regeneration is provided.

\[\dagger\] variable, will be defined by the network and signal format (RZ/NRZ)
2.3.7 Discussion and comparison of optical switching architectures

The properties of a 100 x 100 switch (Table 2.1) and a 1000 x 1000 switch (Table 2.2) are presented. In the columns for system insertion loss and SNR those numbers are highlighted in bold, which indicate a failure in meeting system requirements, and in italic if targets were missed by less than 10%. In this case the exclusion criteria are a system insertion loss > 30 dB requiring optical amplification of the signal, and an OSNR < 15 dB to maintain 10^-9 BER at the output of the switch.

For the 100 x 100 switching fabric, only the Benes and the two Spanke switches meet the requirements to maintain signal quality without additional 3R regeneration inside the switch.

Applying the criteria to a 1000 x 1000 switching fabric, only the two Spanke switches are able to provide the pre-defined insertion loss and OSNR, but they require more than 1 million elements for their realization. The comparison reveals that a low insertion L is critical in particular for large switching matrices and should be aimed to significantly
lower than 1 dB for a single 2 x 2 switching element. The Benes and Extended Baseline architecture require a significantly lower number of switching elements (approx. 10,000) than the Spanke switches and mainly suffer from a low OSNR. Future progress in materials and technology, such as integration in PLCs, could result in a significant improvement of system performance.
### Table 2.1: Summary of results for optical switches of 100 x 100 ports. Bold indicates failure of system requirements (insertion loss < 30 dB, OSNR > 15 dB), italic is used for those which missed the target criteria by less than 10%.

<table>
<thead>
<tr>
<th>Nonblocking characteristic</th>
<th>Number of elements</th>
<th>Control complexity</th>
<th>System insertion loss [dB]</th>
<th>OSNR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crossbar</td>
<td>WSNB</td>
<td>10,000</td>
<td>10,000</td>
<td>205</td>
</tr>
<tr>
<td>N-Stage Planar</td>
<td>RNB</td>
<td>4,950</td>
<td>4,950</td>
<td>105</td>
</tr>
<tr>
<td>Double Crossbar</td>
<td>WSNB</td>
<td>20,000</td>
<td>20,000</td>
<td>106</td>
</tr>
<tr>
<td>Benes</td>
<td>RNB</td>
<td>614</td>
<td>614</td>
<td>16</td>
</tr>
<tr>
<td>Three-Stage Clos</td>
<td>SNB</td>
<td>8,100</td>
<td>8,100</td>
<td>120</td>
</tr>
<tr>
<td>Spanke Type 1</td>
<td>SNB</td>
<td>19,800</td>
<td>1,329</td>
<td>21</td>
</tr>
<tr>
<td>Spanke Type 2</td>
<td>SNB</td>
<td>9,900</td>
<td>664</td>
<td>31</td>
</tr>
<tr>
<td>Jajszczyk</td>
<td>SNB</td>
<td>22,300</td>
<td>1,329</td>
<td>40</td>
</tr>
<tr>
<td>Extended Baseline</td>
<td>SNB</td>
<td>14,750</td>
<td>3,586</td>
<td>22</td>
</tr>
</tbody>
</table>

### Table 2.2: Summary of results for optical switches of 1000 x 1000 ports. Bold indicates failure of system requirements (insertion loss < 30 dB, OSNR < 15 dB), italic is used for those switches which miss the target criteria by less than 10%.

<table>
<thead>
<tr>
<th>Nonblocking characteristic</th>
<th>Number of elements</th>
<th>Control complexity</th>
<th>System insertion loss [dB]</th>
<th>OSNR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crossbar</td>
<td>WSNB</td>
<td>1.00 M</td>
<td>1.00 M</td>
<td>2005</td>
</tr>
<tr>
<td>N-Stage Planar</td>
<td>RNB</td>
<td>0.50 M</td>
<td>0.50 M</td>
<td>1005</td>
</tr>
<tr>
<td>Double Crossbar</td>
<td>WSNB</td>
<td>2.00 M</td>
<td>2.00 M</td>
<td>1006</td>
</tr>
<tr>
<td>Benes</td>
<td>RNB</td>
<td>9.47 k</td>
<td>9.47 k</td>
<td>23</td>
</tr>
<tr>
<td>Three-Stage Clos</td>
<td>SNB</td>
<td>0.76 M</td>
<td>0.76 M</td>
<td>1020</td>
</tr>
<tr>
<td>Spanke Type 1</td>
<td>SNB</td>
<td>2.00 M</td>
<td>19.93 k</td>
<td>28</td>
</tr>
<tr>
<td>Spanke Type 2</td>
<td>SNB</td>
<td>1.00 M</td>
<td>9.97 k</td>
<td>44</td>
</tr>
<tr>
<td>Jajszczyk</td>
<td>SNB</td>
<td>2.25 M</td>
<td>19.93 k</td>
<td>60</td>
</tr>
<tr>
<td>Extended Baseline</td>
<td>SNB</td>
<td>14.75 k</td>
<td>55.80 k</td>
<td>32</td>
</tr>
</tbody>
</table>

List of acronyms:
RNB: Rearrangeably nonblocking
WSNB: Wide-sense nonblocking
SNG: Strictly nonblocking
2.4 Summary and conclusions

This chapter explored the limits on processing and switching, key functionalities for all dynamic network architectures.

It was shown that optical processing has been tested for small scale-routing (2x2) and switching purposes for speeds up to 100 Gb/s. Despite its speed which would allow operation at the optical line rate, optical processing cannot be integrated with current technology in the same manner as transistors on electronic chips. Similar results hold for optical buffering, which either as FDL or recirculating loop buffer never reaches the size nor the performance of electronic RAM chips. Optical processing and optical buffering are, therefore, not considered as potential building blocks of large-scale optical network architectures, in the foreseeable future, and were not analyzed as part of this thesis.

The scalability of electronic switching fabrics is not only limited by the throughput of individual switch chips and the need for optical interconnections within the switch fabrics, but also by software required to reconfigure the entire switch fabric. It was shown that Moore’s Law on the number of transistors per chip will hold also for at least the next decade. The heat dissipation problem, however, is assumed to slow the increase in obtainable processor frequencies. The growth rate of the maximum number of instructions/second per chip over the past decade was 60% p.a. This figure, however, deviates from the growth rate of optical transmission capacity (>100% p.a.) over the same period (Fig. 1.1). Since the transistor count per chip is predicted to follow Moore’s Law, but the growth rate for the operating frequency per chip is forecast to halve from 2005, this finding puts further strain on electronic processing and requires continued effort in parallelizing electronic functionality such as processing. For future network architectures it will, therefore, be mandatory to use the available processing and memory capacities as efficiently as possible.

For the limits in fast optical switching technology, eight switching architectures were analyzed with respect to their blocking characteristic, number of required switching elements (intersections), control complexity (drivers), insertion loss and signal-to-noise ratio (SNR). It was shown that only the three out of eight (Spanke, Benes and Extended Baseline) architectures would scale to 1000x1000 ports.
Taking into account both the merits and the constraints of the different switching technologies, the following switch architecture has emerged from the analysis in this chapter as a possible candidate for implementation in future dynamic optical networks.

Due to the problem of realizing electronic routers with a large port count, electrical buffering and processing are carried out at the network edge, distributed over a larger number of smaller edge routers. These edge routers are connected to the optical core network, with core nodes consisting of optical switches operating at moderate reconfiguration speeds (millisecond to submillisecond) allowing reconfiguration faster than WRONs, but slower than OPS, and eliminate buffering and packet processing in core nodes. This leads to the switching concept which was proposed in the WR-OBS network architecture, whose performance and scalability are investigated in the next 3 chapters.
2.5 References


Scalability of processing and switching


[Non84] -, A nonblocking optical interconnection network using directional couplers, Proc. IEEE Conf. on Global Communications (Globecom’84), 1984, pp. 26.5.1 – 26.5.5

Scalability of processing and switching


Chapter 3 Traffic statistics and adaptive burst aggregation

3.1 Introduction

Buffers, whether electrical or optical, were shown to be an integral part of the switching architectures used in future packet-switched network architectures as discussed in chapter 1. In the context of OPS (Fig. 1.4), optical buffers are located both at the input and the output of the switch, to delay the packet during header processing, and for contention resolution when two packets are competing for the same output port. In OBS networks (Fig. 1.5), electronic buffers are located at the edge of the network to aggregate several packets into a single burst, which is subsequently transmitted through the core network without any buffering. Contention of bursts in the core could be resolved by full wavelength conversion. In the WRON (Fig. 1.3) architecture, no buffering is required in the core network as wavelength channels are provide end-to-end connectivity.

The understanding of the packet (or burst) queueing process in the core network is, therefore, an essential part of any network design, as the queueing process directly determine the performance of applications supported over the network through either delay or packet loss. For instance, the quality of a telephone conversation is significantly impaired for end-to-end delays in excess of 100 ms [Kes97].

The performance of buffers is inevitably connected to the statistical process which describes the arrival process and the length distribution of packets or burst. The discipline which is concerned with the analysis of such queueing problems is referred to as queueing theory, discussed in the following as essential to the results in this chapter:

The queueing theory and terminology based on Markovian models [Pap91, Bol98] has been widely used in the context of telephony networks (described in Appendix A), since the Markov model was found to describe well the nature of telephony traffic and blocking on links. As analytical formulae have been derived for performance parameters such as queueing delay, the Markovian model was initially also adopted for applications in OPS [Dan98], optical buffering [Zho98], and optical burst switching [Yoo99] - Erlang’s loss formula was already introduced in the performance analysis of OBS JET networks in section 1.2.3. Under these assumptions, the average blocking
probability depends solely on the mean values of the call arrival and call holding process, but not on the variance or other statistical characteristics of the traffic process. In the mid-1990s, studies on selected links carrying data traffic showed, however, that the statistical processes describing them were not Markovian (a list of probability density functions, PDF, to describe such processes can be found in Appendix B). [Lel94] showed in an investigation of Ethernet traffic that the traffic showed self-similar characteristics. These findings were confirmed by investigation of data traffic traces on other links [Pax95, Wil97] and also in the analysis of video streams [Gro96]. The notion of self-similarity, mathematically defined in Appendix C, refers to the fact that the traffic trace, if averaged over orders of magnitude of timescale, still shows that same pattern. The Hurst parameter $H$ is a dimensionless metric, often cited in literature, defining the degree of self-similarity with $0.5 \leq H \leq 1$. For $H = 0.5$ a self-similar process is short-range dependent (SRD). A conventional Poisson arrival process falls in this category. A stochastic process is long-range dependent (LRD) for $H > 0.5$ (details see Appendix C), linking the terms of self-similarity and long-range dependence for $H > 0.5$.

Unfortunately there is no agreed definition for the term burstiness, which is related to the covariance structure or higher-order moments of a traffic stream or traffic source. Its meaning varies (see for example [Cal99, Dan98, Wis01]), and it is used even for description of non-LRD processes [Cal99, Dan98].

The finding of traffic traces with self-similar behaviour was important since it could seriously affect the performance of queueing systems, generating packet loss or delays which can be orders of magnitudes higher than in conventional queueing systems [Add98] – an example by way of simulation is presented in section 3.6. In [Lel94] it was already shown that a LRD stochastic process can be generated through the superposition of several individual sources with heavy-tailed distribution (definition Appendix C). A common implementation of LRD traffic [Lik95, Boo02] is the superposition of a number of ON-OFF sources (section 3.3) whose ON and OFF periods are Pareto distributed (Appendix B). The parameter $\alpha$ (with $1 < \alpha < 2$) used to characterize the Pareto distribution relates to the Hurst parameter as

$$H = (3-\alpha)/2,$$  \hspace{1cm} (3.1)

leading to $0.5 < H < 1$, allowing to vary the degree of self-similarity of the traffic sources used.
All queueing results discussed to this point are based on the assumption that queues are emptied at a continuous rate $C$. This can be thought of as the line rate of communication systems, e.g. 10 Gb/s, and describes the rate with which data leaves the buffer. This assumption holds for OPS networks, but not for OBS as discussed below.

Next recent results are summarized on the impact of LRD traffic streams (formed from ON-OFF sources) on the queueing process, how this affects the buffer design in packet networks [Boo02]. Recent literature related to these findings is discussed.

The aim of this chapter, however, is in the extension of conventional queueing theory to cover the queueing process in OBS networks. The main difference to conventional queueing models (as discussed above) is the fact that in OBS networks packets are aggregated into bursts subject to a minimum burst size or maximum aggregation time [Oh02]. But they are not immediately forwarded on the next free slot of the output link, so the assumption of a constant depletion rate does not hold any longer. Two key problems must, therefore, be solved:

1) The first question is concerned with the calculation of the burst size and the burst size distribution for fixed edge delay (equivalent here to the aggregation delay) and as function of properties of the input process, whether heavy- or light-tailed (defined in Appendix C). This problem is addressed analytically.

2) The second, related question is concerned with the calculation of the edge delay for a given buffer size $B$ such that the buffer meets a pre-defined overflow probability irrespective of the input traffic process. Since in real networks memory size is finite, this problem is key for the design in real networks. Results based on a *moderate deviation principle* (defined in Appendix C, see also [Wis01/1]) are applied to propose and test an *adaptive burst assembler*, whose functionality is tested by simulation.

The results of 1) and 2) are general in the sense that they apply to OBS in general. For the WR-OBS network architecture, the results obtained in this chapter are applied twofold:

The precise knowledge about the burst size and burst size distribution as a function of traffic statistics is essential in the context of WR-OBS. The burst and burst size distribution are the input parameters for the network performance analysis carried out in
chapter 4. A chart (Fig. 4.2) visualizes the connection between the queueing analysis in this chapter and the network performance parameters.

For a traffic input process with constant parameters, the adaptive burst assembler would provide a constant value for the edge delay. The inverse, \(1/t_{\text{edge}}\) then denotes the (also constant) frequency with which in the WR-OBS architecture wavelength requests are transmitted between the edge router and the network control node. This significantly simplifies the queueing and scheduling problem of wavelength requests as discussed in chapter 5.
3.2 ON-OFF source description

For traffic generation in this work ON-OFF sources were used since they are versatile, allowing to generate very different traffic processes. They are essential particularly for the generation of LRD traffic [Lik95, Boo02].

ON-OFF sources, shown in Figure 3.1, are used for traffic generation since the probability functions of the ON and OFF period can be independently chosen to model a wide variety of traffic statistics as it appears in packet networks. This is particularly true for the generation for self-similar traffic: asymptotically self-similar traffic with the characteristics of fractional Brownian motion (FBM) traffic is generated by superimposing a large number of sources whose ON and OFF periods are both Pareto distributed [Lik95]. The source continuously alters between the ON and the OFF state; the probability density functions (PDF) for the duration of the ON and the OFF periods can be generally independent, allowing to simulate a wide range of traffic inputs. Appendix B describes the statistical properties of the most widely used discrete and continuous processes used to describe traffic statistics. An output (e.g. a packet, burst or traffic flow) is generated during the ON period, and the source remains idle during the following OFF period before a new output is generated. For the fluid model as considered in this work, the bit rate with which the source sends was chosen to be constant for all ON periods. This correlates with the technical constraint that the line rates in real systems are fixed, e.g. at 2.5 Gb/s or 10 Gb/s.

![Figure 3.1: ON-OFF source for traffic generation of variable length packet and variable length gaps. When active, the source operates with a constant output bit rate b](image)

The timing with a fluid source can be either discrete or continuous, given by the assumptions \( t \in \{\ldots, -1, 0, +1, \ldots\} \) (integer), and \( t \in \{\ldots, -\infty, +\infty\} \), respectively. In the case of the discrete approach, the resulting packet lengths and interarrival times are multiples of the smallest time or data unit, for instance one bit or byte as these are the smallest entities occurring in communication systems. In chapter 3.4 on the simulation results the performance was simulated using discrete sources operating at the resolution...
Traffic statistics and adaptive burst aggregation

of a one byte. The discrete case is appropriate to investigate the performance of individual packets and short queues.

Discrete models applied in the packet generation by individual users (e.g. workstations or TCP connections), whereas continuous models are more appropriate to capture the statistical behaviour of the generated bit rate per unit time, and also in cases when traffic from many sources is multiplexed into a single traffic stream at a higher bit rate. In this work a discrete source was used to reflect the impact that individual packets have on the aggregation process.

The traffic load and the spectral density of the ON-OFF source are a prerequisite for the calculation of the burst size and the burst size variation in OBS networks in section 3.4. The mean load generated by the source is equivalent to the mean bit rate entering the queue. It is important for the calculation of the mean burst size in section 3.4, and therefore, also the basis for the calculation of the network performance parameters in chapter 4. The mean load $\rho$ is a function of the mean of the ON and OFF periods, $E\{P_{ON}\}$ and $E\{P_{OFF}\}$, respectively, and given by:

$$\rho = \frac{E\{P_{ON}\}}{E\{P_{ON}\} + E\{P_{OFF}\}}$$  (3.2)

In the burst aggregation process of OBS networks, as discussed in section 3.4, the statistical properties of the ON and OFF process determine the burst size variance through the autocorrelation function (ACF, [Pap91]) of the source process. The stochastic variations in the ON and OFF processes are characterized by the ACF in the time domain, or by the power spectral density of the source in the frequency domain. For cases where the ACF is difficult or even impossible to be calculated analytically, here a new procedure is suggested, detailed in Appendix D. First, the power spectral density is calculated based on a procedure given in [Lae97]. Next, an inverse Fourier transform is applied to obtain the ACF of the process. This process can be carried out either analytically or numerically using a fast Fourier transform in case an analytical solution is unobtainable. As a result, this procedure allows to calculate the burst size and its variation for a wide variety of traffic statistics – in many cases before this process [Xio00, Oh02] was carried out solely by the means of simulation.
3.3 Queueing systems with small and large buffers for OPS networks

In OPS networks (Fig.1.4) every node is equipped with optical buffers for contention resolution. The packets queueing within the buffers are forwarded to the output link at the line rate, e.g. 10 or 40 Gb/s (equivalent to constant depletion of a queue). The dimensioning of these buffers depends on the traffic input process: For SRD traffic, Markovian models as those described in Appendix A hold, and analytical results are available for delay, buffer overflow and resulting packet loss. Both analytical results and simulation results [Cal98, Dan97] show that under SRD traffic the packet loss rate (PLR) decays exponentially as a function of increasing the length L of the FDL buffer: PLR \( \propto e^{-L} \). Adding wavelength conversion to the FDL buffer helps to reduce the PLR further [Dan97], but the PLR still decays exponentially.

When applying LRD traffic as observed in traces of data links [Lei95] to FDL buffers, however, the decay of the PLR with FDL size is slower than exponential [Lam02], i.e. increasing the buffer size has less effect on the reduction of the PLR as in systems with SRD traffic. Mathematically, the decay in electrical ATM networks for LRD traffic processes is no longer exponential, but algebraic [Lik95]: PLR \( \propto L^{-m} \), with \( 0 < m < 1 \), a large number ON-OFF sources with heavy-tailed (Pareto) ON periods to generate the LRD traffic trace, and for large buffer length \( L \to \infty \). These results showed that under LRD traffic input buffer lengths would have to be increased by orders of magnitude to achieve the same PLR (or overflow probability) as for SRD input. This effect was the main cause for the concern that LRD traffic caused with respect to the buffer design in OPS networks. A second problem with respect to LRD traffic was that analytical results were (if at all) obtainable only when the models were scaled to the limit. In [Lik95] the results are precise only for \( L \to \infty \), and for a large (ideally infinite) number of traffic sources. The lack of analytical models still is a major obstacle in the OPS network design process when considering LRD input traffic statistics.

A closer inspection of the results in [Lik95] shows that the sub-exponential decay of the PLR was caused by the heavy-tailed traffic in the ON periods, but that the distribution of the OFF periods contributed through its mean value only. So the statistical properties of the interarrival time were negligible compared to the properties of distributions for the active periods, which dominate the queue performance. More systematic study of the effects of varying ON and OFF traffic distributions on the buffer overflow process
and as a function of the buffer lengths was carried out over the last years. [Boo02] provides the most comprehensive summary of results to date, and their impact on buffer scaling in OPS networks. The results listed in Table 3.1 are for the buffer overflow probability, here also referred to as loss curve I(b), where b is the buffer size for a single traffic source. The mathematical derivation of the figures given in Table 3.1 is provided in Appendix C.

<table>
<thead>
<tr>
<th>Light-tailed ON</th>
<th>Light-tailed OFF</th>
<th>Light-tailed ON</th>
<th>Light-tailed OFF</th>
<th>Heavy-tailed ON</th>
<th>Heavy-tailed OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small buffer</td>
<td>I(b) (\propto \sqrt{b} + O(b))</td>
<td>I(b) (\propto b)</td>
<td>I(b) (\propto \sqrt{b} + O(b))</td>
<td>I(b) (\propto b)</td>
<td>I(b) (\propto \text{sublin})</td>
</tr>
<tr>
<td>Large buffer</td>
<td>I(b) (\propto b)</td>
<td>I(b) (\propto \text{sublin})</td>
<td>I(b) (\propto \sqrt{b} + O(b))</td>
<td>I(b) (\propto b)</td>
<td>I(b) (\propto \text{sublin})</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of the loss curve I(b) scaling as a function of the buffer size and the ON and OFF traffic characteristics [Boo02]. The sublinear regime refers to functions of the form \(\ln(b)\) or \(b^x\) with \(0 < x < 1\).

The results in Table 3.1 were obtained under the following assumptions:

- A large deviations principle (LDP, definition Appendix C and [Wis01/1]) was assumed. The LDP is a part of queueing theory which is particularly useful for the investigation of buffers with LRD input. The LDP is based on the theorem of the most likely trajectory to overflow, i.e. it considers only the most likely of all rare events which causes the buffer to overflow (as a rare but large burst would overflow a buffer). An analogy from a field other than communications may help to clarify the effect: The LDP was used in [Boo02] also for the prediction of insurance claims of rare events with potentially a high damage. Predictions on the damage caused by earthquakes would serve as an example.

- The loss curve I(b) is normalized to the scaling parameter n (number of sources superimposed):

\[
I(b) = - \lim_{n \to \infty} \frac{1}{n} \ln \{p_n(b, c)\}
\]  

(3.3)

where \(p_n(b, c)\) is the steady-state probability that the buffer exceeds the level \(n \cdot b\), i.e. the overflow probability in the presence of n sources.

- The results can be grouped for the small and large buffer regime (defined in Appendix C), and for light-tailed and heavy-tailed traffic statistics of the ON and
OFF periods. The term small in this context refers to buffers which quickly overflow (for one traffic source this would be equivalent to a buffer size holding only a few packets), whilst the term large describes (mathematically) the limit of \( I(b) \) for \( b \to \infty \). An example showing these limits is provided in the Appendix (C.5.2) when using the Pareto distribution to describe the heavy-tailed process.

In more general terms, a buffer could be considered to be small if its size was considerably smaller than the bandwidth-delay product of the application that it serves. According to [Lak97], the bandwidth-delay product is (not strictly, though) defined as the product of the round-trip time delay and the minimum bit rate available (bottleneck link). A large buffer would then be defined as a buffer which can hold at least the traffic volume accumulating over one network round-trip time.

N.B. The term round-trip time in this context refers to the end-to-end delay for an application (e.g. TCP end-to-end connection), whilst the notion round-trip time is used in chapter 4 to describe the time for which a wavelength is not used for transmission. This includes the time for the acknowledgement to travel from the control node back to the edge router, plus the initial propagation delay of the burst transmission.

The small buffer limit applies to queueing systems with short but strict delays (e.g. voice traffic in ATM), where packets must leave quickly to meet latency requirements. A large buffer would not be beneficial to the application, since it could not prevent packets from violating their latency requirements even if they were buffered.

The key result from the small buffer limit is fact that the buffer overflow probability scales in the same manner irrespective of the ON and OFF traffic statistics used. This result is important since it shows that for the limit of short buffers, LRD traffic has no impact on the buffer performance. Also, the queueing process can be described by the results achieved for SRD input, the overflow probability only depends on the mean values for the ON and OFF process [Boo02]. Hence, the ON and OFF processes light-tailed, i.e. a Markov chain model as in Appendix A holds.

N.B. The large deviations principle here results in a buffer scaling \( I(b) \approx b^{0.5} + O(b) \). The linear term \( O(b) \) will eventually take over the scaling such that the decay eventually reaches \( I(b) \approx b \) (as in the large buffer regime and as earlier stated in this paragraph).

The results in [Cao02] demonstrate by detailed analysis the practical importance of the small buffer limit. In this reference it was shown that the queueing process for a short buffer could be described well by the conventional Poisson model when a large number
of traffic traces were multiplexed, in line with previously discussed results in [Boo02].
An example used in [Cao02] provides some figures on the size of a small buffer with
respect to the arrival process: all traffic arrivals had a mean arrival rate of 5.45 cells
(unit length) per unit time, whilst the buffer depth was just in the range of 1...100 cells.
If one cell was equivalent to an ATM cell with 53-byte length, the buffer sizes
considered here are in the kB regime (for one source).

The third column in Table 3.1 shows that the results for the scaling in the large buffer
limit solely depend on the distribution of the ON periods. In communication networks
the ON period is equivalent to distribution of packets or bit rate. For light-tailed ON
periods, the loss curve scales exponentially, \( \lambda(b) \propto b \) – the queueing process would be
sufficiently described by conventional queueing theory. Only for heavy-tailed ON
periods the loss curve scales sublinear – which according to equation (3.3) results in a
sub-exponential decay of the overflow probability. These results confirm the previously
quoted findings of [Lik95], which already showed for the particular case of Pareto-
distributed ON periods that the packet loss rate decays sub-exponentially. The
interarrival time (OFF period) distributions only contribute through their mean, and
their distribution can be neglected when multiplexing a large number of sources. For
practical applications this means that the distribution of the interarrival times do not
affect the queueing performance when a large a number of traffic traces are multiplexed.
This would be important, for instance when investigating user behaviour, where it
would be important to monitor only the statistical properties of the active periods as
those would impact the network performance (e.g. throughput).

The results for the short and large buffer regime as just discussed hold for all traffic
statistics which were proposed to date to be used in the context of communication
networks. These results are qualitative in nature, but lead to the important conclusion
that LRD traffic only affects the buffer and subsequently the network design when the
packet distribution on individual links follows a heavy-tailed distribution.
If it is possible to further limit the assumptions on the traffic sources, as shown in the
following paragraph, a numerically more accurate model can be developed to obtain
quantitative results for the buffer overflow probability and PLR.

In any communication system an important constraint applies: either the bit rate is
limited by the maximum line rate, e.g. 10 Gb/s, or the packet sizes generated by a user
Traffic statistics and adaptive burst aggregation

are limited by the protocols used (e.g. 64kB for IP, 1536 bytes for Ethernet). The results in the previous paragraph were derived under the assumption that for heavy-tailed traffic the values for the packet size or bit rate would be unbounded, for instance when using a Pareto distribution. When limiting even heavy-tailed distributions to an upper limit, this has serious, but beneficial implications on the properties of the corresponding queueing process (Likhanov-Mazumdar estimate, [Lik98]). Under this condition it has been shown that even for heavy-tailed (with upper limit) ON periods the PLR and the overflow probability decay exponentially when a large number of traffic sources are multiplexed and that multiplexing approximately 100 traffic streams was sufficient to fulfill this condition. [Lik98] also proves explicitly that the decay for both PLR and overflow probability becomes sub-exponential when the stochastic functions used to model the ON and OFF distributions are unbounded, in accordance with results from Table 3.1. The details on the Likhanov-Mazumdar estimate can be found in Appendix E. This result particularly applies to the design of OPS networks where buffers are located in each core router, and overprovisioning of memory to cope with LRD traffic would significantly add to the cost of the router (plus space for memory modules). Based on the assumption that currently each line card in electronic routers is equipped with memory designed to hold 100 ms worth of traffic [Cha02], this would account for buffer extension in the range of several Gbytes. The results obtained from the Likhanov-Mazumdar estimate, however, show that this is avoidable subject to the assumption that traffic entering core routers is already highly multiplexed – to reach a bit rate of 1 Gb/s, for instance, it would be necessary to groom the traffic of 1,000 users each provided with 1 Mb/s network access, typical of DSL (digital subscriber line) connections [Reu01].

In conclusion, it was shown in this section that LRD traffic processes in OPS networks can have a detrimental effect on queueing systems through sub-exponential decay of PLR and overflow probability with increasing buffer size. In a network context this would result in the need of significantly larger buffers to counteract the effect. The main impact of LRD traffic, however, is limited to large buffers (in the limit b→∞) where the active periods of the traffic trace (packets or bit rates) follow a heavy-tailed distribution. The Likhanov-Mazumdar estimate proves that multiplexing gain is obtainable in the practically relevant case of bounded traffic generating processes. For the case of short buffers the queueing process is entirely insensitive to the traffic processes used, and the overflow probability decays exponentially as in conventional queueing systems –
Traffic statistics and adaptive burst aggregation

showing that significant multiplexing gain can be achieved. This result is important for the network design since scaling buffer sizes with $O(n)$, where $n$ is the number of superimposed sources, is more than sufficient to achieve low packet loss, ultimately limiting the required amount of memory to be installed in large routers.

### 3.4 Analysis of buffers with periodic emptying - OBS

After a review on the queueing process in OPS networks where buffers are emptied at a constant rate, we now move on to investigate the queueing process for burst aggregation in OBS networks. The fundamental difference with respect to results presented in the previous section is that packets are aggregated before being released into the network. This makes the models used in the previous section to describe the buffering process in OPS networks inappropriate for the OBS context. This section is, therefore, concerned with the mathematical description of the burst aggregation process in OBS networks, also under the assumption of LRD and SRD traffic inputs as for OPS.

The burst aggregation process for OBS has been investigated previously with LRD traffic statistics (see, for example [Xio00, Ge99, Oh02]), but solely by means of simulation and for microsecond timescales. These are several orders of magnitude lower than the delays expected in the WR-OBS architecture (milliseconds). Since it was shown in the previous section that for queueing systems with constant depletion the detrimental effect of LRD traffic would manifest itself for large buffers, it would be necessary to investigate the effect of large buffers (as in WR-OBS) for the burst aggregation under LRD traffic input, particularly since buffers of that size had not been considered in previous studies.

The need for a thorough theoretical understanding of burst aggregation, in addition to pure simulation, is highlighted by resolving the puzzle of contradictory conclusions derived from two studies on burst aggregation in OBS JET type networks (as defined in section 1.2.3) under LRD inputs. It is mentioned here, since it is related to the key question whether the queueing of traffic traces could, in fact, reduce their degree of self-similarity in OBS networks. This problem had not been addressed before.

Next an analytical model is presented which allows to calculate the mean and the variance of the aggregated burst as function of the ON OFF source traffic characteristics – this has not been done before. This is complemented by simulations which were carried out to determine by how much the buffer length used for aggregation is reduced.
Traffic statistics and adaptive burst aggregation

as a function of LRD and SRD dependent traffic statistics. These results are then used for the introduction of the adaptive burst assembler in the final section of this chapter.

Although [Ge00] claimed that the burst aggregation resulted in a reduction of the degree of self-similarity (decrease of Hurst parameter $H$ from 0.9 to 0.74 and 0.65), [Xio00] claimed that no such effect was observed (Hurst parameter $H = 0.8 = \text{const.}$). The latter would comply with theoretically obtained results from queueing theory [Wis01/1] which found that $H$ is unchanged after passing through a queue. The reduction in the degree of self-similarity would be equivalent to decreasing the long-range dependence traffic characteristics, and hence the distribution would ultimately tend towards SRD statistics.

The misleading statement in [Ge00] could be explained through an insufficient analysis of the output traffic trace which appeared to have a reduced degree of self-similarity. [Xio00] found that when determining $H$ over very short periods (10 $\mu$s, equivalent to an average of 25 packets, the same range as in [Ge00]), the Hurst parameter $H$ in fact was reduced, even to values as low as 0.5. It was only when the period over which LRD was investigated was significantly prolonged (up to 1 ms) that the original degree of self-similarity ($H = 0.8$) manifested itself again. This effect was attributed to the fact that the traffic processes used were asymptotically self-similar only (asymptotically here refers to the observation interval tending to infinity), so showing significant degree of LRD when observed over timescales orders of magnitude larger than the actual burst aggregation delay. In summary, the Hurst parameter did not change on the longer timescales for which self-similarity was observed before.

3.4.1 The analytical description of burst size and variation

It is assumed for all OBS networks that buffers are not emptied continuously at rate $C$, but only periodically when bursts are released into the core network after a given edge delay required for aggregating packets into a burst. This requires novel modelling to understand the statistical properties of the burst size such as distribution, mean and variance as a function of edge delay and traffic statistics.

The following constraints specifically apply to WR-OBS networks investigated in this thesis. Due to the round-trip times in wide area networks, edge delays in the range 5...50 ms must be considered. At a mean bit rate of 10 Gb/s, the average burst size becomes 100 – 500 Mbit (approx. 12 and 60 MB) accordingly. For an average packet size of 389.5 bytes observed in measured traffic streams [Xio00], the average number of packets per burst in the WR-OBS architecture becomes high, 34,000 – 170,000 for the
values chosen, which is much larger than for the shorter bursts in OBS JET – [Xio00] assumed a mean burst size of 5...15 kB.

We now derive analytical expressions for the mean burst size and the burst size \( \mu_s \) and the burst variance \( \sigma_s^2 \). For the following calculations an infinite buffer size was assumed. The input to this buffer is a stochastic process \( x(t) \), reflecting the workload (number of bytes) transmitted to the queue per unit time. For consistency with the ON-OFF source model presented in section 3.2, \( x(t) \) is assumed to be discrete. This also reflects that practically the minimum resolution of packet size is one byte.

We start with the integral that reflects the burst size \( s \) aggregated over period \([b-a]\) when accumulating traffic arriving as random process \( x(t) \).

\[
s = \int_a^b x(t) \cdot dt
\]  
(3.4)

Interpreting the above integral as a Riemann integral (note that \( x(t) \) shows random fluctuations) the mean value of the burst size \( s \), \( \mu_s \), is given as follows:

\[
\mu_s = E\{s\} = \int_a^b E\{x(t)\} \cdot dt = \int_a^b \mu(t) \cdot dt.
\]  
(3.5)

The RV \( s^2 \) is required to determine the burst size variance \( \sigma_s^2 \) in equation (3.8):

\[
s^2 = \int_a^b \int_a^b x(t_1) \cdot x(t_2) \cdot dt_1 dt_2,
\]  
(3.6)

Applying the mean operator \( E\{.\} \) on this integral is equivalent to introducing the ACF \( R(t_1, t_2) \) of the traffic process \( x(t) \) into the integral:

\[
E\{s^2\} = \int_a^b \int_a^b E\{x(t_1) \cdot x(t_2)\} \cdot dt_1 dt_2 = \int_a^b \int_a^b R(t_1, t_2) \cdot dt_1 dt_2.
\]  
(3.7)

Finally, the burst size variance becomes:

\[
\sigma_s^2 = E\{s^2\} - \mu_s^2
\]  
(3.8)

If the process \( x(t) \) is stationary, \( \sigma_s^2 \) can also be calculated using the autocovariance function \( C(t_1, t_2) \) [Pap91]:

\[
\sigma_s^2 = E\{|s - \mu_s|^2\} = \int_a^b \int_a^b C(t_1, t_2) \cdot dt_1 \cdot dt_2.
\]  
(3.9)

For symmetrical limits equation (3.4) can be re-written as

\[
s = \int_{-T}^T x(t) \cdot dt.
\]  
(3.10)
Assuming that the process $x(t)$ is stationary, its variance becomes [Pap91]:

$$
\sigma_s^2 = \int_{-T}^{T} \int_{-T}^{T} C(t_1 - t_2) \cdot dt_1 \cdot dt_2 = \int_{-2T}^{2T} \left(2 \cdot T - |\tau|\right) \cdot C(\tau) \cdot d\tau 
$$

(3.11)

where $C(\tau) = R(\tau) - \mu^2_s$ is the covariance to $R(\tau)$. The edge delay, $t_{edge}$, is the timing parameter related to the burst aggregation process in WR-OBS networks. With $T = t_{edge}/2$ equation (3.11) becomes:

$$
\sigma_s^2 = \int_{-t_{edge}}^{t_{edge}} (t_{edge} - |\tau|) \cdot C(\tau) \cdot d\tau .
$$

(3.12)

Key to solving equation (3.12) for the burst size variance is the knowledge about the ACF or the covariance of the traffic process which feeds the buffer. Exact values on the covariance structure may be hard to obtain when several ON-OFF sources are used. For the case of a single discrete ON-OFF source, however, the corresponding ACF and covariance were derived in section 3.2 (with details in Appendix D), allowing to determine the burst size and the burst size variance analytically as a function of the distribution of the ON and OFF periods. In Appendix D an example is given how to apply the results of this section in practice.

The characterization of the burst size distribution solely by its mean and variance are an essential prerequisite for the analytical modelling of the burst aggregation process. In the following section it is shown by way of simulation that the burst size distribution can be well approximated by a normal distribution – and the results of this chapter would then be used to provide the mean and the variance for this normal distribution.

### 3.4.2 Simulation results

After deriving an analytical expression for mean and the variance of the burst size in equations (3.5) and (3.12), respectively, now the burst aggregation process is investigated by way of simulation. First, the PLR is determined as function of the edge delay and various traffic statistics, in order to understand how LRD traffic statistics will reduce the usable buffer length for burst aggregation. This is important in the WR-OBS networks context since for a constant buffer size, a reduced usable buffer length translates into a higher burst sending rate, i.e. on average shorter bursts have to be sent, or sent with a higher frequency. This leads to a higher rate of wavelength requests to the network control node (chapter 5). Hence, techniques to use the buffer at the network edge as efficiently as possible are key in the network design process.
It is also shown that the burst size distribution for long aggregation delays, as in the WR-OBS network, can be approximated by a normal function even for heavy-tailed traffic inputs. This significantly simplifies the analytical modelling as the mean and variance of the burst size, as derived in the previous section, can be used directly for the characterization of the burst aggregation process.

The traffic statistics of the arriving packets has to have a significant impact on the burst aggregation and buffering and the resultant network performance, and most importantly, the edge delay, $t_{\text{edge}}$ and the packet loss rate, PLR. To analyze this, the following simulations were carried out using a single FIFO queue. The incoming traffic was generated using an ON-OFF source, as detailed in section 3.2, at the input of the edge router with independent probability density functions (PDF) for the ON-state, $P_{\text{ON}}$, and the OFF-state, $P_{\text{OFF}}$. Call arrivals and call holding were modeled by Poisson inter-arrival time and exponential call holding times, but it was discussed earlier in this chapter that this model is inappropriate for the description of data traffic [Lel94, Pax95]. Simulations using a single source were carried out for different scenarios to calculate the distribution of the burst size, $L_{\text{burst}}$, and the resulting PLR for a finite length buffer [Due02]. The PDFs applied for traffic modeling included Pareto, Poisson, and fixed packet length and packet inter-arrival time distributions. A minimum packet length of 50 bytes, approximately the size of a short IP packet (40 bytes IPv4, 60 bytes IPv6) or an ATM cell (53 bytes) was assumed, in combination with a buffer size $B = 400$ Mbit (47.7 MB) for an average input bit-rate $b_{\text{in}} = 10$ Gb/s into a single buffer. The distribution of addresses is uniform.

### 3.4.3 Fragmentation

To reduce the required header processing, future networks might operate with packets which are significantly longer than minimum IP packet size. The decreased granularity or *fragmentation* will typically be determined by the network protocol and the optimum level of fragmentation requires analysis. Here, different levels of packet fragmentation were modelled, and the results shown for values ranging from 50 bytes to 5 kB. The low value corresponds to the current data networks, in which 40 byte TCP/IP acknowledgements account for more than 50% of the total traffic. Longer packets, however, may simplify the processing and forwarding functions, and future applications for data transfer or multimedia applications may make use of longer IP packets which map to the minimum packet lengths up to 5 kB.
3.4.4 Edge router simulation results

Figure 3.2 shows the resultant burst size distribution as a function of $t_{\text{edge}}$, and resulting PLR - for a minimum packet size of 5 kB with $b_{m} = 10$ Gb/s and an average load of 0.1 (i.e. max. access buffer bandwidth 100 Gb/s) for the following packet and inter-arrival time distributions:

I) Pareto ($\alpha = 1.5$) packet length distribution, Pareto ($\alpha = 1.5$) inter-arrival time distribution  
II) Fixed length packet sizes, Pareto inter-arrival time distribution ($\alpha = 1.5$)  
III) Fixed length packet sizes, Poisson inter-arrival time distribution

![Figure 3.2: Simulation results for burst size and PLR as a function of $t_{\text{edge}}$ and a mean input bit-rate $b_{in} = 10$ Gb/s. The burst size increase is plotted for an infinitely large buffer size, but PLR for a finite buffer size $B = 400$ Mbit (47.7 MB). Bars indicate 95% confidence level for Pareto distributed packets and interarrival times, with minimum packet size of 5 kB.](image)

In all three cases, it can be seen that the mean burst size increases linearly. However, the burst size distribution for a given $t_{\text{edge}}$ does not follow the same behaviour. For burst statistics on finite timescales (as for a given $t_{\text{edge}}$) a possible measure of the burstiness is the variance of the burst size distribution. Packet loss variation according to the burstiness of the input traffic for cases I - III is shown by the PLR curves in Figure 3.2. As a blocking-free core network, and, therefore, no packet loss in the core was assumed throughout this paper, packet loss refers to those packets lost due to buffer overflow in edge routers. The largest deviation of the burst size distribution for a given $t_{\text{edge}}$ was observed for case I, indicated by bars for a distribution with 95% confidence level. For the calculation of the burst size distribution an infinite buffer size was assumed. For the calculation of the PLR, the buffer size then was bounded to $B = 400$ Mbit (47.7 MB).
For finite simulation time, an average PLR of $10^{-6}$ was reached for edge delays of 27.5, 31.5 and 38 ms for cases I - III. The results can be compared to the case of a continuous bit-rate (CBR; also referred to as the fluid traffic model) with an achievable edge delay of 40 ms before packet loss occurs. The application of a CBR traffic model allows the development of an analytical model independent from the actual traffic statistics, which can be applied to derive bounds for parameters. The PLR graphs in Figure 3.2 show that traffic with heavy-tailed packet and interarrival time distribution significantly reduces the maximum allowable $t_{\text{edge}}$. To meet a specific PLR, e.g. $10^{-6}$, the maximum allowable $t_{\text{edge}}$ before releasing a burst would be < 28 ms. This finding is important for all applications and network services whose quality is determined by a low PLR, such as voice transmission. It is emphasized that the value of 40 ms in this example is the upper limit before packet loss occurs; for time critical applications the buffer can be emptied at a faster rate, whilst best-effort type traffic would experience longer edge delays, after which there is no further delay except the propagation time.

More detailed analysis of the variation of the PLR for case I is shown in Figure 3.3. Increase in PLR is observed for $t_{\text{edge}} = 28.8$ ms. Figure 3.4 (plotted for the same values) shows the probability $P$ for the PLR to exceed a given threshold $X$ for $t_{\text{edge}}$ varying in the range from 28.8 ms – 41.6 ms.

*Figure 3.3: Deviations of the PLR for Pareto packet, Pareto inter-arrival time statistics ($\alpha = 1.5$) for $B = 400$ Mbit (47.7 MB) and minimum packet size of 5 kB. For clarity, frequency values > 1000 were omitted. The results show that the heavy-tailed input traffic statistics result in a significant distribution of the PLR. In Fig. 3.2 only the mean PLR was plotted. Selected traces are re-plotted in Fig. 3.4*
A comparison with the average PLR in Figure 3.2 shows that for an edge delay of 33.6 ms, for which an average PLR of $3.83 \times 10^{-3}$ was calculated, PLR > 0.08 appears with a probability of 1%. The results signify that the variation in the PLR must be taken into account to accurately characterize the QoS of a lightpath.

The effects of packet fragmentation on the PLR and maximum allowable edge delay were also analyzed by using the same statistics as in case I, but with a minimum packet length of 50 bytes, 0.5 and 5 kB, respectively. The resultant PLRs are shown in Figure 3.5 and result in maximum allowable $t_{edge}$ of 27.5, 34 and 36 ms, respectively, achieving mean PLR < $10^{-6}$.

The same figure shows that for aggregation of packets over timescales significantly longer than the packet length that the burst size distribution can be approximated by a normal distribution. For the assumed buffer size of 400 Mbit (47.7 MB), the PLR values derived from an approximation with a normal distribution are in good agreement with those obtained from simulation for minimum packet sizes ≤ 5 kB. The result that the burst size distribution is normal is important as it proves that the central limit theorem can be applied, simplifying the modelling of the burst aggregation process over timescales. These results extend previous analysis of shorter aggregation times in OBS-JET type networks, where in [Xio00] the burst sizes were claimed, but not formally proven, to tend to a normal distribution.
This result significantly simplifies the analysis and enables the scaling of the mean and the variance of the burst size with the edge delay, for the basic stochastic processes for the packet length and packet inter-arrival time. With the burst size distribution approaching a normal distribution with mean and variance \( \{ \mu_s, \sigma_s^2 \} \) from section 3.4.1, the PLR can then be explicitly calculated using the error-function [Pap91] for a given buffer size \( B \).

![Figure 3.5: Simulation results for the PLR as a function of the edge delay for \( B = 400 \text{ Mbit (47.7 MB)} \) and a mean input bit-rate \( b_{in} = 10 \text{ Gb/s} \), for different levels of packet size, 5 kB (dash), 0.5 kB (dot) and 50 bytes (dash-dot). The PLR calculated from burst size distribution (assumed normal) is shown by solid line.](image)

3.5 **Adaptive burst assembler**

From the simulation results of the previous section (Fig. 3.2) it can be seen that, in order to achieve a particular packet loss rate, the respective edge delay (aggregation time) varies as a function of the input process. For applications a precise knowledge of this aggregation time is, therefore, key for guaranteeing performance. The burst assembler process can operate in two different modes: In the first instance, bursts would be aggregated for a time known a-priori, the edge delay. This is relevant in all latency-sensitive applications, where the edge delay is adjusted such that pre-defined deadlines are met irrespective of the buffer filling. In the second mode, a burst would only be send when a particular buffer filling ratio was reached to maximize the use of available resources. The latency requirements of best-effort type applications such as datagram...
transmission are relaxed, at least compared to voice transmission, and would allow for more flexible times for emptying buffers at the network edge.

It is necessary to consider the queueing problems at the network edge since in the WR-OBS network architecture this is the place where packet loss occurs. The admission control ensures that bursts enter the network only when the required resources (lightpaths) are available. Data is lost at the network edge when the capacity in the core network is exhausted and newly arriving packets cannot be admitted to the buffer as this is already full. The assumption here is that in all real systems buffers are of finite size, especially since for the WR-OBS scheme burst sizes on the order of Megabytes are assumed. The related performance metrics are buffer overflow, and packet loss rate, respectively. For the burst aggregation process it is key to design the burst aggregation process such that pre-defined packet loss or buffer overflow probabilities are met. Although the packet loss rate is a more precise performance measure than the overflow probability, it can often not be quantified, except for given parameters and under specific assumptions to the traffic models as in [Lik98]. In queueing theory the overflow probability is almost exclusively used as a metric. As discussed in section 3.4, for queueing theory, the more relevant problem is the scaling of buffer overflow with increasing buffer size and number of traffic flows applied, e.g. whether it is necessary to increase the buffer size linearly for a linear increase in the number of traffic flows applied. The more accessible metric is the overflow probability usually considered in queueing theory, which will also be used as metric in the context of burst assembly. An additional advantage of adjusting the burst size and the burst aggregation time is the possibility of smoothing the distribution of burst requests. In chapter 1 it was shown that these requests were required for lightpath reservation. A smooth, and preferably deterministic, arrival of burst requests to the network management and the control nodes reduces the probability of lightpath requests being rejected, as will be shown later in chapter 5.

Based on moderate deviations theory (as in Appendix C) the overflow probability \( P(L_B > B) \) for the buffer under investigation is given as \([Wis01/1]\):

\[
P(L_B > B) = \frac{e^{-\gamma}}{\sqrt{\pi \cdot \gamma}}
\]  

(3.13)

where the parameter \( \gamma \) is derived as the metric for buffer overflow and defined as:

\[
\gamma = - \ln(\text{overflow probability}) = \frac{(B - \mu \cdot t)^2}{2 \cdot V_t}
\]  

(3.14)
Traffic statistics and adaptive burst aggregation

with buffer size $B$, mean bit rate $\mu$, aggregation time $t$ and variance $V_t$ measured over period $t$. $V_t$ depends on the traffic process fed into the queue and can only be assumed to be a linear function of $t$ when the traffic process is Gaussian, $V_t = \sigma^2 \cdot t$, where $\sigma^2$ is the variance of the traffic process per unit time. In this case, $\gamma$ becomes:

$$\gamma = \frac{(B - \mu \cdot t)^2}{2 \cdot \sigma^2 \cdot t}$$

(3.15)

The traffic process is characterized by its mean and variance only, not by higher-order moments as required by large deviations theory, similarly to the estimates for heavy traffic theory and applicability of the central limit theorem. The estimate does not require a full characterization of the incoming traffic process and, therefore, significantly simplifies the traffic characterization. The estimate is applicable to a wide range of statistics as only their means and variances determine the queueing process.

The result for the overflow probability was then used to design a novel feedback control loop to control the burst assembly process as shown in Figure 3.6. The system is designed such that a specified overflow probability $P_{\text{target}}$ is maintained by adjusting the edge delay $t_{\text{edge}}$ according to the traffic mean and variance measured during the assembly of the preceding burst. One burst assembly cycle is the minimum time to react to traffic changes since the parameters do not vary after the start of the aggregation process.

The moderate deviations estimate is based on the mean and the variance only, but neglects higher-order moments. This is important for practical implementation since the mean and the variance are relatively easy to obtain. In addition, these values must be frequently updated - milliseconds for the WR-OBS architecture, so the time for processing the incoming data stream and characterize it with respect to its statistical properties is limited.

![Figure 3.6: Feedback control mechanism for burst assembly based on measurement of the traffic mean and variance](image)

Figure 3.6: Feedback control mechanism for burst assembly based on measurement of the traffic mean and variance
Depending on the load and traffic source model it was shown that the results obtained through the moderate deviations estimate could deviate by more than one order of magnitude from the actual value [Wis01/1]. However, the moderate deviations estimate is the only estimate available to date which covers all traffic types and load ranges. Since the buffer overflow might not be sufficiently suppressed, the moderate deviations estimate (3.15) is combined with an empirical correction of the measured overflow. So the moderate deviations estimate of \( t_{edge} \) is refined by the actually measured traffic statistics as detailed in the description of the algorithm below. The code for the control algorithm looks as follows:

\[
\begin{align*}
t_{edge\_old} &= t_{edge}; \\
if (P_{ovflw} < P_{target}) \\
\{ \\
\quad t_{edge} &= \frac{B + \gamma_{\text{ovflw}}(1-\mu)}{\mu} \cdot \frac{1}{\mu} \cdot \sqrt{\gamma_{\text{ovflw}} \cdot (1-\mu)} \cdot (2B + \gamma_{\text{ovflw}} \cdot (1-\mu)) \\
\} \\
t_{edge} &= \omega \cdot t_{edge} + (1-\omega) \cdot t_{edge\_old}; \\
if (w > B) \\
\{ \\
\quad t_{edge} &= (1 - \sqrt{P_{ovflw}}) \cdot \frac{B}{\mu} \\
\}
\end{align*}
\]

The algorithm can be described in more detail as follows:

1) The previous value of \( t_{edge} \) is copied into a new variable to be used with the exponentially weighted moving average.

2) If the overflow probability (\( P_{ovflw} \)) is below the specified overflow probability (\( P_{target} \)), the moderate deviations estimate is applied to calculate \( t_{edge} \).

3) An exponentially weighted moving average is used to provide memory and to smooth fluctuations and adjust \( t_{edge} \). The parameter \( \omega \) is usually chosen such that \( 0.05 \leq \omega \leq 0.2 \).

4) If buffer overflow occurs (workload \( w \) > buffer size \( B \)), the edge delay is adjusted, with the adjustment weighed by the square root of the overflow probability. This ensures that for larger overflow probabilities the aggregation time is significantly reduced, since the overflow most probably was induced by a load change. This empirical approach proved to be a reliable way of adjusting to buffer overflow for all traffic statistics applied.
Figure 3.7: Aggregation delay and burst size for overflow probability of $10^{-4}$ for a buffer size of 4 Mbit. The mean bit rate was increased in a step-like manner from 20 to 27, and 27 to 32 Gb/s after 300,000 and 600,000 simulations. Inset: fit of the burst size distribution with a normal distribution.

The feedback control loop was tested with heavy-tailed input traffic, the results of which are shown in Figure 3.7 in terms of aggregation delay and burst size. The aggregate of four sources with Pareto ON and OFF periods was applied to a buffer of size 4 Mbit. The parameters of the Pareto distribution were chosen as $\alpha = 1.5$, $A = 320$ bit – a minimum packet size of 320 bits corresponds to the shortest 40-byte TCP/IP acknowledgements observed in real networks. The distribution was finally truncated for values exceeding 12,000 bits, the equivalent of a maximum Ethernet packet of approximately 1,500 bytes. A smaller buffer than in previous examples (400 Mbit) was chosen to ease the computational effort. To model an example of traffic volatility, the mean bit rate of the traffic entering the queue was increased from 20 to 27 Gb/s, and 27 to 32 Gb/s, respectively, after 300,000 and 600,000 simulations for a system with a maximum input bit rate of 40 Gb/s (multiplexed from 4x10 Gb/s). The target overflow probability of $10^{-4}$ was obtained in all cases except for abrupt changes of the load: Spikes in the burst size plot indicate the short-term overflow of the buffer which is corrected within one burst cycle. It can be seen that the resultant aggregation delay (< 200 $\mu$s for the above parameters) and its control are stable over long windows and are able to adjust to a new traffic load within one cycle.
The burst size distribution during the last third of the simulation (32 Gb/s mean bit rate) is shown as inset to Figure 3.7. It can be well approximated with a normal fit with mean 3.86 Mbit and standard deviation 34 kbit – proving that the buffer space is efficiently used.

The burst aggregation time can be regarded as nearly constant during the last third of the simulation, with a mean value of 123.5 μs and a standard deviation of 0.5 μs. These values also prove that it is possible to obtain a very narrow, in fact nearly deterministic distribution of the aggregation time, which also allows for a nearly constant burst sending time and burst frequency. The value $1/t_{\text{edge}}$ can be considered as the desired burst frequency $[t_{\text{Wis01/2}}]$, indicating how often the buffer needs emptying.

The significance of the desired burst frequency as a result of a thorough edge burst aggregation design is key for the efficient operation of the control node in chapter 5. As shown there, the incoming requests can be scheduled more efficiently if the request sending rate is constant, and this will increase the throughput of the network control node. Ultimately the network architecture scalability increases (more nodes can be accommodated), and performance requirements, especially latency, are easier to meet than for a completely random arrival of wavelength requests to the control node.

### 3.6 Summary and conclusions

For all OBS architectures the design of the burst aggregation process is critical to avoid delays and packet loss at the network edge. This is particularly important for the WR-OBS network architecture where the main sources of degradation are the queueing delay and packet loss due to overflow at the network edge. The correct application of queueing theory is, therefore, key in the analysis and subsequent design of dynamically operated optical networks.

The contributions of this chapter are as follows:

First we introduce the characteristics of ON-OFF sources, which due to their versatility are frequently used as input sources in the investigation of queueing process, and particularly to generate LRD traffic streams.

The impact on LRD traffic was first determined for OPS networks. It was shown that the detrimental effect on the buffer overflow probability is limited to the case when the active periods (i.e. packets) of a traffic trace show a heavy-tailed distribution, and the buffer size is large. Otherwise the results obtained with SRD traffic hold. A buffer is considered to be large if its size was equal or larger than the bandwidth-delay product of
the application it serves. Accordingly, a buffer is regarded to be small if the buffer size is considerably smaller than the bandwidth-delay product. The qualitative analysis of the effect of SRD and LRD traffic was complemented by the more accurate Likhanov-Mazumdar estimates for packet loss rate and overflow probability. The significance of the results is that through the achievable multiplexing gain the required increase in memory to be installed in routers can be limited to scale with \( O(n) \) (or even less according to [Cao02]) even for LRD traffic, where \( n \) is the number multiplexed traffic traces. The small buffer limit introduced in this context would apply to queueing systems which, in the case of ATM cells (53 bytes) provide buffer in the kB regime for a single traffic source.

For OBS networks a new analytical model was developed to calculate the mean and the variance of the burst size when aggregated at the network edge. Linked with the simulation results for WR-OBS, this model allows to analytically calculate the PLR at the edge since the burst size can be modeled by a normal distribution. The simulations further showed the impact of heavy-tailed input traffic on the PLR and the reduction in usable buffer space, leading to the design of a novel adaptive burst assembler.

The adaptive burst assembler provides an accurate estimate for the edge delay on millisecond timescale as a function of the buffer size and the input traffic statistics. The mean and variance of the incoming traffic were shown to be sufficient to provide an estimate for the buffer overflow probability, hence the approach could be implemented in networks where these parameters must be frequently (milliseconds) measured and updated. The adaptive burst assembler is a key instrument not only to provide buffer performance guarantees at the network edge, e.g. overflow probability, but it also generates wavelength requests in the WR-OBS network at a constant rate. The benefits of this for the network design process are discussed in the next paragraph.

The results achieved in this chapter are used in the two next chapters of this thesis. The performance analysis (chapter 4) is based on the values for the mean of the traffic arrival process. The results obtained for the adaptive burst assembler are used in chapter 5, where the wavelength request generation process provides an input to the central network control and management.

Queueing theory under heavy-tailed traffic input is still a topic of mathematical research, and more results are expected to appear in the near future. From the point of view of communication networks the lack of explicit numerical results not only for the buffer overflow probability, but also packet loss rate for finite buffer systems remain open research areas.
3.7 References


Chapter 4 Performance analysis of the WR-OBS network architecture

4.1 Introduction

The aim of this chapter is to analyze the conditions under which dynamic WR-OBS (Figs. 1.12-1.14) would bring significant operational advantages over a much simpler but less adaptable quasi-static logical fully-meshed WRON (Fig. 1.3). In all OBS network architectures with one-way reservation (Fig. 1.5) wavelengths are solely used as transmission channels, but the wavelength-routing capability offered by the WDM network is not exploited. It was in the context of the WR-OBS architecture that for the first time the concept of burst switching was combined with dynamic wavelength allocation, first shown in [Due00]. To the best of our knowledge, there is to date only one other group known to have begun working on the combined problems of OBS and RWA [Wan02].

The goal of this chapter is to quantify upper bounds for the achievable network parameters, namely the edge delay, bandwidth utilization, wavelength reuse and round-trip time, and give design rules on the speed requirements for dynamic routing and wavelength assignment algorithms to make a dynamic core network practical. This work is closely related to the discussion of optical switching speeds (chapter 2) and the network control and management (chapter 5), which is concerned with finding a fast solution to the RWA problem. The results obtained in this chapter can then be applied to optimize the design rules of future optical network architectures and quantify the operation regimes which best make use of the static or dynamic network architectures.

The chapter starts with a summary on the relevant features of the edge router and edge aggregation process as relevant to the performance analysis of the WR-OBS network. Timing constants relevant to the burst aggregation, signaling and burst transmission process are introduced, linking the burst aggregation process, the wavelength request signalling, and network performance. The core of this chapter focuses on the performance results: an upper bound is obtained for a continuous bit rate model, whilst the deviations of these parameters are derived as a function of the burst size distribution, and their implications for the for network design are investigated.
4.2 Network architecture and edge router model

4.2.1 Network and edge router architecture assumptions

The edge router and burst aggregation process were analysed in the previous chapter, and the results on the burst aggregation process provide the input for the performance analysis of the WR-OBS network architecture carried out in this chapter. This chapter focuses on how the burst aggregation time in combination with the average input bit rate into the buffer impacts the network performance parameters, and which savings in resources are available through dynamic network operation when compared to a static WRON. For the analysis here it is sufficient to consider the mean arrival rate of traffic to the buffer located at the edge of the network (continuous bit rate [CBR] traffic). The mean arrival rate is the product of mean load $\rho$ (eqn. 3.2) and the line rate at which the buffer is connected to the incoming traffic. The electronic switch located in the edge router provides statistical multiplexing under the assumption of an uniform destination address distribution. The buffers are dimensioned to store data arriving on the order of 100 MB. This value is typical of conventional electronic routers which are designed to hold about 100 ms worth of traffic per line card [Cha02], confirmed by the values specified for line cards in commercial routers (e.g. Cisco 12416 router with 128 MB upgradable to 256 MB).

In this chapter, first the network control layer is described, especially with respect to the amount of control information required to operate the network and size of control packets, followed by a detailed description of the timing of the burst sending process, building on the results previously discussed in chapter 1.3. The relation between the burst aggregation process and the network performance parameters is then visualized in Figure 4.2. The novel performance parameters introduced to quantify the benefits of dynamic wavelength allocation are the wavelength holding time, bandwidth-per-wavelength, lightpath utilization and wavelength reuse. The wavelength reuse is the key metric in quantifying the wavelength savings compared to a WRON. The level of achievable savings depends significantly on the edge delay, the round-trip time, and the ratio of the core bit rate to the input bit rate into the buffer, highlighting the strong link between the queue design and the network performance.
4.2.2 Network control and management layer

In the proposed architecture, several edge routers are connected to one optical core router. This simplifies the scalability problem that electronic routers face when their throughput is scaled to the Tb/s regime. Assuming a network with \( N = 100 \) edge routers and a core network with \( M = 20 \) core routers, this results in add/drop traffic of the order of \( 5 \) Tb/s per core router for a logical fully-meshed architecture with a maximum bit rate per lightpath of 10 Gb/s (input bit rate to the buffer, \( b_{in} \)), and a potentially higher bit rate for the optical core (core bit rate, \( b_{core} \)), as already mentioned in section 1.3. Buffers at the edge separate the different speeds of the access and the core layer, and the ratio \( b_{core}/b_{in} \) plays an important role in the analytical description of network performance, as shown in chapter 4.3. The buffers at the edge adapt. Since wavelengths in the WR-OBS architecture can be shared between different edge routers, it is expected that the number of required tuneable lasers at the edge is significantly smaller than in the WRON, and those savings would be expected be on the same order as the wavelength reuse factor (RUF, section 4.3.5). The WR-OBS architecture would, therefore, allow to drastically reduce the component count, or alternatively allow to gradually increase of the installed capacity in line with growing demands.

OBS networks also require new or improved components for dynamic network operation, namely fast tuneable lasers operating in the C-band for DWDM applications [Cha02], and burst-mode receivers with large dynamic range (\( > 10 \) dB) and fast clock and data recovery units for bit rates in excess of 10 Gb/s. This part of the work is described separately in chapter 6.

The out-of-band signaling on a separate wavelength requires one control packet per request and acknowledgement, plus several control messages to be sent to core nodes, in the case of switch re-configuration. Control packets contain information about the origin and destination addresses, the CoS, and the amount of data accumulated when the request was sent to estimate the traffic arrival rate. Assuming control packets of maximum 250 bits (chapter 4.2.3) to be sent every 10 ms, for a network with \( M = 100 \) edge routers and 3 CoS, the capacity of the control network would total to approximately 1.5 Gb/s, or 3 million requests and acknowledgements per second. The requests then have to be processed by the RWA algorithm in the central node, and scheduled for QoS provisioning. Details addressing both problems can be found in [Mig01, Koz02], and further investigation on the implications of different RWA algorithms on the network scalability can be found in chapter 5. The results presented
here are for an ideal RWA – in real systems a trade-off between processing speed and blocking probability must be found, and the number of required wavelengths also becomes dependent on the network topology.

Different burst aggregation mechanisms for WR-OBS have been studied in [Mig01]. The analysis in this thesis focuses on the limited-burst-size scheme (LBS): The control node estimates the traffic arrival rate from the packets accumulated when the request packet is sent, and hence the burst size is limited to the value as it is sent in the acknowledgement. With this approach it is possible to avoid the release phase required for bursts of unknown size. For bursts of unknown size a special control message would be needed which is sent to the control node when the burst transmission is finished, to free up resources. This messaging would require at least the propagation delay between the sending core router and the control node (milliseconds). One example where this occurs are ATM networks, which require a set-up and release phase for each connection. In contrast, the LBS scheme avoids this additional overhead, allowing it to operate more efficiently.

4.2.3 Timing diagram for burst aggregation and transmission

The burst aggregation process and most important timings are summarised in Figure 4.1. The burst aggregation cycle can be described as follows. $t_{edge}$, as defined in section 1.3, is the elapsed time between the time of the arrival of the first bit of the first packet to the buffer queue until the entire burst is released into the network. The arriving packets are aggregated in the buffer until triggered either by a threshold indicating potential buffer overflow or a timeout signal for delay-sensitive data, as discussed in chapter 3. This occurs when the wavelength request signalling packet is sent to the control node. The propagation delay for this control packet is $t_{prop,sig}$.

It is assumed that the signalling packet contains information on the source and destination edge routers, the CoS, and the quantity of data in the buffer, required to estimate the wavelength holding time $t_{wht}$, defined as the time necessary to empty the buffer and transmit the data between edge routers.
The following example formats of control packets and bursts to be used in the WR-OBS network architecture are proposed:

1. **Wavelength request (199 bits, 25 bytes):**
   - Request identifier (3 digits for edge router, followed by a unique number from 1-10000 assigned by the router [successfully finished requests/transmission numbers go back into a number pool to be reused], 7+14 = 21 bits)
   - Sender (1024 edge routers, 10 bits)
   - Destination (1024 edge routers, 10 bits)
   - (latest) time to start transmission: hh:mm:ss:ms:μs (5+6+6+10+10 = 39 bit)
   - burst length [bytes] (64 GB, 36 bits)
   - CoS (8, 3 bits)
   - traffic mean arrival rate (bit/s) (40 bits) and standard deviation (bit/s) (40 bits)

2. **Acknowledgement (123 bits, 16 bytes):**
   - Request identifier (21 bits)
   - Sending edge router (1024 edge routers, 10 bits)
   - receiving edge router (1024 edge routers, 10 bits)
   - wavelength (100, 7 bits)
   - time to start transmission: hh:mm:ss:ms:μs (5+6+6+10+10 = 39 bit)
   - burst length/transmission duration [bytes] (64 GB, 36 bits)

3. **Refused connection (80 bits, 10 bytes):**
   - Request identifier (21 bits)
   - Sending edge router (10 bits)
   - receiving edge router (10 bits)
   - time for re-trial of transmission: hh:mm:ss:ms:μs (5+6+6+10+10 = 39 bit)
4. **Control message to optical crossconnects (216 bits, 27 bytes):** Control message broadcasted to all routers after request granted, containing a list of all routers along the path and the respective input/output port combinations to be used.
   - Crossconnect identifier (100, 7 bits)
   - Crossconnect input port (1000, 10 bits)
   - Crossconnect output port (1000, 10 bits)
   - Assuming a maximum of 8 hops in large networks [Bar97], the above number of bits must be multiplied accordingly, resulting in a total of 216 bits

5. **Burst header format (<1,000 bits, burst in the MB regime):**
   - Preamble for synchronization (few hundred bits, to be determined by technology used; answers provided by experiments in chapter 6)
   - Burst (former request) identifier (3 digits for edge router, followed by a unique number from 1-10000 assigned by the router [successful request numbers go back into pool], 7+14 = 21 bits)
   - Burst length/transmission duration [bytes] (64 GB, 36 bits)
   - FEC for protection and allowing relaxed dynamic range and jitter specs on the transmission path (strongly dependent on coding scheme used)

It should be noted that the length of the control information packets is smaller than TCP/IP header information (40- and 60-bytes for IPv4 and IPv6, respectively). The burst header format size is dominated by the need for synchronization of each burst. Taking into account a minimum clock recovery time of 40 ns as shown in the HORNET project [Shr00], this accounts for a synchronization period equivalent to 400 bits at 10 Gb/s.
### Figure 4.1: Timing diagram showing key timing parameters in the burst aggregation and burst transmission processes (limited burst size scheme, LBS). The timing constants are defined in more detail in Table 4.1 below.

<table>
<thead>
<tr>
<th>Timing constant</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{edge}}$</td>
<td>max. delay, i.e. the time the first packet spends in the buffer before the burst is released into the network</td>
</tr>
<tr>
<td>$t_{\text{prop,sig}}$</td>
<td>propagation delay for a control packet sent from the edge router to the network control for wavelength reservation</td>
</tr>
<tr>
<td>$t_{\text{proc}}$</td>
<td>processing time, i.e. time between arrival of control packet and decision on lightpath and wavelength</td>
</tr>
<tr>
<td>$t_{\text{prop,ack}}$</td>
<td>propagation delay between the sending and receiving edge router for sending the acknowledgement for a wavelength reservation $t_{\text{prop,ack}} = t_{\text{prop,sig}}$ assuming the control packets take the same route between the sending edge router and the control node</td>
</tr>
<tr>
<td>$t_{\text{WHT}}$</td>
<td>wavelength-holding time, i.e. the total time for which a wavelength is set up</td>
</tr>
<tr>
<td>$t_{\text{prop,net}}$</td>
<td>propagation delay for signal travelling from sending to receiving edge router across the core network</td>
</tr>
<tr>
<td>$t_{\text{trans}} = L_{\text{burst}}/b_{\text{core}}$</td>
<td>transmission time of the burst</td>
</tr>
<tr>
<td>$t_{\text{RTT}} = t_{\text{prop,ack}} + t_{\text{prop,net}}$</td>
<td>Time during which the acknowledgement is sent and before the first packet arrives at the receiving edge router</td>
</tr>
</tbody>
</table>

**N.B.** The definition of the term *round-trip time* deviates from the one used in IP networks [Kes97], which refers to the round-trip for end-to-end applications.

*Table 4.1: Definition of timing parameters used for the description of the WR-OBS architecture*
Processing the wavelength request requires time $t_{\text{proc}}$, followed by an acknowledgement packet to be returned to the requesting edge router, with an additional delay $t_{\text{prop,ack}}$. Concurrently with the transmission of $t_{\text{prop,ack}}$, a wavelength channel is reserved, setting the start of $t_{\text{WHT}}$. In parallel, the burst aggregation continues until an acknowledgement from the control node of a confirmed wavelength reservation is received. In this chapter we assume that the burst assembly terminates at the point the acknowledgement packet from the controller reaches the edge router, although alternative schemes have also been analyzed [Mig01]. This allows the burst aggregation to continue in parallel with the processing of the wavelength request, thus decreasing the overall delay although the final burst size would have to be estimated by monitoring the buffer filling statistics. Packets arriving subsequently to the receipt of the acknowledgement packet are designated to the next burst.

It takes a finite propagation time across the network $t_{\text{prop,net}}$, for the first bit to arrive at the destination edge router, so the reserved wavelength is idle and not used for data transmission for the period $t_{\text{RTT}}= t_{\text{prop,ack}} + t_{\text{prop,net}}$. The time to complete the burst transmission is $t_{\text{trans}} = L_{\text{burst}}/b_{\text{core}}$, so that the wavelength holding time is given by $t_{\text{WHT}} = t_{\text{RTT}} + t_{\text{trans}}$. In principle, the wavelength holding time could be fixed either by the maximum edge delay or by streaming data in which case $t_{\text{WHT}}$ would be less predictable but the lightpath utilization would increase [Mig01]. The maximum latency or upper bound on the maximum transmission time that packets experience between entering the core network at the source and leaving the destination routers is

$$\text{Latency}_{\text{max}} = t_{\text{edge}} + t_{\text{prop,net}} + L_{\text{burst}}/b_{\text{core}}. \quad (4.1)$$

The arrival of the acknowledgement packet from the control node sets the start of the subsequent burst assembly and the cycle repeats.

From the analysis of the timings involved in burst assembly and transmission it is clear that the network efficiency depends on the processing speed of the network controller. Minimization of $t_{\text{proc}}$ can be achieved by applying fast dynamic routing and wavelength assignment algorithms. Efficient algorithms already exist for the optimization of static and dynamic WRONs, see for example [Bar97, Mok98], and have also been applied in the design of WR-OBS [Mig01]. It was assumed in this work that a wavelength will always be available, under the conditions of an ideal RWA algorithm. This is, of course, an idealizing assumption which does not hold in practice since dynamic RWA do not give as efficient results as static RWA algorithms where traffic demand is known $a$-
priori. The term efficiency here refers to a low blocking probability, which corresponds to the minimum number of wavelengths required to satisfy all demands set by the traffic matrix. The design of fast dynamic RWA algorithms with low blocking probability is an open research area, see, for instance, [Mig02] and references therein. For a given network topology and optimized route look up and wavelength allocation algorithms $t_{\text{prop,ack}}$ and $t_{\text{proc}}$ are known \textit{a-priori} and the timings of wavelength requests can be adjusted by the edge routers to meet latency and PLR criteria.

4.3 Core network performance model and results

In this section, an analytical model to calculate network performance parameters is derived for some of the results of chapter 3, to extend the analysis of burst aggregation and to quantify the achievable wavelength reuse and lightpath utilization in OBS networks for this fast circuit-switched architecture where the lightpaths are set up only for the time required to transmit the content of a single buffer between two edge routers, and then released for subsequent requests. This time includes an overhead required for lightpath set-up, and propagation delays, and was previously introduced as round-trip time, $t_{\text{RTT}}$. Following the results of the previous section, Figure 3.2, in this section it is assumed that burst sizes increase linearly, equivalent to the case of CBR traffic arriving to the buffer, and for which there is no variation in the burst size. Then, for a constant load and CBR traffic, the burst size, $L_{\text{burst}}$ is proportional to the edge delay and the input bit rate $b_{\text{in}}$, so that $L_{\text{burst}} = b_{\text{in}} \cdot t_{\text{edge}}$.

For clarity and simplicity, the analysis in this section is based only on mean values for all parameters. However, for an arbitrary burst size distribution, the PDF can be derived as described in Appendix F, and the related deviation be calculated either explicitly from the PDF, or estimated for small deviations by the means of error progression.
4.3.1 Calculation flow chart for derivation of network performance parameters from queueing results

The flow chart in Figure 4.2 shows how the values for the performance parameters investigated in this chapter are derived from the burst size distributions investigated by both simulation and analytical techniques based on queueing theory in chapter 3, indicating the close link between the two areas of research. All parts originating from chapter 3 on queueing theory and burst aggregation are grey shaded in Figure 4.2.

The burst size mean and variance are determined from the traffic generated by ON-OFF sources either by analytical means using advanced queueing theory (section 3.4). Alternatively these values are determined by simulation (section 3.5). Based on the mean burst size, the mean traffic rates is used in the continuous bit rate (CBR) model used for most of this section, leading to upper bounds on the performance parameters. It is a key advantage that the mean performance parameters do not depend on the traffic statistics used, so the results apply to a wide variety of input traffic statistics. It was shown, however, that the input traffic characteristic determines the burst size variation (section 3.4 or 3.5). The variation of the performance parameters in this chapter is derived from the burst size variation by calculating it from the PDF of the performance parameter. Details of this procedure and formulae for the PDF of the performance parameters are given in Appendix F. Alternatively, for small deviations, it is possible to calculate the variance of performance parameters from the burst size variation using the means of derivation of error based on a normal distribution, also detailed in Appendix F.
### Performance analysis of the WR-OBS network architecture

#### Input traffic statistics (ON-OFF source)

<table>
<thead>
<tr>
<th>Queueing Theory</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ON-OFF source statistics</td>
<td>B = const. or $t_{edge} = const.$</td>
</tr>
<tr>
<td>- mean</td>
<td>-</td>
</tr>
<tr>
<td>- autocorrelation function</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Burst size distribution (chapter 3)

- Buffers adapt input bit rate ($b_{in}$) to core bit rate ($b_{core}$)

<table>
<thead>
<tr>
<th>Mean burst size</th>
<th>Variance of burst size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{BR}$ traffic</td>
<td>$\text{traffic variation}$</td>
</tr>
</tbody>
</table>

#### Performance parameters

- wavelength holding time ($t_{WHI}$)
- bandwidth per wavelength ($B_{Per}$)
- lightpath utilization ($U$)
- reuse factor ($RUF$)
- performance parameter ($P$)

#### Deviation ($\sigma^2$, 95% level)

PDF for new RV $Y$ from $PDF_{burst}$

$$f_Y(Y) = \frac{f_{burst}(burst)}{\frac{\partial g}{\partial burst}},$$

where $Y = g(burst)$ is one of the performance parameters.

Then, if $\sigma/\mu \ll 1$, the following approximation holds (based on normal distribution):

$$\sigma_Y^2 = \frac{\partial g}{\partial burst} \cdot \sigma_{burst}^2.$$

#### Combine results for mean of performance parameters and their PDFs

**Figure 4.2**: Diagram for the calculation of the results on burst size distribution during the edge aggregation process, obtained by simulation or analytical model as specified in chapter 3 (grey shaded), are used to derive network performance parameters and parameter deviations in this chapter.
4.3.2 Wavelength holding time

Once a burst is assigned to a free wavelength, this wavelength will be reserved and is used until the buffer content is transmitted from the source to the destination edge router. The wavelength holding time, $t_{\text{WHT}}$, shown in Figure 4.3, can be thought of as equivalent to the call-holding time in circuit-switched networks. It is given by:

$$t_{\text{WHT}} = t_{\text{RTT}} + \frac{L_{\text{burst}}}{b_{\text{core}}} = t_{\text{RTT}} + \frac{1}{A} \cdot t_{\text{edge}} \tag{4.2}$$

where $t_{\text{RTT}}$ is the idle time before the burst reaches the destination edge router plus the time for the acknowledgement, and $A$ is the core bit rate to input bit rate ratio, $A = b_{\text{core}}/b_{\text{in}}$. For small values of $A$, the data transmission time $t_{\text{trans}}$ can be in the range of tens of milliseconds, so that $t_{\text{RTT}} \ll t_{\text{WHT}}$. Time $t_{\text{RTT}}$ starts to affect the service quality when the values of the $t_{\text{trans}}$ are comparable to $t_{\text{RTT}}$, and dominates the wavelength holding time for high core bit rates such as $b_{\text{core}} = 100$ Gb/s as shown in Figure 4.3-a for $t_{\text{RTT}} = 2, 5, 10$ ms. Figure 4.3-b shows the effect for $t_{\text{RTT}} = 5$ ms and a variation of $b_{\text{core}}$ from 20 – 100 Gb/s, where for $b_{\text{core}} = 20$ Gb/s the $t_{\text{WHT}}$ is significantly longer than for $b_{\text{core}} = 100$ Gb/s.

![Figure 4.3: Wavelength holding time, $t_{\text{WHT}}$, as a function of the edge delay, for $b_{\text{core}} = 100$ Gb/s, $b_{\text{in}} = 10$ Gb/s and $t_{\text{RTT}} = 2, 5, 10$ ms (a), and $t_{\text{RTT}} = 5$ ms and $b_{\text{core}} = 20, 40, 100$ Gb/s (b). Bars show 95% confidence level.](image)

4.3.3 Bandwidth-per-wavelength

A parameter following from equation (4.2) is the bandwidth-per-wavelength, $B_{\text{per}}$, which indicates the effective bandwidth of a lightpath used for transmission of data between edge routers:

$$B_{\text{per}} = \frac{L_{\text{burst}}}{t_{\text{WHT}}} = \frac{b_{\text{in}} \cdot t_{\text{edge}}}{t_{\text{RTT}} + \frac{1}{A} \cdot t_{\text{edge}}} \tag{4.3}$$
The influence of $t_{\text{RTT}}$ on $B_{\text{perλ}}$ is shown in Figure 4-a for $b_{\text{core}} = 100 \text{ Gb/s}$ and $t_{\text{RTT}} = 2, 5, 10 \text{ ms}$. The increase in bandwidth for the identical values $t_{\text{edge}}$ is reduced for higher $t_{\text{RTT}}$; for $t_{\text{RTT}} = 10 \text{ ms}$ values remain below 50 Gb/s for 100 Gb/s physical bit rate. Figure 4-b shows the effect of bandwidth saturation for $t_{\text{RTT}} = 5 \text{ ms}$ and core bit rates varying between 20 - 100 Gb/s. The significance of the results is that $B_{\text{perλ}}$ remains significantly smaller than the optical line rate for $t_{\text{edge}} \leq 40 \text{ ms}$, especially for high $b_{\text{core}}$ such as 100 Gb/s.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Bandwidth-per-wavelength $B_{\text{perλ}}$ for $b_{\text{core}} = 100 \text{ Gb/s}, b_{\text{in}} = 10 \text{ Gb/s}$ and $t_{\text{RTT}} = 2, 5, 10 \text{ ms (a)},$ and for $t_{\text{RTT}} = 5 \text{ ms}$ and $b_{\text{core}} = 20, 40, 100 \text{ Gb/s (b).}$ Bars show 95\% confidence level.}
\end{figure}

### 4.3.4 Lightpath utilization

Relating the bandwidth-per-wavelength, $B_{\text{perλ}}$, to the physical bit rate in the core, $b_{\text{core}}$, leads to the dimensionless parameter $U$, the utilization which describes the efficiency with which the lightpath is utilized:

$$U = \frac{B_{\text{perλ}}}{b_{\text{core}}} = \frac{t_{\text{edge}}}{A \cdot t_{\text{RTT}} + t_{\text{edge}}} \quad (4.4)$$

Maximizing the use of available resources is key for the network operator, implying that utilization $U$ must be maximized. Figure 4.5 shows the dependence of $U$ on $t_{\text{edge}}$, $A$, and $t_{\text{RTT}}$ for the following range of values: $0 \text{ ms} \leq t_{\text{edge}} \leq 200 \text{ ms}$, $0 \leq A \leq 100$, and for $t_{\text{RTT}} = 2$ (Figure 4.5-a) and 10 ms (Figure 4.5-b), as calculated in the previous section. As can be seen from Figure 4.5, the highest bandwidth utilization is achieved for low values of $A$ and high edge delays ($> 50 \text{ ms}$). Utilization also increases for smaller $t_{\text{RTT}}$ but the same values of $t_{\text{edge}}$ and $A$, as shown for $t_{\text{RTT}} = 2 \text{ ms}$ in Figure 4.5-a as compared to $t_{\text{RTT}} = 10 \text{ ms}$ in Figure 4.5-b.
4.3.5 Wavelength reuse factor

In high-speed networks it can be assumed that \( b_{\text{core}} \gg b_{\text{in}} \) results in \( t_{\text{WHT}} \ll t_{\text{edge}} \), i.e. the time required to aggregate a burst is significantly larger than the time to transmit it. In the case of dynamic wavelength allocation an unused wavelength can be assigned to another edge router, and the resultant increase in the wavelength reuse can be defined as a wavelength reuse factor, \( \text{RUF} \), defined as:

\[
\text{RUF} = \frac{t_{\text{edge}}}{t_{\text{WHT}}} = \frac{A \cdot t_{\text{edge}}}{A \cdot t_{\text{RTT}} + t_{\text{edge}}} = A \cdot U
\] (4.5)

For consistency, the variation of \( \text{RUF} \) is plotted in Figure 4.6 for the same range of values as for the utilization \( U \). For comparison to a static WRON, Figure 4.6 shows the values for \( \text{RUF} = 1 \). This is justified by the assumption that in a static WRON a given lightpath is established for a long period, but not shared between different edge routers.

In an optical network with dynamic wavelength assignment, this is equivalent to a lightpath permanently assigned between two edge routers, i.e. \( t_{\text{edge}} = t_{\text{WHT}} \). For \( \text{RUF} < 1 \) the WR-OBS network would theoretically require more wavelengths than in a static WRON to satisfy all demanded connections, and, therefore, values for \( \text{RUF} < 1 \) represent the region of network instability where the total input load exceeds the network throughput.

Despite the potential savings in terms of the number of wavelengths, it should be noted that the actual number of wavelengths required also depends on the physical topology and routing strategy, as well as the wavelength allocation algorithm [Bar97, Mok98], as shown in chapter 5.

The variation of the mean \( \text{RUF} \) for \( t_{\text{RTT}} = 2 \) and 10 ms are shown in Figs 4.6-a and 4.6-b. It can be seen that the \( \text{RUF} \) increases with both \( t_{\text{edge}} \) and \( A \), to values of \( \text{RUF}_{\text{max}} = 50 \).
and 16.7, respectively, for the given range of \( t_{\text{edge}} \) and \( A \) with \( A_{\text{max}} \), \( t_{\text{edge, max}} \) as the maximum values of their given ranges (\( A_{\text{max}} = 100 \), \( t_{\text{edge, max}} = 200 \) ms). It should be noted that for a given \( t_{\text{RTT}} \) upper bounds for the reuse factor can be determined for either constant \( t_{\text{edge}} \) or constant \( A \):

In the case of \( t_{\text{edge}} = \text{const.} \), and increasing \( A \), the data transmission time \( L_{\text{burst}}/b_{\text{core}} \) becomes negligible so that \( t_{\text{WHT}} = t_{\text{RTT}} \) and \( \text{RUF} \leq t_{\text{edge}}/t_{\text{RTT}} \). In the case of \( A = \text{const.} \) the buffer content increases proportionally with \( t_{\text{edge}} \), restricting the reuse factor to \( \text{RUF} \leq A \). For constant \( t_{\text{edge}} \) an increase of \( A \) is only useful for \( A \leq t_{\text{edge}}/t_{\text{RTT}} \), for larger values of \( A \) the reuse factor will only increase marginally. For constant \( A \), an increase in \( t_{\text{edge}} \) is beneficial only for \( t_{\text{edge}} \leq A \cdot t_{\text{RTT}} \).

![Figure 4.6: Mean wavelength reuse factor, RUF, as function of edge delay \( t_{\text{edge}} \) and bit rate ratio \( A \), for \( t_{\text{RTT}} = 2 \) ms (a) and \( t_{\text{RTT}} = 10 \) ms (b)](image)

The significance of the reuse factor as a metric in WR-OBS networks arises from the savings in terms of wavelengths which can be achieved. These savings can be pronounced especially in large networks with numbers of nodes > 40, e.g. USNet and EuroLarge, which require 103 and 66 wavelengths, respectively, for static assignment creating a logical full-mesh [Bar97]. Assuming a channel spacing of 0.4 nm, it would just be possible to realize both networks using a C-band infrastructure [Kei99] for bit rates up to 40 Gb/s (NRZ coding with 0.8 bit/(sHz) spectral efficiency, [Idl02]). For additional demand in these networks it would be necessary to either extend transmission to the S- or L-band, or to light a second fibre network in the C-band. In each case it would be necessary to invest in new transmission equipment such as transponders, amplifiers and dispersion compensation suitable for the second fibre network. If, however, the network infrastructure was dynamically enabled, even a wavelength reuse factor of 2 would be sufficient to continue transmission entirely in the C-band, avoiding the investment in new infrastructure.
4.3.6 Network performance parameter

A comparison between Figs. 4.5 and 4.6 shows that both utilization and reuse factor increase with $t_{\text{edge}}$, but that $U$ is maximum for low values of $A$, whereas $\text{RUF}$ maximizes for high values of $A$. The resulting trade-off for constant $t_{\text{edge}}$ between both parameters can be described by the dimensionless parameter $P$ defined as the product of $U$ and $\text{RUF}$,

$$P = U \cdot \text{RUF}$$

and is plotted in Figure 4.7 for $0 \text{ ms} \leq t_{\text{edge}} \leq 200 \text{ ms}$, $0 \leq A \leq 100$, with $t_{\text{RTT}} = 2 \text{ ms}$ (Figure 4.7-a) and $10 \text{ ms}$ (Figure 4.7-b). For a constant value of $t_{\text{edge}}$, $P$ can be optimized for a set of parameters $\{A, t_{\text{edge}}\}$ such that $U_{\text{opt}} = 50\%$ and $\text{RUF}_{\text{opt}} = 2 \cdot P_{\text{opt}}$ are achievable. Hence the WR-OBS network benefits from good utilization and wavelength reuse. As in preceding graphs, the optimization process also depends on $t_{\text{RTT}}$ as shown in Figs. 4.5-a and 4.5-b. Especially for $t_{\text{RTT}} = 10 \text{ ms}$, $\text{RUF} \geq 1$ must be maintained to avoid network instability.

![Figure 4.7: Network performance parameter, $P$, showing the trade-off between utilization $U$ and wavelength reuse $\text{RUF}$ for constant $t_{\text{edge}}$. $t_{\text{RTT}}$ = 2 ms (a) and $t_{\text{RTT}}$ = 10 ms (b). Optimum values $P_{\text{opt}}$ are shown by dashed line.](image)

4.3.7 Impact of the round-trip time on network performance

To investigate the impact of the $t_{\text{RTT}}$ on the network, both the utilization $U$ and reuse factor $\text{RUF}$ are re-plotted for different values of edge delays ($10, 20, 50 \text{ ms}$) and constant $A = 10$ in Figures 4.8 and 4.9. To ensure that $\text{RUF} > 1$ requires

$$t_{\text{RTT}} < \frac{A - 1}{A} \cdot t_{\text{edge}} \iff t_{\text{RTT}} < t_{\text{edge}} \quad \text{for } A \gg 1.$$  

A key result is that for $A \gg 1$, as in high core bit rate networks, a high reuse factor can be achieved only for $t_{\text{RTT}}$ on the timescale of a few milliseconds. It is important to note that in order to achieve efficient wavelength reuse, the lightpath set-up time must be as small as possible, and for a fixed $t_{\text{edge}}$, $\text{RUF}_{\text{max}}$ is given for instantaneous lightpath set-up ($t_{\text{RTT}} = 0$) as $\text{RUF}_{\text{max}} = A$. Not
only does the wavelength reuse factor decrease with an increasing $t_{RTT}$, but so does the lightpath utilization, which in all cases is less than 50% for round-trip times $t_{RTT} \geq 10$ ms, and drops sharply especially for an edge delay of 10 ms. These results show that the round-trip time, $t_{RTT}$, is a key parameter in the design of OBS networks with dynamic wavelength allocation. The results also define the performance requirements on the dynamic RWA algorithm used for lightpath establishment between edge routers to minimize the overhead of the time $t_{RTT}$. This is necessary to achieve the operational advantage of increased throughput per wavelength under dynamic wavelength operation. For the speed of a RWA algorithm this implies that the RWA decision time must meet even tight delay constraints. Identifying these constraints will help optimize the RWA process carried out in the control node to maximize the number of edge routers and network routes, as described in chapter 5.

These constraints also affect the maximum size over which a WR-OBS network can span. For the speed of light with 200 km/ms, an edge delay constraint of 10 ms translates into a network diameter of < 1,500 km. This is sufficient for all European national and even pan-European networks, but might not be sufficient for US networks.

![Figure 4.8: Wavelength reuse factor, RUF, as a function of the round-trip time ($t_{RTT}$) for $t_{edge} = 10, 20, 50$ ms, and bit rate ratio $A = 10$. Shaded region: network requires more wavelengths than in a static WRON](image-url)
From the analysis carried out it is clear that the constraints for $t_{\text{RTT}}$ limit the network diameter, or the allowable minimum edge delays for efficient network operation. Thus, a WR-OBS scheme brings most advantages for network sizes found in Europe, or metropolitan area-type networks where the lower signaling round-trip times allow wavelength savings from dynamic network operation.

### 4.4 Summary

This chapter described and analyzed the wavelength-routed optical burst-switched (WR-OBS) network which combines the functions of OBS with fast circuit switching by dynamically assigning and releasing wavelength-routed lightpaths over a bufferless optical core. The potential advantages of this architecture compared to conventional OBS are explicit QoS provisioning and - compared to static WRONs - are in fast adaptation to dynamic traffic changes in optical networks and more efficient utilisation of each wavelength channel. The proposed architecture ensures a deterministic delay for the optical packets through a known, pre-defined delay at the edge queueing and burst aggregation, and the propagation in the core network. Moreover it guarantees an acknowledgement of the wavelength-assignment for QoS-determined provisioning and uses dynamic wavelength-routing.

The lightpath utilization and wavelength reuse factor were introduced to characterize OBS networks with dynamic wavelength assignment. It was shown that these parameters could be described as a function of the edge delay, $t_{\text{edge}}$, the round-trip time,
Performance analysis of the WR-OBS network architecture

$t_{\text{RTT}}$, and the ratio of core to input bit rates, $A$, allowing the results to be generalized to cover a wide range of input and core bandwidths. These results also allowed to quantify the operating range for $A$ and $t_{\text{edge}}$ for which increases in lightpath utilization and reuse of a given wavelength channel, and thus increased network throughput, can be achieved relative to static WRONs. However, to attain this performance, the signalling round-trip time for the acknowledgement of dynamic wavelength reservation and wavelength assignment must be much shorter than the edge delay, setting stringent limits on the performance of such networks. The allowable round-trip time delay is related to the maximum network diameter. Whereas the WR-OBS architecture would probably work for maximum network diameters up to 4,000-km with round-trip time delays of 30 ms, it would offer the most significant advantage for networks with smaller diameters, e.g. as found with European network operators and for metropolitan-area size networks. The results can be applied to the design and the dimensioning of wavelength-routed optical burst-switched networks and the optimization of scheduling, control and wavelength assignment in coordination between the electronic and the optical network layers.
4.5 References


Chapter 5 Scalability of the WR-OBS network architecture with QoS considerations

5.1 Introduction

This chapter analyses the scalability of a WR-OBS architecture where end-to-end QoS is required. In particular, it is investigated how the network scales with the number of wavelength requests. This is key not only to the WR-OBS architecture, but to any network architecture with dynamic wavelength assignment. In the design process, conflicting constraints may have to be accommodated: for instance, the blocking probability is expected to decrease with increasing processing time since more potential solutions can be evaluated. This, however, limits the overall number of requests which can be processed per unit time, so a trade-off between the two constraints must be found. The analysis here is closely linked to the results obtained for the adaptive burst assembler, chapter 3.5.

Meeting QoS constraints such as latency and packet loss rate (a prerequisite in all future network architectures) is a key feature of all optical network architectures where network resources are dynamically allocated. The performance analysis in chapter 4 was based on the assumption of an ideal RWA algorithm where a lightpath was always granted. This assumption may not hold in a real network since due to constraints of the topology and the current state of the network, a RWA algorithm might not be able to find a lightpath; the request would be blocked resulting in potential loss of a burst at the edge. Under QoS constraints it is paramount to guarantee that an acknowledgement is granted within a limited amount of time to meet delay constraints. It is assumed that resources, like lightpaths, are assigned across the entire network by the network management, concentrated in a central control node. A centralized approach was assumed for the following reasons:

- Dynamic RWA algorithms proposed to date, centralized or distributed e.g. in [Mig02 and references therein, Mok98], extend the solution of the static RWA problem to the dynamic case by taking into account the current state of the network during the decision process. A centralized database on the wavelengths and routes in use is easier to maintain and faster to update than distributed databases.
• A centralized solution with a single processor would give a worst-case scenario (delay before a decision is made on whether or not a lightpath can be established), where the network scalability is determined by the ability of fast processing of incoming requests at a single decision point.

In the WR-OBS networks with admission control bursts will not enter the network before an acknowledgement with route and wavelength information is sent. The key parameter is, thus, the latency since applications are affected when delay requirements are violated. However, even best-effort-type traffic can be affected when an overdue acknowledgement results in buffer overflow and packet loss at the network edge, as shown by previously (section 1.2) discussed results on TCP throughput over OBS networks with one-way reservation [Det02]. It is, therefore, key to schedule incoming wavelength requests such that related QoS demands such as latency or packet loss are met irrespective of the request arrival statistic. In section 3 it was shown that the edge routers in WR-OBS networks can be modelled to generate wavelength requests at a deterministic rate, with the rate of the desired burst frequency. Corresponding wavelength requests then arrive at the network control node with a nearly constant rate, simplifying the scheduling of requests as shown in the remainder of this chapter.

5.2 Related work

A key question in the design of networks with two-way reservation is the scalability of the central network control and management system which establishes end-to-end connections across the entire network. The term scalability here refers to the maximum number of nodes which a network architecture can support without violating service constraints such as a maximum end-to-end latency or PLR.

For any dynamic network with end-to-end lightpath assignment it is mandatory to process each wavelength request within a given time, where the upper bound on the processing time ultimately limits the number of edge nodes and classes of service which can be supported. The problem of designing a central controller, similar in nature to the one considered in this work, was investigated within the SONATA project, envisaging one passive optical node with wavelength tuneable sources at the edge, serving 20 million end users spread over 400 passive optical networks via a TDMA/WDMA (Time division multiple access/wavelength division multiple access) [Bin00]. Although the
related design problem was found to be NP-hard, sub-optimal heuristics of the resource allocation problem were shown to provide satisfactory performance [Bia00].

The problem of network scalability with respect to the numbers of nodes supported has not previously been studied in the context of OBS networks. To date, only two studies were concerned with the investigation of one-way reservation schemes in the context of entire networks [Dol01, Mye01], but concentrated on the performance degradation with increasing number of hops (burst loss, [Dol01]), and the number of required wavelengths to support a pre-defined burst loss probability network wide [Mye01].

The function of providing service guarantees in real-time is a task required not only by communication networks, but also in a number of different applications, especially for control purposes. The control is carried out by a microprocessor with scheduling of tasks such that the deadlines of all arriving tasks are met.

To study the effect of real-time scheduling of wavelength requests two types of algorithms were applied to the problems in the WR-OBS control node: The rate-monotonic (RM) algorithm [Leh89] where priorities are assigned at a rate inversely proportional to task periods, and the earliest-deadline-first (EDF) algorithm which assigns priorities inversely proportional to absolute deadline [Bin01]. The RM algorithm is preferred to the EDF algorithm due to reduced efforts in implementation and its ability to meet latency constraints even in the presence of transient overload [Leh89], as described later in this section. The RM algorithm is easier to implement since over a certain period of time requests are expected to arrive at a fixed rate (1/t_{edge}). Hence there is less requirement to re-arrange the scheduling table than for the EDF algorithm, where each individual request is scheduled with respect to all other requests within the queue. Both algorithms were originally designed for the execution of real-time tasks, e.g. in control engineering. For the purposes of operation in a network environment both algorithms were modified: The period between occurrence of a particular wavelength request is shortened by the round-trip time which cannot be used for processing.

In this section, the modified RM and EDF algorithms are applied, and bounds derived to calculate the maximum number of edge routers which can be supported in a WR-OBS network architecture for the latencies of different CoS. In a final step the time required to process a wavelength request is determined as a function of the RWA algorithm, the
number of memory lookups and memory access times as well as the implementation of the database.

The EDF algorithm has been used in the context of WR-OBS networks for request scheduling in a simulation study [Mig01], but has been applied in [Due02] for the first time to determine the network scalability of networks with dynamic wavelength provisioning in general.

5.3 Processor utilization and bounds

The original approach as used for processor scheduling [Bin01] needed to be modified for this work to account for the delays introduced by the round-trip time $t_{RTT}$ in the WR-OBS network context. The following parameters are required to determine whether a task (e.g. a wavelength request) is schedulable:

- $T_i$: Task $i$, where $i = 1, ..., n$ (n denotes the total number of tasks/requests)
- $T_i$: Task period of task $i$ (here processing of a wavelength request)
- $C_i$: Worst case execution time of task $i$, which in the context of the WR-OBS architecture would be equivalent to a wavelength request; please note that the variable $C_i$ in this chapter is different from the buffer depletion rate $C$ used in chapter 3
- $D_i$: Relative deadline
- $U_i = C_i/T_i$: Utilization factor of task $i$ with $0 \leq U_i \leq 1$

$U_p$: Total processor utilization with $U_p = \sum_{i=1}^{n} U_i$, and $0 \leq U_p \leq 1$  \hspace{1cm} (5.1)

For the EDF algorithm, as set of $n$ tasks can be scheduled if and only if

$$U_p = \sum_{i=1}^{n} U_i \leq 1.$$  \hspace{1cm} (5.2)

For the RM algorithm, as set of $n$ tasks can be scheduled provided that

$$U_p = \sum_{i=1}^{n} U_i \leq n(2^{1/n} - 1).$$  \hspace{1cm} (5.3)

For a large number of tasks ($n \to \infty$) the upper bound of the processor utilisation for the RM algorithm becomes:

$$U_p = \lim_{n \to \infty} n(2^{1/n} - 1) = \ln 2 = 0.69$$  \hspace{1cm} (5.4)

Eqn. (5.4) presents a worst case bound assuming that all tasks arrive at the processor at the same time (in-phase arrival) – even under this condition, a utilization of 69% is achievable. This bound can be replaced by a less stringent one, called the hyperbolic
Scalability of the WR-OBS network architecture with QoS considerations

bound [Bin01]. A set of n tasks with individual processor utilization $U_i$ can be scheduled with the RM algorithm if the following relation holds:

$$\prod_{i=1}^{n} (U_i + 1) \leq 2$$  \hspace{1cm} (5.5)

5.4 Application of modified RM and EDF algorithms in the WR-OBS network architecture

In this section it is shown how the RM and EDF algorithms need to be modified for the scheduling of wavelength requests in the WR-OBS network architecture. They are then used to calculate the maximum number of edge routers supported, first for networks were all edge routers are located at the same distance from the control node, as would be the case in a star-coupled network (for example [Bor98]), then for seven meshed network topologies to prove operation over arbitrarily meshed topologies.

5.4.1 Modification of the RM and EDF algorithms and their application in networks with equidistant node spacing

Let $C(i,k)$ be the time to process a wavelength request $\tau_{ik}$, where $i$ indicates the CoS and $k$ the edge router. $C(i,k)$ is a function dependent on the state of the network and the network topology, the processor speed and memory access time, and is here assumed to reflect the worst case state, such that finally $C_{ik} = \max C(i,k) = \text{const.}$ Within network processors for which the RM and EDF algorithms were developed, propagation delays are negligible – which is not true, however, in wide area networks, and hence requires modification of existing theory as described in the following paragraph.

The propagation delay of the wavelength request between edge router and control node must be taken into account and substracted from the period $T_{ik}$ of request $\tau_{ik}$ as a request can only be scheduled when it has reached at the central control node. Hence the period $T_{ik}$ as defined above must be as follows:

$$T_{ik} = t_{\text{edge},ik} - t_{\text{RTT},ik},$$  \hspace{1cm} (5.6)

where $t_{\text{edge},ik}$ denotes the average edge delay (as defined in chapter 1), i.e. the average time between sending two wavelength requests irrespective of the burst aggregation technique, and $t_{\text{RTT},ik}$ is the round-trip time (propagation delay) for the wavelength request to travel from edge router $k$ to the central control node, and for the acknowledgement (or rejection) to travel back to the edge router, given both take the same route in the network. Assuming a control node with a single processor and the
possibility of re-ordering requests within an infinite input queue\(^4\), the network is robust if and only if conditions (5.2) or (5.3) are satisfied:

\[
\sum_{i=1}^{N_{\text{COS}}} \sum_{k=1}^{N_{\text{COS}}} (N_{\text{edge}} - 1) \cdot \frac{C_{ik}}{T_{ik}} = U \leq \begin{cases} N \cdot 2^{1/N} - 1 \leq 1, & \text{for RM} \\ 1, & \text{for EDF} \end{cases}
\]

where: \(N = N_{\text{COS}} \cdot (N_{\text{edge}} - 1)\)

\(N_{\text{edge, max}}\) is the maximum number of edge routers for which the system is stable (no violation of latencies) for a given \(N_{\text{COS}}\). It is further assumed that \(C_{ik} = C = \text{const.}\) for all \(i, k\), round-trip times of all connections are identical (equidistant node spacing), leading to \(t_{\text{RTT, ik}} = t_{\text{RTT}}\), and that the edge delay for every CoS is unique, \(t_{\text{edge, ik}} = t_{\text{edge, i}}\). Applying equation (5.7), \(N_{\text{edge, max}}\) is calculated as:

\[
N_{\text{edge}} = \left[ \frac{1}{2} + \left( \frac{1}{4} + U \cdot \left( C \cdot \sum_{i=1}^{N_{\text{COS}}} \frac{1}{T_i} \right)^{-1} \right)^{-1/2} \right]
\]

For large \(N_{\text{edge}}\), \(N_{\text{edge}} \cdot (N_{\text{edge}} - 1) \approx N_{\text{edge}}^2\), so that eqn. (5.9) can be simplified to:

\[
N_{\text{edge}} \approx \sqrt{U \cdot \left( C \cdot \sum_{i=1}^{N_{\text{COS}}} \frac{1}{T_i} \right)^{-1}} \approx \frac{1}{\sqrt{C}}
\]

The number of edge routers as a function of the processing time per request \(C\) is plotted in Figure 5.1 for \(0.1 \mu s \leq C \leq 10 \mu s\). The selection of these values for \(C\) will be justified in chapter 5.5, where it is shown that \(C\) is mostly limited by the number of memory lookups required. \(N_{\text{edge, max}}\) is shown for the case of 3 CoS under RM and EDF bound as well as 1 CoS (lowest latency \(T_1\) only) with EDF bound. The parameters were chosen as \(t_{\text{RTT}} = 5\) ms, and request periods of different CoS as \(T_1 = 5\) ms, \(T_2 = 15\) ms, \(T_3 = 45\) ms (for edge delays \(t_{\text{edge, 1}} = 10\) ms, \(t_{\text{edge, 2}} = 20\) ms and \(t_{\text{edge, 3}} = 50\) ms, respectively).

---

\(^4\) with identical assumptions as in chapter 4 (250 bits per control packet, 100 nodes, 3 CoS, 10 ms between requests), the required queue size approximately 1 MB for constant emptying, so the assumption of an infinite buffer (e.g. 128 MB) is realistic.
Scalability of the WR-OBS network architecture with QoS considerations

Figure 5.1: Number of edge routers as a function of processing time $C$ and for 3 CoS using the RM and the EDF algorithms with $T_1 = 5$, $T_2 = 15$ and $T_3 = 45$ ms, as well as for 1 CoS (lowest latency $T_1 = 5$ ms only), $t_{RTT} = 5$ ms

For $C = 0.1$ $\mu$s (equivalent to 200 cycles of 2-GHz processor), 3 CoS and parameters as above, the network can support a maximum of 186 edge routers when using the EDF algorithm. For a network without CoS differentiation and $T = T_1 = 5$ ms only, the number of edge routers which can be supported increases to 223. This result implies that the most time-sensitive requests (the highest CoS) determine the network performance. This means that in a network additional classes of service with less stringent delay requirements can be accommodated with a minimal reduction in the allowable number of edge routers. The critical parameter in this analysis is the request processing time $C$, the values of which are derived in the next section.

The significance of the results is that they do not only hold for the WR-OBS architecture investigated in this work, but apply more generally for dynamic optical networks. In particular they represent an upper bound to the scalability of the optical network, and therefore, the speed with which the electronic processing must operate and thus the table lookup times required for the determination of a suitable lightpath by the RWA algorithm. Further improvements in the capacity of electronic processing as discussed in chapter 2.2 would translate into increased scalability of the dynamic network architecture.
5.4.2 QoS and fairness in meshed network architectures

The results as shown in Figure 5.1 were obtained under the assumption that the network architecture is regular in the sense that all edge routers are located at the same distance from the control node so that the assumption $t_{RTT,ik} = t_{RTT}$ holds. This is not true, however, in most real network architectures which are arbitrarily meshed and $t_{RTT,ik}$ is different for each edge router – control node connection. The resulting difference in the period $T_{ik}$ would result in unfairness: edge routers located closer to the control node could send wavelength requests more frequently. The application of the RM and EDF algorithms for scheduling, however, allows to establish fairness even in meshed network architectures. Prioritizing the wavelength requests proportional according to their distance from the control node results in high priorities for requests with short periods $T_{ik}$, i.e. requests from edge routers located at greater distance inherently obtain higher priorities. Thus the control node is able to enforce the same edge delay across the entire network architecture irrespective of the topology used. This was applied to the NSFNet (details in Tab. 5.1) architecture in Figure 5.2-a, which shows the distribution of periods $T_{ik}$ in 5 ms intervals (Figure 5.2-b).

![NSFNet architecture with node number 6 designated as control node; link length given in kilometers (a). Distribution of wavelength request periods $T_{ik}$ in 5 ms intervals (b)](image)

The edge delay in this case was chosen to $t_{edge} = 40$ ms such that it would also hold for the node with the greatest distance from the control node. Due to the topology of the network the values of $T_{ik}$ are widely distributed with mean 18.11 ms and standard deviation 8.11 ms. Figure 5.3 shows the effect the topology has on the periods $T_{ik}$ for the same networks if different CoS are introduced. The 3 CoS introduced were chosen such that the edge delays $t_{edge} = \{40, 50, 80\}$ ms were achieved, allowing for a difference of 10 and 40 ms between the first and the second, and the first and the third CoS, respectively.
Scalability of the WR-OBS network architecture with QoS considerations

It can be seen from Figure 5.3 that due to the large round-trip time distribution the periods $T_{ik}$ of different CoS can overlap, which is especially the case for CoS 1 and 2. This means that a scheduler operating the RM or EDF algorithms assigns a higher priority to those requests of CoS 2 whose period is shorter than requests of CoS 1. This result proves that the central control node is capable of enforcing strict QoS guarantees even in meshed networks by taking into account the effective periods $T_{ik}$ of each request.

5.5 Request processing time $C$

An important parameter in quantifying the maximum number of edge routers supported by the algorithm is the calculation time $C$ required to process a wavelength request as shown in Figure 5.1. We now justify the approach of investigating the number of edge routers for processing times $0.1 \mu s \leq C \leq 10 \mu s$ by taking a closer look at the necessary calculations required to process a wavelength request [Zap03].

The dynamic RWA considered here decouples the search for a route and a wavelength, and keeps an updated database (matrix) on which routes and wavelengths are currently in use [Mok98]. The database for searching a route contains for each source-destination pair the pre-computed shortest path, using e.g. the Dijkstra algorithm [Kes97], which can be stored either as a linked list or an array. Since an array can be searched faster than a linked list, we only consider the use of arrays. It is obvious that the dynamic
RWA algorithm is closely related to the solution of the static RWA problem, e.g. by using pre-computed shortest paths. Due to the proximity to the static case, some findings of the analysis of static WRON are used in quantifying the processing time required [Bar98]. The request processing is carried out in three steps:

1) Firstly, the shortest path is obtained from the route database, an operation which requires a maximum of $L_{\text{route,max}}$ lookups, where $L_{\text{route,max}}$ is the maximum length of a route within a network. Based on the empirical study of real network architectures [Bar98], which relates $L_{\text{route,max}}$ to the number of core nodes ($N_{\text{core}}$), the range of values for $L_{\text{route,max}}$ becomes $\frac{N_{\text{core}}}{5} \leq L_{\text{route,max}} \leq \frac{N_{\text{core}}}{3}$. For a worst case estimation, $N_{\text{core}}/3$ was assumed.

2) The identification of a wavelength, which is not used along the shortest path, requires a maximum of $W$ lookups, where $W$ is the number of wavelengths in the network and also the number of wavelength matrices. This is based on the assumption that for each wavelength a separate matrix exists with entries on whether the wavelength is used on that link. Although $W$ depends largely on the topology and connectivity of the network, it has been shown in [Bar98] that the upper bound for the number of wavelengths required in a static WRON scales with $O(N^2)$. Due to the number of matrix lookups the route colouring problem is the most computational intensive in the case of a single processor machine. This complexity can be reduced by distributing the task using a multi-processor approach such that each wavelength matrix is equipped with its own processor and a copy of the lookup table. This approach would be justified by the reduction in required processing time.

For a single processor, in the worst case $L_{\text{route,max}}$ entries in each of the $W$ wavelength matrices must be checked, totaling in $W \cdot L_{\text{route,max}}$ lookup operations. This value is reduced by factor $W$ if a multi-processor machine is used.

3) If the search for route and wavelength is successful, the wavelength database requires updating, which takes a maximum of $L_{\text{route,max}}$ operations to update each link used.

In summary, the total maximum processing time $t_{\text{RWA,max}}$ for the single- and multi-processor case is calculated as:

\begin{align*}
\text{a) single processor:} & \quad t_{\text{RWA,max}} = (2 \cdot L_{\text{route,max}} + W \cdot L_{\text{route,max}}) \cdot t_{\text{mem.access}} \quad (5.11) \\
\text{b) multi processor:} & \quad t_{\text{RWA,max}} = (3 \cdot L_{\text{route,max}}) \cdot t_{\text{mem.access}} \quad (5.12)
\end{align*}
Equations (5.11) and (5.12) are based on the assumption that $t_{\text{RWA, max}}$ is dominated by the electronic memory lookup times rather than being limited by the operating frequency of the microchip. The access speed to electronic memory grows at a lower rate [McK02] than Moore's Law discussed in section 2.2: for DRAM, the increase in access speed is approximately 7% p.a. [McK02], whilst the number of operations per chip increases by approximately 60% p.a (Fig. 2.3). The current lookup speed of SRAM of 5 ns is equivalent to about 10 processor cycles of in a state-of-the-art chip operating at 2 GHz. The value of 5 ns was chosen as a compromise between current SRAM technology with access times of 10 ns [Cha02], and experimentally demonstrated SRAMs with access times < 1 ns (see, for example [Nam00] and references therein).

With $L_{\text{route, max}}$ scaling $O(N^2)$, the routing and wavelength assignment processing problem grows $O(n)$, i.e. linearly with the number of core nodes, when a multi-processor system is used, but scales order $O(N^3)$ in the case of a single-processor machine. This result is confirmed by analyzing the $t_{\text{RWA, max}}$ for a number of existing network architectures, and is shown in Figure 5.4 as function of the number of core nodes, assuming $L_{\text{route, max}} = N_{\text{core}}/3$ and $t_{\text{mem, access}} = 5$ ns (internal SRAM).
Figure 5.4: Upper bound for the processing time of the RWA algorithm for the case of a single processor (solid line) with the number of wavelengths \( W \) as parameter. The multiprocessor case is independent of \( W \) (dashed line). Values for real network architectures for both cases are given as well (open symbol – single processor, full symbol – multiprocessor). Linear fit to values for real networks, single processor case indicates scaling \( O(N^{0.73}) \).

Figure 5.4 shows the results for the case of control nodes equipped with single- or multi-processor machines. The processing time was determined for seven real network architectures [Bar98], summarized in table 5.1.

Analysing the numbers for the case of a single-processor system shows that the RWA processing time scales with \( O\left(N_{\text{core}}^{2.71}\right) \) (linear fit: 2.73±0.19 for 95% confidence interval), approximately in accordance with the predicted scaling of \( O\left(N_{\text{core}}^{3}\right) \) based on the upper bound for the required number of wavelengths. The number of nodes in the real networks under investigation is in the range 11…46. For these numbers the upper bound on the processing time we set \( C = t_{\text{RWA, max}} \), and so that values fall approximately in the range 0.1 … 10 \( \mu \text{s} \), justifying that this is the relevant set of data for which values were plotted in Figure 5.1.
### Table 5.1: Network parameters used to quantify $t_{WRA,\text{max}}$ in Figure 5.4 for seven given network topologies [Bar98]

<table>
<thead>
<tr>
<th>Network</th>
<th>Number of core nodes ($N_{\text{core}}$)</th>
<th>Connectivity ($\alpha$)</th>
<th>Number of wavelengths in static WRON (W)</th>
<th>$t_{WRA,\text{max}}$ single proc. [µs]</th>
<th>$t_{WRA,\text{max}}$ multi proc [µs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>USNet</td>
<td>46</td>
<td>0.07</td>
<td>108</td>
<td>8.43</td>
<td>0.23</td>
</tr>
<tr>
<td>Euro large</td>
<td>43</td>
<td>0.10</td>
<td>103</td>
<td>7.53</td>
<td>0.22</td>
</tr>
<tr>
<td>ARPANet</td>
<td>20</td>
<td>0.16</td>
<td>33</td>
<td>1.17</td>
<td>0.10</td>
</tr>
<tr>
<td>UKNet</td>
<td>21</td>
<td>0.19</td>
<td>21</td>
<td>0.81</td>
<td>0.11</td>
</tr>
<tr>
<td>EON$^5$</td>
<td>20</td>
<td>0.20</td>
<td>18</td>
<td>0.67</td>
<td>0.10</td>
</tr>
<tr>
<td>NSFNet</td>
<td>14</td>
<td>0.23</td>
<td>13</td>
<td>0.35</td>
<td>0.07</td>
</tr>
<tr>
<td>Euro Core</td>
<td>11</td>
<td>0.45</td>
<td>4</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

#### 5.6 Functionality of the RM algorithm under transient overload

In real systems, it will not necessarily be the case that all tasks arrive simultaneously – the worst case assumption made in eqn. (5.4). Furthermore, computation times $C_i$ and task periods may vary randomly, a situation that might result in transient overload of the processor [Leh89]. From a QoS provisioning perspective it is important to determine the degree of overload for which all tasks can be scheduled subject to their delay constraint. The following calculation provides the framework for carrying out this analysis, extending the analysis from section 5.4. Within the context of the WR-OBS network architecture the term task, used in the following paragraph, represents the wavelength request send from one edge router for one specified CoS. The parameter $i$ is used to indicate the maximum number of tasks with $i = 1 \ldots N_{\text{edge}}(N_{\text{edge}}-1)-N_{\text{CoS}}$.

For the investigation of transient overload it is necessary to check whether all tasks of a given set $\{\tau_1, \ldots, \tau_i\}$ can be scheduled within time $t \in \{0, T_i\}$. The search can be simplified by restricting only on the processing requirements arising at scheduling points $S_i$, defined as:

$$S_i = \left\{k \cdot T_j | j = 1, \ldots, i; k = 1, \ldots, \left\lfloor \frac{T_i}{T_j} \right\rfloor \right\}$$

---

$^5$ European Optical Network
where the $S_i$ represent $T_i$ and all multiples of the periods of all other tasks $\tau_j$ with period $T_j$ with $j < i$ and $T_i > T_j$. For instance, if we consider a system with 3 tasks with periods $T_1 = 4, T_2 = 7, T_3 = 22$, the respective scheduling points of the task set are (time in arbitrary units):

- $S_1 = \{4\}$
- $S_2 = \{4, 7\}$
- $S_3 = \{4, 7, 8, 12, 14, 16, 20, 21, 22\}$

So for $S_3$ it is mandatory to check whether all tasks of type $\tau_1$ and $\tau_3$ can be scheduled within the period of $S_3$.

To check whether a task set $\{\tau_1, \ldots, \tau_i\}$, with $1 \leq i \leq n$ and $T_i < T_{i+1}$ can be scheduled, eqn. (5.4) is modified as follows. Define the cumulative processing demand of a task set $\{\tau_1, \ldots, \tau_i\}$ to be $W_i(t)$ over period $\{0,\ldots,t\}$:

$$W_i(t) = \sum_{k=1}^{i} C_k \left\lceil \frac{t}{T_k} \right\rceil. \quad (5.14)$$

The average processing demand over the same period for the same task set then becomes:

$$L_i(t) = \frac{W_i(t)}{t}, \quad (5.15)$$

leading to the following results:

1. A particular task $\tau_i$ can be scheduled by the RM algorithm for all times $t$ if and only if the minimum average processor utilization can be maintained below 1:

$$L_i = \min_{wS_i} L_i(t) \leq 1 \quad (5.16)$$

2. The entire task set $\{\tau_1, \ldots, \tau_n\}$ can be scheduled if and only if the maximum processor utilization $L_i$ as given in eqn. (5.14) remains below 1 for all instances:

$$L = \max_{1 \leq i \leq n} L_i \leq 1 \quad (5.17)$$

The level of sustainable overload the processor can tolerate depends on the statistical distribution of the periods $T$ with PDF $f_T(T)$ and the computation time $C$ with PDF $f_C(C)$, leading to the concept of the breakdown utilization [Leh89]. The breakdown utilization was defined as the processor load for which the latency requirements of a certain task (wavelength request) are no longer met. The exact value of the breakdown utilization depends on $f_T(T)$ and $f_C(C)$, so varies with topology (determining $T$), as well as with the RWA algorithm and its software implementation (determining $C$). Equations (5.16) and (5.17) would provide the bounds against which to check, for instance by
incorporating them into a programme, whether the latency requirements for a given network and an envisaged set of latencies for the different CoS within this network could be met.

5.7 Summary and conclusions

This chapter analysed the achievable scalability of the WR-OBS network architecture from the point of view of network control implementation. A central network control and management was assumed which also carries out the routing and wavelength assignment. Key for the operation under strict latency requirements is that requests arriving at the central node are processed such that the delay requirements can be met. It was shown that modifications of the rate monotonic and earliest deadline first algorithms, developed for real-time processing, allow scheduling of a large number of requests such that these constraints are met, and allow scalability for up to 190 core nodes and 3 classes of service. Within limits, the rate monotonic algorithm allows for operation under transient overload, i.e. burst arrivals of requests.

Work so far has been concentrated on the case of centralized network control and management due to the limitations of dynamic RWA algorithms. The design of distributed control schemes with high wavelength reuse and low blocking probability remains a challenge in this area. The actual bottleneck to the electronic functionality is not the processing itself, growing with Moore’s Law as shown in chapter 2. The real limitation comes from the memory access times required for table lookups, so improvements in the pipelining of requests and more efficient memory technologies are expected to provide significant enhancements in terms of network scalability for all types of dynamic networking.
5.8 References


Scalability of the WR-OBS network architecture with QoS considerations

Symposium (LCS 2001), University College London, UK, 10-11 Sep. 2001, pp. 163-166


Chapter 6  Experimental investigation of tuneable lasers for optical burst switching

6.1 Introduction

The implementation of all dynamic network architectures and specifically WR-OBS requires new sources to enable rapid assignment of electronic bursts to wavelengths at the edge of the network, fast switching matrices in the network core and receivers which are capable of extracting information from received bursts. Chapter 2 was concerned with the scalability of optical switching matrices, so the focus in this chapter is on the investigation of the properties of fast tuneable lasers used as the key component located in routers at the network edge.

The initial rationale for the development of tuneable lasers was sparing, i.e. replacement of DFB lasers in WDM systems, saving the need to double the number of lasers at all wavelengths installed in the edge routers [Col00]. However, the emergence of fast tuneable lasers makes these ideal for application as sources in dynamic wavelength-routed optical networks. The application of tuneable lasers in dynamic optical networks or even in the context of packet-switching is driven by research interests. The design challenge here is the achievement of fast switching, i.e. switching and wavelength stabilization on timescales less than a millisecond. In chapters 3 and 4 the burst lengths envisaged for the WR-OBS architecture are most likely in the range of 5...50 ms, so sub-ms operation would provide a negligible overhead for laser tuning.

In this work, new fast tuneable lasers were investigated for their applicability to the WR-OBS network architectures and possible impairments such as wavelength inaccuracy, limited tuning speed, limited side mode suppression ratio (SMSR) and power fluctuations. The SMSR is defined as the difference in power between the lasing wavelength and the main side mode, typically measured in dB [Mor99]. Four lasers, three sampled-grating distributed Bragg reflector (SG-DBR) and one grating-assisted codirectional coupler with rear sampled grating reflector (GCSR) laser were investigated. For details on the design and the operation of SG-DBR and GCSR type lasers, as well as their difference, see, for example, [Ama98].
The fast tuneable lasers used in the experiments had a high output power (> 3 dBm for the SG-DBR laser) and operated across the entire C-band. The different requirements for OPS and OBS networks with respect to the laser tuning speed were reflected in experiments with timescales on micro- to milliseconds, which were a new operating regime for fast tuneable lasers not covered before. Previous work solely focused on fast switching in the nanaosecond regime [Gri01, ODo01].

To date, there has not been systematic experimental investigation of optical burst-switching in a network context. However, a number of areas applicable to OBS have already been investigated in the course of optical packet-switching, in particular:

- the implementation of transmitters for packet transmission and the use of fast tuneable lasers for rapid mapping of packets or short bursts of fixed size (800 ns) rather than longer bursts [Gri01]
- optical, electrical or hybrid packet header extraction and processing to determine the forwarding address and switch output port, also in the context of label switching (simplified header information)
- implementation of optical buffering, or techniques to avoid them (chapter 2)
- fast optical switches and switching matrices (chapter 2)
- burst-mode receiver with fast clock recovery and an increased dynamic range to cope with amplitude fluctuations [Rub02]

The Hybrid Opto-electronic Ring Network (HORNET) project [Shr00] investigated the implementation of fast optical transmitters using rapidly tuneable lasers (GCSR type) as transmitters, and techniques for fast clock and data recovery using a single wavelength system operating at 2.5 Gb/s. By overdriving the current inputs to a GCSR laser wavelength switching times < 15 ns were demonstrated for constant packet durations of 250 ns (equivalent to 625 bits). Considering a clock recovery time of 40 ns, the useable packet length is 210 ns or approximately 65 bytes. The fast clock recovery in [Shr00] was achieved, as in the experiments in this work, by applying an RF mixing technique followed by a high-Q filter to extract the operating frequency. The focus of the work described here was on the investigation of the impairments caused by temperature induced wavelength drift when these sources were operated with longer bursts. Previous projects were more device centric, e.g. [ODo01, Sim02], and by solely concentrating on fast switching the impairments occurring on other timescales were neglected.
A novel switch architecture [Gri01] was proposed based on an arrayed-waveguide grating (AWG) combined with tuneable transmitters to form a 32x32 switching matrix operating at 40 Gb/s per channel. The switch is operated in synchronous manner with timeslots of approximately 1 µs (800 ns payload plus variable training sequence and guard band between slots), where slots are centrally assigned by a scheduler. The scheme is conceptually close to the WR-OBS architecture. The main difference is the short distance between ingress and egress of the switch (few metres), such that the propagation delay for signalling is negligible and does not affect performance. Also, the synchronous operation of the switch differs from the asynchronous establishment of end-to-end lightpaths, which is more demanding as a solution for the RWA problem must be found in an arbitrarily meshed network. The transmitters used in [Gri01] were (unspecified) multi-section tuneable lasers, tested for a maximum switching speed of 50 ns, but not tested for the occurrence of thermal drift for burst lengths longer than microseconds as measured in the experiments. The BER in this experiment was recorded only for continuous transmission at 40 Gb/s with a PRBS word length of $2^{31}-1$, and the resulting penalty of 1 dB only reflects the impairments caused by the AWG, but not those caused by the laser switching process.

The limits to the tuning time (or switching speed) of multi-section semiconductor tuneable lasers were experimentally investigated in detail in [ODo01, Yu02] for SG-DBR and GCSR type lasers. The lasers were only tested under continuous wave (cw) operation, but not with modulation of bit patterns. The maximum achievable switching time for these lasers was measured to be 25 ns. A frequency plan for 2,000 channels with 2 GHz spacing was devised; however, the SMSR in this case was as low as 25 dB. For an SMSR of 40 dB (typical for the SG-DBR lasers used in the experiments described in this thesis), the number of channels dropped sharply to 124 since the higher SMSR could not be achieved by most of the channels. By restricting the maximum switching currents used in the experiment to ≤ 10 mA, any temperature induced effects were avoided.

Experimental results in [Bha02], related to the previously mentioned switch architecture in [Gri01], investigated for a GCSR laser the minimum guard band between two consecutive packets required to avoid impairments such as power fluctuations. The minimum guard band was found to be 45 ns for packets with a constant length of 1 µs and less than $10^{-8}$ packet loss rate.
An experimental packet switching demonstrator was also built in the course of the WASPNET project [Hun99], based on an AWG with wavelength converters and FDL buffer for fixed length packets (412 ns packets with 300 ns payload at 2.5 Gb/s and SCM header) [Gui00/1, Gui00/2]. Successful transmission of packets was demonstrated for a cascade of 14 such nodes with a BER less than $10^{-9}$ (equivalent to Q-factor > 16 dB as reported in [Gui00]). In a previous, related recirculating loop transmission experiment, including wavelength conversion between two wavelengths modulated at 2.5 Gb/s with PRBS $2^{21} - 1$, penalty-free operation was reported for a cascade of 25 nodes (30 nodes for power penalty of 1 dB) [Tza99], demonstrating that AWGs could be a key component in future optical networks due to their low loss and crosstalk, important for achieving a large number of hops in an OBS or OPS network without regeneration.

A burst-mode transceiver with a fast tuneable laser in the transmitter and a burst mode receiver has been reported [Rub02] and experimentally demonstrated to operate in back-to-back mode for bit rates of 10 Gb/s. Tuning times for an unspecified tuneable laser of less than 50 ns were reported, using a pre-distortion technique to decrease the tuning time. For transmission fixed size cells were assumed, with a 150 ns long guard band for laser tuning and clock recovery, and payload sizes between 400 ns and 4 μs, leading to an overhead of 27%...3%, respectively. Penalties < 0.2 dB were reported in [Rub02] when comparing the BER of the switched signal with the case of continuous transmission.

Although transient effects in EDFAs caused by variations in the power and the number optical channels have been a concern in the past (see, for example, [Ric97, Dim99]), compensation techniques such as gain clamping have been proven to be a suitable solution for this problem (see, for example [Fen02]), and were, therefore, not investigated in this thesis.

In the experiments carried out as part of the thesis work, the performance in terms of wavelength stability, output power and SMSR was investigated for fast tuneable lasers (SG-DBR and GCSR). The timescales covered in this work stretched over a much broader range than in the OPS experiments described, e.g [Gri01]. For timescales ranging from 1 μs – 100 ms, in the experiments reported here wavelength drift was identified as a key issue when high switching currents. This directly affected the BER
and Q-factor measured for bursts at 10 Gb/s and gives rise to increased penalty with short bursts, as detailed later in this chapter. To improve the switching speed of the device fast wavelength locking was demonstrated using a novel optical injection locking phase-locked loop (OIPLL) technique for an SG-DBR laser [Sil02/2]. In the conclusions the implications of the experimental results on the optical network design are discussed.

6.2 Laser characterization and wavelength selection

The initial step in preparation of the experiments was the characterisation of the 3 lasers used in the switching experiment: two sampled-grating distributed Bragg reflector (SG-DBR) lasers and one grating-coupled sampled reflector (GCSR) laser. An additional, already characterised SG-DBR laser [Sil02/2] was used in the wavelength locking experiments since it had deliberately not been fitted with an optical isolator, unlike the lasers used for the experiments discussed in sections 6.2-6.6 who all contained optical isolators (30 dB). All lasers consist of four sections as shown in Figure 6.1. The main difference between the GCSR and SG-DBR type lasers is the fact that the front section of the GCSR laser (also referred to as coupler) is designed as a broader, but tuneable filter rather than generating a comb of wavelength. The lasing wavelength is obtained by matching the filter with one of the wavelengths generated by the rear grating.

The principle of operation, however, is identical to both types of lasers: A unique lasing wavelength is selected by the front and rear gratings (a filter section referred to as ‘coupler’ replaces the front section in the case of the GCSR), amplified when traveling through the gain section, and fine tuned in the phase section of the device.

The advantage of four-section devices over other types of tuneable semiconductor lasers such as temperature-controlled DBR or three-section devices [Ama98] is the wide tuning range, which was > 40 nm (1525-1570 nm) for all devices. During the characterisation process, the output wavelength, the optical output power and the SMSR were measured as functions of the four currents. As the values vary from

![Figure 6.1: Principle of a four-section fast tuneable semiconductor laser](image-url)
device to device, it is necessary to carry out a thorough characterisation for each individual laser. For the lasers described here, the characterisation process was carried out using remotely controlled current sources and an optical spectrum analyzer (OSA) for precise measurement of wavelength, output power and SMSR.

All lasers (SG-DBR and GCSR) were characterized for their wavelength, output power and SMSR as a function of the input currents for constant temperature (25 °C default value). The front and rear current were varied whilst phase and gain current remained constant. The lasers were characterized using a matrix of 101x101 points, with the resolution adjusted to the maximum allowed current settings.
6.2.1 SG-DBR laser

To obtain all wavelengths, the SG-DBR laser must be operated with currents which are higher than those required for the GCSR laser described in next section. The current setting for the front grating current was $I_{\text{front}} \leq 100 \text{ mA}$, for the rear grating $I_{\text{rear}} \leq 120 \text{ mA}$, at 1 mA resolution each, and for constant currents into the gain (100 or 150 mA) and phase section (5 mA) of the device. Figs. 6.2-4 show the output wavelength (resolution 0.1 mA), the optical output power and the side-mode suppression ratio (SMSR) for SG-DBR laser #2.

Switching with high current differences resulted in significant temperature variation and corresponding wavelength drift as detailed in chapter 6.7. This problem could be avoided by using lower current values, so that the device was re-characterized for current settings with $I_{\text{front}}, I_{\text{rear}} \leq 20 \text{ mA}$ and 0.2 mA resolution, again for constant gain (100, 150 mA) and phase (5 mA) currents. In this case, however, not all ITU wavelength channels across the C-band were available.

![Diagram showing wavelength versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA. The intra-modal boundaries were chosen to 0.4 nm (50 GHz).]
Experimental investigation of tuneable lasers for optical burst switching

Figure 6.3: SG-DBR laser #2: Optical output power versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA

Figure 6.4: SG-DBR laser #2: Side-mode suppression ratio (SMSR) versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, resolution 1 mA
6.2.2 GCSR laser

The current values used for characterization of the GCSR laser were $I_{\text{front}} \leq 10 \, \text{mA}$, $I_{\text{rear}} \leq 40 \, \text{mA}$, with a resolution of 0.1 and 0.4 mA, respectively for constant gain (100 mA) and phase current (5 mA) settings. The results of the characterization are shown in Figs. 6.5-7 for the wavelength, the output power and the SMSR.

A comparison to the SG-DBR laser (Figs. 6.2-4) shows that the GCSR laser achieves operation over the same range of wavelengths with a significantly lower current for the coupler and rear grating section, hence easing the impairments caused by wavelength drift during switching, as shown later. The characterization shows a more regular pattern compared to the SG-DBR laser: Figure 6.5 shows that the wavelength of principal mode groups (‘fans’) decreases with increasing front grating current, as the wavelengths within the mode groups also decrease with increasing rear grating current. This structure was less regular for the SG-DBR laser, most likely due to imperfections during device fabrication process.

![Figure 6.5: Wavelength versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear)](image-url)
Experimental investigation of tuneable lasers for optical burst switching

Figure 6.6: Output power versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear)

Figure 6.7: Side-mode suppression ratio (SMSR) versus front and rear grating current for constant gain (100 mA) and phase (5 mA) current, GCSR laser, resolution 0.4 mA (front) and 0.1 mA (rear)
The GCSR laser, however, has a lower output power (< -3 dBm) and lower SMSR (> -40 dB) than its SG-DBR counterpart, which would affect the performance in DWDM systems, e.g. by reduced extinction ratio or the need for higher gain optical amplifiers.

Power versus current measurement: When the gain current of the GCSR laser is constantly increased from 0-100 mA in the experiment, the wavelength shows a mode jump, and also a change in the output power as plotted in Figure 6.8-a. A closer inspection of the linear regime from 15-30 mA leads to a dP/dl ratio of approximately 0.0067 mW/mA (Figure 6.8-b) at 25°C.

![Figure 6.8: dP/dl for GCSR laser (a) with power dip due to mode hopping. Linear fit to the ramp (b) to determine the dP/dl for gain currents of 15-30 mA](image)

6.2.3 ITU wavelength selection

To extract a unique set of wavelengths, especially to comply with the ITU-T frequency specification (Appendix G), the measurement data was searched such that for each frequency only a particular wavelength with respective current settings was selected, subject to selection criteria such as output power, output power range or minimum SMSR.

The result of the selection process is a reduced data set; in most cases, however, this will not be sufficient, as several current combinations would satisfy the pre-defined conditions. Power uniformity is not only an important criterion for minimizing impact on transmission, but also to reduce transient processes caused by EDFAs for varying input power. Figs. 6.9-11 show the wavelengths of all three lasers under test which comply with a 50-GHz spacing: For the SG-DBR laser, the selection criteria were SMSR > 40 dB and -3 dBm < output power < 0 dBm, whilst for the GCSR laser the criteria had to be relaxed to SMSR > 30 dB and -6 dBm < output power < -3 dBm. From all possible data sets, the one which provided the highest output power was selected. According to these specifications, not all wavelengths in the C-band were
available in all three lasers, indicating that the devices used still suffer from imperfection. For the purposes of the envisaged experiments, all lasers provided a sufficient number of wavelengths to carry out the switching and burst-to-wavelength assignment experiments.

**Figure 6.9:** Selection of wavelength for the SG-DBR laser #1 according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 40 dB (a) and -3dBm ≤ output power ≤ 0 dBm (b)

**Figure 6.10:** Selection of wavelength for the SG-DBR laser #2 according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 40 dB (a) and -3dBm ≤ output power ≤ 0 dBm (b)
Experimental investigation of tuneable lasers for optical burst switching

Figure 6.11: Selection of wavelength for the GCSR laser according to ITU-frequency grid with 50 GHz spacing, with a minimum SMSR of 30 dB (a) and -6 dBm \( \leq \) output power \( \leq -3 \) dBm (b)
6.3 Switching characteristics – rise/fall time & wavelength drift

6.3.1 Introduction and experimental set-up

A transient effect in tuneable lasers is the occurrence of wavelength instability and wavelength drift whilst changing wavelengths as the device passes through a number of modes (Figs. 6.2 and 6.5 depicted the wavelength characteristics of the SG-DBR and GCSR laser, respectively), depending on the individual wavelength characterization plot. Our investigations showed that during the switching between wavelengths A and B, intermittent wavelengths are excited; during this period, in a DWDM system, this potentially leads to crosstalk in other channels, resulting in additional penalties or even channel breakdown. So the parameters of interest in this case are the number (power) of transient wavelengths excited as well as the total switching time for wavelength stability. Whereas the electronic switching process based on the sweeping electrons takes approximately 15 ns [ODo01], the thermally induced wavelength drift within the device takes considerably longer. This is mainly due to the time required for the temperature gradient between different sections of the device to stabilize. As shown in section 6.4.2, this process can take up to 50 ms. The transient wavelength drift due to thermal effects was investigated for all four lasers. To simulate real network conditions, the gain and the phase current were held constant, and changes were only made to the front (coupler) and rear grating currents, and the laser alternates periodically between two wavelengths. The experimental set-up used is shown in Figure 6.12.

Figure 6.12: Set-up for tuneable laser switching experiments to analyze temperature dependent wavelength drift
Triggering of the oscilloscope and the modulation signal to the laser diode controllers (LDC) was provided by a delay generator (Stanford Research Systems).

As the wavelength resolution of conventional optical spectrum analyzers (OSA) was insufficient (0.07 nm), a Fabry Perot interferometer (FPI, also referred to as an etalon) was used. The Micron Optics FPI used in the experiment had a free-spectral range FSR = 150 GHz (1.2 nm), a finesse F = 571 and a 3-dB optical bandwidth $\Delta f = 263$ MHz. The DC control voltage of $0 \rightarrow 10V$ was provided by the auxiliary digital-to-analogue converter (DAC) output of an EG&G 5210 lock-in amplifier.

An optical bandpass filter with 3-dB bandwidth of 3 nm was used to remove any potential wavelength ambiguity during the switching process. The optical signal was received by a DC-coupled photo-diode (located in the FPI housing) and analysed using a Tektronix TDS 210 digital oscilloscope.

6.3.2 SG-DBR high current (up to 130 mA)

As the wavelength drift occurred on millisecond timescales, burst lengths were increased to 90 ms and 110 ms in the two wavelengths, corresponding to a repetition rate of each cycle of 5 Hz. The switching process was observed for a wavelength of 1540.7 nm for the following current settings:

- $I_{\text{gain}} = 150$ mA
- $I_{\text{front}} = 1.1 \pm \{10...40\}$ mA
- $I_{\text{rear}} = 3.2 \pm \{10...90\}$ mA
- $I_{\text{phase}} = 0$ mA

The front and rear grating current were changed simultaneously such that the maximum current change would occur, $\Delta I = \Delta I_{\text{front}} + \Delta I_{\text{rear}}$. Figure 5.13 shows the wavelength drift for a burst duration of 90 ms and $\Delta I = 90$ mA, resulting in a wavelength drift of approximately 0.25 nm.
Figure 6.13: Switching of SG-DBR laser #2 to wavelength 1540.7 nm, using a burst duration of 90 ms. The electrical control signal (blue) is shown for reference.

Figure 6.14 shows the wavelength drift and related fall times for all values of $\Delta I$ investigated for a range of $30 \text{ mA} \leq \Delta I \leq 130 \text{ mA}$. As previously observed, wavelength drift is almost negligible for current changes $\Delta I \leq 30 \text{ mA}$. The wavelength drift increases approximately linearly for total current changes in the range from 30...90mA. Beyond 90 mA, the front current was fixed to avoid observed wavelength instability, and only the rear grating current increased. The wavelength drift observed in the experiment reaches 0.25 nm for $\Delta I_{\text{tot}} = 90 \text{ mA}$, which is an unacceptable value for WDM systems with close channel spacing such as 50 GHz (0.4 nm). The fall times observed are $> 30 \text{ ms}$ in all cases, showing some independence from actual total current change, and only increase to a value $> 40 \text{ ms}$ for the highest total current in the experiment of 130 mA.
6.3.3 GCSR low current (40 mA) operation

A similar set of experiments as described in the previous chapter was carried out to investigate temperature induced wavelength drift for a GCSR laser. Due to the maximum upper current limits for safe operation of the device, only the coupler current was changed by 40 mA for the following settings:

- $I_{\text{gain}} = 100$ mA
- $I_{\text{coupler}} = 7.8...47.8$ mA
- $I_{\text{rear}} = 11.1$ mA
- $I_{\text{phase}} = 0$ mA

The maximum wavelength drift observed during the experiment was 0.05 nm for a wavelength of 1519.9 nm and a rise time of approximately 10 ms. Comparison to the results for the SG-DBR laser (Figure 6.14) also shows a wavelength drift of approximately 0.05 nm for a total current change of 40 mA.

The conclusion from the results is that the GCSR laser is operational over the whole wavelength range without any significant wavelength drift due to the lower current input in the device compared to the values of the SG-DBR laser. The results also emphasize that low current operation is a prerequisite to achieve fast laser switching times with minimum penalty due to wavelength drift.
6.4 Compensation schemes for wavelength drift

There exist a number of solutions to compensate for the temperature induced wavelength drift observed previously:

- For front and back currents < 20 mA each, and constant gain current, the wavelength drift due to temperature change is negligible for all types of lasers. This is within the typical operating range of the GCSR laser, but presents a strong constraint for the SG-DBR lasers whose front and back currents were specified to operate up to 100 and 150 mA, respectively. Even under these settings it was not possible to address all ITU channels (as specified in Appendix G).

- Compensation by switching currents such that the total input current in the device remains constant. In the case of constant gain and (negligible) phase current, a constant sum of front and rear grating current limits the number of obtainable wavelengths.

  In the characterisation charts for wavelength/output power/SMSR the obtainable values would be located on a diagonal, resulting in a restriction to the obtainable wavelengths and power/SMSR combinations.

- Use of wavelength locker; but requires round-trip time in feedback loop, and adds complexity to the device and the control electronics. An example for the improvements achieved by wavelength-locking using an optical injection locking technique are described in chapter 6.7.

6.5 BER/Q and penalty measurements for different burst lengths

BER, penalty and Q-factor measurements were carried out for the SG-DBR laser #2 to investigate the impairments introduced by burst transmission in comparison to a continuously transmitted optical signal.

The characterisation in combination with a Mach-Zehnder (MZ) modulator is shown as an eye diagram in Figure 6.15. The following parameters were measured:

Bias: -6.96 V
RF signal pk-pk: 1.11 V
PRBS $2^{31}$-1
$P_{Rx,in}$: 1.2 dBm
$\lambda$: 1540.7 nm
Extinction ratio: 10.23 dB
6.5.1 BER analysis

The set-up of the BER analysis is shown in Figure 6.16. Wavelengths generated by the tuneable laser switch periodically to generate bursts of 2 µs – 100 ms length. This range covers burst duration as expected in WR-OBS (millisecond timescale, chapter 4), as well as short bursts as expected in architectures such as OBS-JET (Tab. 1.2). As the pulse pattern generator (PPG) in the experiment was not operated in burst-mode, the cw light is continuously modulated using an IOC MZ modulator at a bit-rate of 10 Gb/s. With a typical laser output power of 0 dBm, the output power after the modulator is −10 dB as a result of internal losses in the polarization controller (PC, 1 dB) and modulator (9 dB). The optical signal is then amplified using an EDFA, followed by an optical attenuator and narrow bandpass filter (NBF, $f_{3-dB} = 2$ nm) to filter the desired wavelength and to adjust the input power to the receiver.

In the receiver the optical signal is split by a 3-dB coupler; one output is used for optoelectronic conversion. The second output is used for fast clock recovery using a high-Q filter (narrow electrical bandpass filter) operating at 10 GHz, enabling clock recovery in < 100 ns, which is significantly faster than a phase-looked loop (PLL), which was shown to have a minimum settling time of ≥ 300 ns [Mir99]. At 10 Gb/s, 100 ns of traffic still account for about 1,000 bits before synchronization would be achieved.
Experimental investigation of tuneable lasers for optical burst switching

The BER analysis was carried out for bursts switching between 1547.8 nm and 1554.2 nm, a wavelength pair which had been identified to have a low difference in steady-state powers \(0.2 \text{ dB}, P_{1547.8 \text{ nm}} = -0.76 \text{ dBm}, P_{1554.2 \text{ nm}} = -0.56 \text{ dBm}\). Figure 6.17 shows the BER at 1547.2 nm as a function of the receiver input power for both the continuous case and for burst lengths from 2 \(\mu\text{s}\) to 100 ms. For constant receiver input power the BER decreases significantly with burst length, e.g. at a receiver input power of \(-15.25 \text{ dBm}\) from \(10^{-10}\) for the continuous case to \(10^{-4}\) for a burst length of 2 \(\mu\text{s}\). Deterioration is also visible in the eye diagrams in Figure 6.18, with the largest eye opening obtained for the continuous case, and closing of the eye for the burst operation with a maximum burst length of 100 ms (b) and a minimum burst length of 2 \(\mu\text{s}\) (c). In the latter case distortions occur both on the mark and the space level of the optical signal. Figs. 6.19 and 6.20 show the same parameters (BER and eye diagrams) for the second wavelength of 1554.2 nm, confirming in principle the same observations as for the first wavelength. From the eye diagrams for 1554.2 nm in Figure 6.20 significant distortion of the ‘1’-level is observed, arising either from wavelength instability or EDFA noise.
Figure 6.17: BER as a function of receiver input power for continuous transmission and burst lengths of 2 µs – 100 ms at 1547.8 nm
Figure 6.18: Eye diagrams for burst generation for the continuous case (a), a maximum burst length in the experiment of 100 ms (b), and a minimum burst length of 2 μs (c) at 1547.8 nm
Figure 6.19: BER as a function of receiver input power for continuous transmission and burst lengths of 2 µs at 1554.2 nm
Figure 6.20: Eye diagrams for burst generation for the continuous case (a), a maximum burst length in the experiment of 100 ms (b), and a minimum burst length of 2 µs (c) at 1554.2 nm
6.5.2 Q-factor and penalty measurements

The set-up used is the same as for the BER measurements described in Figure 6.16. The data signal is modulated onto the optical carrier provided by the SG-DBR laser at a bit rate of 10 Gb/s. The typical output power of the modulated signal of $-8 \ldots -10$ dBm is amplified by an EDFA/attenuator such that the optical input power into the photodiode is approximately $-3$ dBm. As the PPG cannot be gated the optical signal is modulated continuously whilst the wavelength of the tuneable laser is switched periodically to generate bursts, with burst lengths varying between 50 and 900 μs.

Figure 6.21 shows the required receiver input power as a function of the burst length required to achieve $10^{-9}$ BER for a wavelength of 1540.4 nm and PRBS $2^1-1$ word length. The penalty with respect to continuous transmission is approximately 3.7 dB maximum for 100 and 200 μs burst length, and gradually decreases to the same level as the continuous case for 800 and 900 μs long bursts.

![Figure 6.21: Receiver input power as function of burst length from 100 – 900 μs using the SG-DBR laser #2 with PRBS $2^1-1$ and extinction ratio of approx. 10 dB. The dashed line indicates performance for continuous transmission of signal](image)

Results for the measurement of the Q-factor using the SG-DBR laser #2 are shown in Figure 6.22 for a wavelength of 1540.8 nm and PRBS $2^1-1$, with burst lengths from 100 – 900 μs. The relationship between Q-factor and BER was already described in equation (2.6).
For practical use it is key to determine the optimum BER and the corresponding optimum decision voltage setting in the electrical receiver. As very low BERs (<10^{-12}) take a long time to measure\(^6\), a method was devised which allows to calculate the optimum Q-factor and hence the optimum threshold voltage. This \textit{Q-factor measurement} is based on systematic variation of the decision voltage at the receiver such that the BER is high (here 10^{-5}...10^{-10}) due to errors recorded either on the mark or the space pulses as the decision threshold is varied [Nor02]. Based on the inversion of equation (2.6), and using linear interpolation the optimum Q-factor and the corresponding threshold setting are determined (tail extrapolation [Thi00]).

For burst lengths of 900 – 300 \text{μs} the optimum Q-factor remains in-between values of 15 – 17. The excursion at 200 \text{μs} of a Q-factor of 22 is most probably within the experimental uncertainty, as the value decreases to approximately 16 for 150 \text{μs}. The Q-factor significantly reduces for 100 \text{μs} to a value of 11, and was not measurable for 50 \text{μs}, indicating that for the experimental configuration used the minimum acceptable burst length in this case was 100 \text{μs}. Since the current changes involved were very low (10 mA), the effect of temperature induced drift was assumed to be negligible (Fig. 6.14). Instead, it is assumed that this limitation reflects significant rise and fall times of the laser diode current controllers used (approx. 4.5 \text{μs} for 70-kHz modulation bandwidth).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{q_factor_burst_length.png}
\caption{Optimum Q-factor versus burst length for 100 – 900 μs using SG-DBR laser #2 at 1540.8 nm and a \textsuperscript{2}7-1 PRBS}
\end{figure}

\textsuperscript{6} At 10 Gb/s, a minimum of 1Tbit of data is required to measure a BER < 10^{-12}
6.6 *Rapid wavelength stabilisation of SG-DBR Laser using optical injection locking*

In the previous sections of this chapter it was shown that the SG-DBR lasers under test were prone to temperature induced wavelength drift when the operating current during a switching process changed by more than 20 mA (Fig. 6.14). In the transmission experiment at 10 Gb/s, this resulted in a significant penalty (Fig. 6.17 and 6.19), especially for short bursts. Hence, for operation in a WDM environment, fast and accurate wavelength stabilisation would be required to eliminate such penalties from the system. As depicted in Table 1.2, burst durations are envisaged to range from the microsecond to millisecond regime, depending on the implementation used. For optical packet-switching, where datagrams contain only a few hundred bytes (e.g. 389.5 bytes mean, chapter 3.6.1), the value is reduced to nanoseconds — and conflicting with minimum switching times observed on the order of 10...50 ns [Gri01].

Compensation techniques as briefly discussed in chapter 5.5 were initially aimed at compensating long-term drifts (months to years) due to ageing. Only with the advent of fast switching techniques it became clear that also the process of wavelength stabilisation (also referred to as *locking*) should occur on timescales smaller than those of the envisaged bursts. At the time of writing the fastest commercially available locking technique based on a FPI is reported to stabilise wavelengths in the sub-ms regime (INT1100 module, Intune, Ireland).

The results in this section were obtained by switching experiments using an SG-DBR tuneable laser in combination with fast locking of the laser to a pre-defined wavelength through an optical injection phase-locked loop (OIPLL) [Sil02/1, Sil02/2]. The objective of the experiments discussed in the following was to show that the OIPLL could provide wavelength locking and stabilization with respect to a reference wavelength with locking times significantly faster than 1 ms. It could be shown in fact that the wavelength drift occurring on a timescale > 50 ms for the free-running tuneable laser could be reduced to locking time < 10 μs when the OIPLL was used. This value holds when the drifting wavelength entered the locking range of the OIPLL. However, an additional delay was observed between the trigger signal and the actual change in wavelength which — depending on the gain settings — was in the millisecond regime, resulting in overall switching delays in the millisecond regime.
The following section gives a description of the measurement set-up and the software used to monitor the wavelength drift and locking process are to follow. A set of experiments was carried out to establish whether the OIPLL was able to achieve stable wavelength locking for burst lengths up to several milliseconds, and whether the OIPLL would be able to provide wavelength lock faster than commercially available techniques.

6.6.1 Experimental set-up and control circuit for stabilisation

The experimental set-up in back-to-back configuration used for the wavelength stabilisation experiments is shown in Figure 6.23. A four-section SG-DBR (Marconi Caswell, now Bookham Technology) was continuously switched between two wavelengths by modulating the injection current into the rear section of the device, whilst the device was externally temperature controlled to 19°C. The current settings were as follows:

(constant) Gain current $I_g = 115$ mA
(constant) Front current $I_f = 16.2$ mA
(constant) Phase current $I_p = 0$ mA (matched)
(switched) Rear current $I_r = 5 \ldots 45$ mA

The current modulation was initially obtained from a 0.4 V peak-to-peak (pk-pk) modulation from a digital delay generator (DG, Stanford Research Systems), and fed into the external modulation input of the laser diode controller (LDC, ILX-3207B) with a conversion rate of 100mA/V. The switching time period was 400 ms, with 200 ms in each of the ‘High’ and ‘Low’ states, resulting in a switching frequency of 2.5 Hz. This low frequency was required to observe temperature drift phenomena on the order of 10’s milliseconds. Due to temperature instability of the SRS delay generator, the instrument was exchanged for an Agilent 33120A signal generator operated with a 1.3 Hz square wave and 0.4 V pk-pk amplitude. A further reduction in frequency was required for the wavelength to settle down completely to the target wavelength, allowing approximately 384.6 ms in both the ‘High’ and ‘Low’ state.

Operation of the 2nd order OIPLL [Sil02/2]: An optical frequency comb generator (OFCG) provides the reference wavelength (here: 1531.7 nm), which is fed into the SG-DBR laser via an optical circulator. The output wavelength is tapped (10%), and after opto-electronic conversion, fed into the microwave part of the optical injection phase-locked loop of the OIPLL. Once the tuneable laser wavelength entered the locking
range of the OIPLL – values were measured here were 18...25 GHz – an error signal ($V_{\text{error}}$) was generated at the IF output of the RF mixer, functioning as a phase comparator. The error signal is fed into the loop filter (first-order, Fig. 6.24), which provides the control signal ($V_{\text{gain}}$) for adjustment of the current of the laser gain section (100 mA/V conversion efficiency).

![Figure 6.23: Experimental set-up for fast wavelength-locking experiments using an SG-DBR laser; locking is obtained by using an optical injection phase-locked loop (OIPLL, 2nd order - loop path in blue), which uses an optical frequency comb generator (OFCG) as a reference](image)

Feeding the control signal of the OIPLL back into the gain section of the SG-DBR laser rather than to the phase section was chosen due to the larger locking range obtainable. For the settings chosen in the experiment, the laser showed linear behaviour around a current of 115 mA with a chirp of 882 MHz/mA and a useable range (maximum locking range) of 80 GHz between mode hops. The corresponding useable range for the phase section was maximum 5 GHz, which would have resulted in a much smaller locking range.

The actual locking range (18...25 GHz) of the OIPLL circuit was lower than the maximum locking range of 80 GHz. This design was chosen deliberately since a comb generator with dense spacing of frequencies (25 GHz, could be adjusted down to 18 GHz [Sil02/2]) was used for maximum spectral efficiency in DWDM systems. Increasing the locking range beyond the frequency spacing would result in instability of the OIPLL circuit.

The monitoring was carried out by using the same FPI filter as during the switching experiments (chapter 6.4). The wavelength resolution obtained in the experiment was
1.12 pm/mV. Prior to the FPI a tuneable optical bandpass filter (BPF) was installed to remove wavelength ambiguity introduced by the FPI. The optical bandwidth of the BPF was 2 nm. The output signal from the FPI is fed into a DC-coupled photodiode (PD) and observed on a digital oscilloscope (Tektronix TDS210), which is triggered from the same source as the rear grating modulation. Both the lock-in amplifier and the oscilloscope are connected to computer for remote control via GPIB/LabView. It records the signal trace of the oscilloscope for each wavelength setting, and generates a colour-graded 3D graph on the screen, where wavelength is plotted over time and intensity is indicated by the colour. This 3D plot is saved in matrix format on the hard drive.

The data is further processed to extract the timing for the maximum voltage (intensity) and the respective timing of each trace. This information is used to generate the wavelength vs. time and maximum power vs. time plots (both 2D) on the screen. At the end of the measurement, also the electrical control signal used for switching is recorded (for reference), and stored together with the wavelength vs. time and maximum power vs. time plots in a second file.

6.6.2 OIPLL loop filter design

The feedback signal into the modulator input of the current controller for the gain section of the laser was provided by a non-inverting integrator [Tit93], shown in Figure 6.24, and operating as a first-order filter. The resulting OIPLL circuit is, therefore, of second-order [Mir99].

\[
V_{\text{gain}}(t) = \frac{2}{RC} \int_0^t V_{\text{error}}(\tau) \cdot d\tau
\]

*Figure 6.24: Non-inverting integrator [Tit93] used as PLL loop filter in the OIPLL experiment, for provisioning of a feedback signal into the modulator input of the current controller of the gain section of the laser.*
Figure 6.25 shows the control signal \( (V_{\text{gain}}) \) which fed into the modulation input at the gain section current controller. The OIPLL was configured such that the time required to achieve wavelength lock was minimised.

The control circuit initially provides incremental changes to the current fed into the gain section of the laser. When the wavelength enters the capture range of the OIPLL, the wavelength is rapidly matched to the reference wavelength. This is indicated by the rapid increase ('jump') in the control signal depicted in Figure 6.25. Once the wavelength is matched to the reference wavelength, small changes are applied for the effect of temperature drift which occurred on the timescale of several tens of milliseconds. The noise of the observed in the experiment was caused by the high bandwidth of the feedback control circuit (approximately 1 MHz), and reduced in the plot in Figure 6.25 by averaging the trace 16 times. The results in section 6.7.3 indicate, however, that despite the noise in the feedback control circuit the wavelength locking was not affected and remained stable for the entire burst duration. The noise observed on the control signal is typical of second-order PLL operation, and could be significantly reduced by installing an additional capacitor in the loop filter [Mir99], leading to a third-order PLL circuit.

![Figure 6.25: Electrical control signal fed into the modulation input \( (V_{\text{gain}}) \), also converted in frequency deviation based on a chirp of 88 MHz/mV. The signal was averaged 16 times. The trigger signal (13 Hz) is shown for reference](image)
The control signal plotted here as voltage, was converted to a frequency offset by multiplying the conversion rate 100mA/V of the current controller with the chirp of 882 MHz/mA measured for the operating point, resulting in a wavelength change of 88.2 MHz/mV. Hence the changes in the control signal correspond to frequency changes in the GHz regime, and the total frequency deviation (minimum to maximum) observed in Figure 6.25 is approximately 17.5 GHz, compatible with the previously identified locking range values of 18...25 GHz.

6.6.3 Experimental results on microsecond wavelength-locking

Optimizing the control set-up resulted in better switching times with minimum OIPLL loop settling times. The experiments were carried out using the same set-up as shown in Figure 6.23. The current settings of the tuneable laser were the same as used previously, but the wavelength of the tuneable laser was adjusted by varying the temperature of the laser to ensure that it would fall into the measurement window of the scanning FPI. The front and rear sections of the tuneable laser were biased using a battery to reduce impact of noise on the wavelength stability.

Both the stability and speed of the wavelength locking mechanism are shown in Figure 6.26 plotted over a range of 0...5 ms when compared to a free-running laser. The wavelength modulation was driven by a square wave generator with 100 Hz frequency, resulting in effective switching windows of 5 ms per wavelength. The free-running laser (feedback switched off) showed a total wavelength drift in excess of 0.1 nm for 40 mA switching current over a period of 50 ms, whereas the tuneable laser shows fast locking (<100 μs) and high wavelength stability. The same trace was recorded with increased resolution the results of which are shown in Figure 6.27-a. The wavelength vs. time plot confirms the fast stabilisation of the wavelength, whereas the power plot in Figure 6.27-b indicates that power stabilisation is achieved after approximately 17 μs (3 μs delay plus 14 μs rise time). The total switching speed was below 20 μs.
Figure 6.26: Wavelength drift of the free-running tuneable laser versus the case when wavelength locking with OIPLL is operational. Wavelength measured with 2 pm resolution.

Figure 6.27: Wavelength locking for the same wavelength as in Figure 6.26 (a), and corresponding power reading and trigger signal (b), indicating that the switching and wavelength stabilization process are carried out in < 20 µs.
The system performance slightly deteriorated when the switching speed was increased to 1 and 10 kHz, shown in Figs. 6.28 and 6.29, respectively.

For a switching speed of 1 kHz the time before lock was achieved was approximately 40 μs (Figure 6.28-a), with power stabilisation achieved after approximately 50 μs (Figure 6.28-b). The wavelength and power plot also indicate that before final lock was obtained the OIPLL may have locked to a wavelength neighbouring the reference wavelength.

Similar values for wavelength locking (Figure 6.29-a) and power stabilisation (Figure 6.29-b) were observed in the experiments with 10 kHz switching times, resulting in a ‘burst’ length of approximately 50 μs. In particular the graph for the optical output power (Fig. 6.29(b)) shows a rise/fall time of approximately 10 μs. This value is mainly due to the bandwidth limitations of the current sources (1 MHz modulation bandwidth) and the feedback control circuit, implying that wavelength locking could not be achieved at fast switching speeds.

![Figure 6.28: Wavelength locking (a), and corresponding power reading and trigger signal (b) for a switching speed of 1 kHz](image-url)
Experimental investigation of tuneable lasers for optical burst switching

Figure 6.29: Wavelength locking (a), and corresponding power reading and trigger signal (b) for a switching speed of 10 kHz
6.7 Summary and conclusions

In this chapter experiments and results were described on the investigation of fast tunable lasers for applications as sources in edge routers of OBS networks. Two types of lasers, SG-DBR and GCSR, were used.

The main finding was that in the SG-DBR laser a temperature-induced wavelength drift was observed, occurring on millisecond timescales and for switching currents > 20 mA. Wavelength deviations of > 0.2 nm were observed for current changes in excess of 90 mA, leading to detrimental effects in WDM systems. The problem of wavelength instability, both through drift and during the switching process itself, was observed in BER and Q-factor measurements. The measurements were carried out systematically for a much wider range of burst length (2 μs – 100 ms), and showed increase of penalty with decreasing burst length. The experiments demonstrated that fast tuneable lasers are compact devices which can offer fast wavelength provisioning in future dynamic network architectures. Improvements are expected in the fast switching stabilisation, key for DWDM operation, and in the output power and SMSR (especially for the GCSR laser) as required in optical communication systems operating at bit rates of 10 Gb/s and beyond.

To compensate for the induced wavelength drift, an optical injection phase-lock loop was used in a novel approach for fast wavelength stabilisation. For an SG-DBR laser it was shown that locking times < 15 μs were obtained, limited by the electronic feedback circuit and the external current controller only. Faster operation and noise reduction in the control signal is expected from upgrading the OIPLL loop filter from first- to second order. At lower RF frequencies, a third-order lowpass Bessel filter was already shown to provide a low noise output and to improve loop setting times by 25% [Mir99] over conventional design. Fast digital to analogue (DAC) converters as current sources for the tuneable laser would also be key to obtain the switching speeds in the nanosecond regime reported by other research groups [Klo02]. Finally, novel tuneable laser designs (digitally switched DBR [DS-DBR] laser) were proposed with simplified wavelength selection [Rei02], helping to simplify the complex laser characterization process.

The experimental results described in this chapter have implications for the design of OPS and OBS networks. SG-DBR lasers with high output powers as required for DWDM transmission systems showed significant temperature induced wavelength-drift when operated with switching current differences > 30 mA. The drift could become
larger than 0.2 nm and occur over timescales of several ten milliseconds. This could result in significant interchannel crosstalk and increased BER when used in DWDM systems [Kei99]. This detrimental effect could be successfully eliminated by a novel scheme using an OPLL technique with wavelength locking times < 15 μs; this value can potentially further reduced to values below 50 ns achieved by other groups, when improving the set-up as discussed above. However, 50 ns seems to be the minimum achievable switching time measured independently by several research groups for switching times from any channel to any other channel in tuneable lasers [ODo01, Gri01, Rub02], confirmed by switching-related packet loss measurements in [Bha02]. Assuming 50 ns to be a hard limit for the technically achievable tuning time would have a number of severe implications on the network design:

- Assuming a minimum overhead for the guard band of less than 10%, the minimum packet length in OPS networks would become 500 ns. At a bit rate of 40 Gb/s this corresponds to 2500 bytes, i.e. the packet must be longer than a maximum Ethernet frame (1,518 bytes) or longer than 47 ATM cells (53 bytes). This would mean that for efficient bandwidth utilization even in OPS networks most likely some form of grooming would be required – leading to a network operation scheme similar to OBS!

- OBS networks, and particularly WR-OBS (Tab. 1.2), would be able to tolerate the limitations in the switching time since they assume burst lengths to be in the microsecond regime. For the WR-OBS network architecture with burst lengths in the millisecond regime (chapter 4) even wavelength-locking times of 15 μs as reported in section 6.7 would result in an overhead of less than 1%.

- Assuming a guard band of 50 ns between packets or bursts would also mean that fast optical switches used in OPS networks (chapter 2) would have to operate on the same timescale, i.e. the header lookup and switch reconfiguration process should take no longer than 50 ns.

- Finally, the limit of 50 ns would provide the target value for fast clock and phase recovery in packet or burst mode receivers for OBS and OPS networks.
6.8 References


Experimental investigation of tuneable lasers for optical burst switching


Chapter 7  Summary and conclusions

The work in this thesis describes the requirements for future optical network architectures which will have to accommodate both time-critical and data-centric traffic in the same physical layer whilst achieving significant gain in reusing resources from aggregation of traffic at the network edge.

The analysis of network traffic growth showed that the growth rates in the US backbone increased steadily at approximately 100 ± 10% p.a. over the past decade. In terms of volume, data traffic was predicted to match voice traffic at the end of the year 2002. However, given also that voice contributes over-proportionally to operators’ revenues, a network architecture is required which not only focuses on a data-centric approach, but provides the performance required to operate simultaneously significant amounts of delay-sensitive traffic.

The static wavelength-routed optical networks (WRON), optical packet switched (OPS) and optical burst switched (OBS) networks were discussed as potential candidates for dynamic future architectures. Based on their shortcomings, the wavelength-routed optical burst switched (WR-OBS) architecture was proposed, and its functionality analysed.

The performance of electronic and optical processing and switching technologies to support this traffic growth and new applications was investigated in chapter 2. Electronic processing of information grew at a steady 60% p.a., which does not match traffic growth rates in excess of 100% p.a. over the last decade. In the analysis of optical switching fabrics it was shown that optical switching matrices composed of individual 2x2 elements can be scaled to port sizes of 1000x1000, but that this is achievable only for a few architectures.

The burst aggregation process at the network edge was investigated in depth in chapter 3 of this thesis. Although edge delays are clearly dependent on traffic statistics, the goal was to devise a mechanism which automatically adapts to the incoming traffic statistics without explicit knowledge of their statistical nature. This aim was achieved by devising an adaptive burst mode assembler whose performance was successfully tested. The chapter also provides a review on short-range dependent (SRD) and long-range dependent (LRD) traffic statistics, and their impact on network performance, showing that for real network scenarios multiplexing gains are achievable even under
self-similar input. The numerically most precise results to date were the Likhanov-Mazumdar estimates (App. E) under strict bounds for the underlying traffic processes.

The operational advantages of a dynamically operated optical network were investigated in chapter 4. A number of new performance measures were introduced, namely the wavelength holding time, the lightpath utilization and the wavelength reuse factor. It was shown that for optimum performance a compromise is to be found between the lightpath utilization and wavelength reuse, leading to a novel performance parameter. Based on the analytical model, it was shown that significant gains in resource utilization could be obtained with wavelength reuse factors up to 40, though these values depend on the round-trip time between edge router and network control node as well as the bit rates used. For large networks (diameter > 1,500 km) it was shown that the static WRON provides better performance than the dynamic architecture.

The scalability of the dynamic network architecture critically depends on the ability of fast wavelength provisioning. As shown in chapter 5, real-time systems theory was applied for the first time in the context of rapid end-to-end lightpath assignment. Dynamic network scalability of up to 190 nodes for 3 CoS was achieved, depending on the calculation time per wavelength request. This processing time was shown to be obtainable when carrying out the route and wavelength table lookup process in a parallel manner. Parallel processing was shown to be key since the lookup problem for serial processing scales O(N^3), where N is the number of nodes, leading to unacceptable performance.

The practical constraints under which dynamic network architectures operate were experimentally investigated in depth in chapter 6 of this thesis. Fast tuneable lasers (SG-DBR) were operated at switching speeds in the range of 1 μs to match the expected burst specifications. Especially the high currents required for switching in the SG-DBR lasers resulted in significant wavelength drift over timescales relevant to the WR-OBS architecture (0.2 nm over 50 ms), leading to significant penalties in BER measurements at 10 Gb/s, where the maximum usable burst length was limited to 150 μs. Fast optical injection locking was demonstrated to provide wavelength stabilization within 15 μs, limited by the electronic switching speed of the set-up.
The following original contributions were made by this thesis to the field of dynamic optical networks, supported by the publications listed in the next chapter. Numbers refer to the key references in that chapter.

- A novel design for a dynamic optical network architecture was proposed, termed wavelength-routed optical burst switching (WR-OBS) [8].
- The investigation of the queueing processes in OPS and OBS networks under short- and long-range dependent traffic statistics showed that the impact of LRD dependent traffic is limited to the large buffer regime, and can be successfully compensated by multiplexing [16].
- New dynamic network performance parameters were introduced to quantify the achievable gain in resource utilization compared to static network architectures. For the WR-OBS architecture significant gain in reuse (up to factor of 40) was achieved over a static WRON [11, 14].
- Acknowledgements used in the wavelength reservation process limit the maximum network diameter for which dynamic network architectures provide operational gain over static networks. The analytical model for network performance evaluation provides these limits [14].
- It was shown that a central network control node can be scaled to 190 core routers, whilst meeting strict latency requirements, when using a request scheduling mechanism based on the rate monotonic and earliest-deadline first algorithms [16].
- The problem of temperature drift in fast tuneable SG-DBR and GCSR lasers was systematically investigated over timescales from micro- to milliseconds, both for continuous wave (cw) and modulated operation at 10 Gb/s to establish transmission penalties.
- Fast wavelength locking during the switching process of a fast tuneable laser was achieved with a new technique using an optical injection phase-locked loop (OIPLL). Locking times < 15 μs were obtained, with potential further reduction through improved laser diode current sources [18].
The research work presented in this thesis could be extended in the following areas:

Numerically more accurate models for the prediction of packet loss and delay in the presence of heavy-tailed input traffic statistics would be beneficial for the modelling of the queueing process at the network edge. In the case of OPS networks the models would be derived for a queue with constant depletion, whilst for the case of OBS networks the queue is only emptied periodically, which needs to be reflected in modified models.

*Analytical models* will be key for the evaluation of future network architectures due to the computational complexity of Monte Carlo modelling techniques. An essential extension to the model of the WR-OBS network architecture in this thesis would include the investigation of the performance of higher-layer applications such as TCP. Due to the delay introduced by burst aggregation process it is expected that the burst aggregation delay and the TCP window size would need to be adjusted to achieve high throughput. These problems would also require confirmation by simulation for selected cases and network topologies.

Future work in the area of *network scalability* would focus on novel, potentially distributed dynamic RWA algorithms with fast processing and low blocking probability. Research in this area would include practical implementation aspects, particularly whether the required lookup tables can be fitted into memory with fast access times such as SRAM.

Future *experimental work* will concentrate on further reduction of the switching times for the tuneable lasers into the nanosecond regime by improving the current supply to the laser; pulsed overdrive could be used for faster carrier sweeping of the active section of the laser. Finally, the burst switching experiments should be carried into the *network context* by performing recirculating loop experiments, using advanced AWG and free-space router techniques as large-port routing devices.
Chapter 8  List of publications

The following is a list of publications that arose as a result of the research described in this thesis.

Refereed International Journals and Conferences


Other Conferences with Published Abstracts


Appendix A - Queueing theory

A.1 Single station queueing systems

This appendix introduces the key notations and characteristics of queueing systems with a single queue of finite or infinite length and m servers as shown in Figure A.1, and used in chapter 3. The application in the WROBS network architecture is in the burst aggregation process at the network edge. This appendix provides a comprehensive overview of the results achieved for queueing systems based on Markovian models [Bol98]. These have been widely used for the description of voice traffic, but have also been applied for preliminary description of optical packet [Dan98, Hun99] and burst-switching networks [Yoo00]. The queueing models described in this appendix apply to problems in OPS networks with short-range dependent (SRD, section C.2) input traffic. Subject to a high degree of multiplexing and conditions to the buffer size, the models described in this appendix can also be applied to problems involving long-range dependent traffic as subsequently shown in Appendices C and E.

\[
\begin{align*}
\text{Figure A.1: Service station with single queue and m servers} \\
\text{Any queueing system is analyzed subject to a number of assumptions, listed briefly: The arrival process describes the period between two consecutive tasks (in communication systems: bursts or packets, or requests), and the process is described by the mean interarrival time } T_A = E\{A\} = 1/\lambda, \text{ where } E\{\cdot\} \text{ denotes the mean of random process } A, \text{ and } \lambda \text{ the mean arrival rate. Arriving tasks are processed with service rate } \mu \text{ in each server, with } T_B = E\{B\} = 1/\mu, \text{ where } B \text{ is a random process for the service time distribution. If there is no free server when a new task arrives, it is stored in the queue. The queueing discipline also determines the system performance, as described in the following section.}
\end{align*}
\]
A.2 Kendall's notation and different types of queueing systems

Kendall's notation is widely used to describe queueing systems with a single queue:

\[ A/B/m/K - \text{queueing discipline} \]

where \( A \) indicates the arrival process, \( B \) the service time process, \( m \) is the number of servers available, and \( K \) the queue length in the case it is finite (and discrete). Often a shortened version of the form \( A/B/m \) is used assuming infinite queue length and FCFS queueing discipline. The symbols listed in Table A.1 are used to describe the statistical properties of the independent and identically distributed (i.i.d) processes \( A \) and \( B \).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Exponential distribution</td>
</tr>
<tr>
<td>( E_k )</td>
<td>Erlang distribution with ( k ) phases</td>
</tr>
<tr>
<td>( H_k )</td>
<td>Hyperexponential distribution with ( k ) phases</td>
</tr>
<tr>
<td>( C_k )</td>
<td>Cox distribution with ( k ) phases</td>
</tr>
<tr>
<td>D</td>
<td>Deterministic distribution, i.e. constant interarrival or service time</td>
</tr>
<tr>
<td>G</td>
<td>General distribution</td>
</tr>
<tr>
<td>GI</td>
<td>General distribution with independent interarrival times</td>
</tr>
</tbody>
</table>

*Table A.1: Widely used symbols to describe the arrival and service time processes in queueing systems [Bol98]*

The different queueing disciplines are listed in Table A.2, but the FCFS/FIFO is the most widely used discipline.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Queueing discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCFS</td>
<td>First-come-first-served (FIFO)</td>
</tr>
<tr>
<td>LCFS</td>
<td>Last-come-first-served</td>
</tr>
<tr>
<td>SIRO</td>
<td>Service-in-random-order</td>
</tr>
<tr>
<td>RR</td>
<td>Round Robin</td>
</tr>
<tr>
<td>PS</td>
<td>Processor Sharing</td>
</tr>
<tr>
<td>IS</td>
<td>Infinite number of servers</td>
</tr>
<tr>
<td>Static Priorities</td>
<td>Priorities permanently assigned to a class of jobs</td>
</tr>
<tr>
<td>Dynamic Priorities</td>
<td>Priorities even for the same class change with time</td>
</tr>
<tr>
<td>Preemption</td>
<td>If LCFS or a priority discipline are used, the current job is interrupted and replaced by a newly arriving task</td>
</tr>
</tbody>
</table>

*Table A.2: Abbreviation and descriptions of widely used queueing disciplines [Bol98]*
In all systems with an infinite buffer (as was assumed in chapter 5 for the queueing of wavelength request packets), the waiting time (queueing delay) is the key parameter. In systems with finite buffer size, the metric is packet loss. Performance measures generally used in the queueing analysis:

- Probability of the number of jobs in the system \( \pi_k \)
  \[ \pi_k = P[k \text{ tasks in the queueing system}] \]

- Utilization \( \rho \), for \( m \) servers in the system
  \[ \rho = \frac{\lambda}{m \cdot \mu} \quad \text{with } \rho < 1 \]  
  \( \lambda = m \cdot \rho \cdot \mu \)

- Throughput \( \lambda \)
  \[ \lambda = m \cdot \rho \cdot \mu \]

- Response time \( T \): also known as sojourn time, i.e. the total time a task spends in the queueing system

- Waiting time \( W \): Time a task spends in the queue until it is ready to be serviced

  Response time = waiting time + service time

  The following formula holds for the means:
  \[ T = W + \frac{1}{\mu} \]  
  \( T = \frac{W}{1 - \rho} \)  
  \( W = \frac{1}{1 - \rho} \)

- Queue length \( Q \), i.e. the number of tasks in the system (assuming that all tasks/bursts/packets/requests are of equal length)

- Number of jobs in the system, \( K \). The mean number of jobs in the system, \( \bar{K} \), and the mean queue length \( \bar{Q} \) are given by Little's theorem based on the mean arrival rate \( \lambda \):
  \[ \bar{K} = \lambda \cdot T \]  
  \[ \bar{Q} = \lambda \cdot W \]

In the following a listing of widely used queueing systems is provided, which are relevant for the modelling of buffering in OPS networks.

### A.2.1 M/M/1 system

The M/M/1 queueing system is the most popular since it describes the call arrival and call holding process in telephone networks, where both processes are negative exponentially distributed, and the queueing discipline is FCFS. The system can be modeled as a birth-death process using a continuous time markov chain (CTMC)
process. Assuming that \( \lambda < \mu \), the steady-state probability \( p_0 \) for the system being empty is:

\[
\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k} = 1 - \frac{\lambda}{\mu}
\]  

(A.5)

The probability for \( \pi_k \) for \( k \) tasks in the system is

\[
\pi_k = \left( 1 - \frac{\lambda}{\mu} \right) \left( \frac{\lambda}{\mu} \right)^k
\]  

(A.6)

The queueing parameters are calculated as:

\[
\bar{K} = \frac{\rho}{1 - \rho}
\]  

(A.7)

\[
\bar{T} = \frac{1}{1 - \rho}
\]  

(A.8)

\[
\bar{W} = \frac{\mu}{1 - \rho}
\]  

(A.9)

\[
\bar{Q} = \frac{\rho^2}{1 - \rho}
\]  

(A.10)

The formulae also hold for the M/G/1-PS and the M/G/1-LCFS preemptive system. The M/M/1 system is often used as a reference model for switches with infinite queue sizes.

### A.2.2 M/M/∞ system

\[
\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k} = e^{-\frac{\lambda}{\mu}}
\]  

(A.11)

\[
\pi_k = \frac{\left( \frac{\lambda}{\mu} \right)^k}{k!} \cdot e^{-\frac{\lambda}{\mu}}
\]  

(A.12)

\[
\bar{K} = \frac{\lambda}{\mu}
\]  

(A.13)

\[
\bar{T} = \frac{1}{\mu}
\]  

(A.14)
A.2.3 M/M/m system

This queueing system operates with a constant number of servers \(m\) as supposed to an infinite number as was assumed in the previous paragraph:

\[
\pi_0 = \frac{1}{\sum_{k=0}^{m-1} \frac{(m \cdot \rho)^k}{k!} + \frac{(m \cdot \rho)^m}{m! \cdot (1 - \rho)}}
\]

(A.15)

\[
\pi_k = \begin{cases} 
\pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda}{(i+1) \cdot \mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \cdot \frac{1}{k!}, & 0 \leq k \leq m \\
\pi_0 \cdot \prod_{i=0}^{m-1} \frac{\lambda}{(i+1) \cdot \mu} \cdot \prod_{i=m}^{k-1} \frac{\lambda}{m \cdot \mu}, & k \geq m 
\end{cases}
\]

(A.16)

Steady-state probability for a task to wait is

\[
P_m = P(K \geq m) = \sum_{k=m}^{\infty} \pi_k = \frac{(m \cdot \rho)^m}{m!(1 - \rho)} \cdot \pi_0
\]

(A.17)

\[
\overline{K} = m \cdot \rho + \frac{\rho}{1 - \rho} \cdot P_m
\]

(A.18)

\[
\overline{Q} = \frac{\rho}{1 - \rho} \cdot P_m
\]

(A.19)

\(\overline{T}\) and \(\overline{W}\) are derived using Little’s theorem, and for \(\overline{W}\) the distribution function can be determined as

\[
F_W(x) = \begin{cases} 
1 - P_m, & x = 0 \\
1 - P_m \cdot e^{-m \cdot \mu (1 - \rho) x}, & x > 0 
\end{cases}
\]

(A.20)

A.2.4 M/M/1/K finite capacity system

One of the few systems with finite buffer size for which buffer overflow and resulting loss of tasks in the system becomes important. The maximum number of tasks in the system is limited to \(K\), otherwise the task is lost. Steady-state probability for \(k\) jobs in the system \((k = K)\) with \(a = \lambda/\mu\) (where the stability criterion \(\lambda < \mu\) does not have to be fulfilled as tasks can be refused to be admitted to the queue):

\[
\pi_k = \begin{cases} 
(1 - a)^k \cdot a^k, & 0 \leq k \leq K \\
\frac{1 - a^k}{1 - a^{k+1}}, & k > K 
\end{cases}
\]

(A.21)

The mean number of tasks is:
Note that the utilization $\rho < \lambda/\mu$ for a finite capacity queueing system, i.e. the utilization is lower than compared to systems with infinite buffer size due to packet loss. This result is significant as for the WR-OBS architecture, available memory size in edge routers is limited, and packet loss might occur due to buffer overflow. Under these circumstances it is key to design the queueing system such that overflow is minimized.

The results of the $\text{M}/\text{M}/1/K$ queueing system are limited, though, as they are only concerned with negative exponential distributions and are not valid for heavy-tailed traffic per se. However, the results for the $\text{M}/\text{M}/1/K$ system could provide an estimate for the lower bound of packet loss when input to the queueing system is composed of many bursty traffic flows with finite maximum bandwidth [Lik98]. An extension to the $\text{M}/\text{M}/1/K$ model discussed here is the $\text{M}/\text{M}/K/K$ queueing system, which is the basis for calculating the blocking probability or burst loss using the Erlang loss formula as shown in chapter 1.2.5.

### A.2.5 M/G/1 system

The $\text{M}/\text{G}/1$ system is applicable for the analysis of OPS networks since the arrival process is negative exponentially distributed and, therefore, light-tailed. The service times are generally distributed, and can take a wide variety of distributions, and could therefore be used to model traffic sources which are burstier (higher coefficient of variation) than a Poisson source.

Steady-state probability is

$$\pi_k = (1 - \rho) \cdot \rho^k$$

$$K = \rho + \frac{\rho^2}{1 - \rho} \cdot \frac{1 + c_B^2}{2}$$

$$Q = \frac{\rho^2}{1 - \rho} + \frac{1 + c_B^2}{2}$$

For some special cases the mean queue length can be calculated explicitly:

$$Q_{\text{M}/\text{M}/1} = \frac{\rho^2}{1 - \rho}$$

$$Q_{\text{M}/\text{D}/1} = \frac{\rho^2}{2 \cdot (1 - \rho)}$$
where $c_B$ is the coefficient of variation for the service time distribution, defined as ratio of standard deviation and mean of the stochastic process $B$.

Figure A.2 shows the comparison of the queueing systems $M/D/1$, $M/M/1$, $M/G/1$ ($c_B = 0.5$), $M/G/1$ ($c_B = 1.0$), $M/G/1$ ($c_B = 2.0$), $M/G/1$ ($c_B = 3.0$) using the mean number of tasks as a function of the utilization $\rho$. The number of jobs in the queue is minimum for the $M/D/1$, where no deviation of the service time occurs, and maximum for the $M/G/1$ system with a large coefficient of variation in the service time ($c_B = 3$ here).

![Figure A.2: Comparison of the mean number of tasks for different queueing systems as a function of utilization $\rho$ ($M/D/1$, $M/M/1$ and $M/G/1$ for $c_B = 0.5, 1, 2, 3$)](image)

**A.2.6 GI/M/1 system**

For calculating the results of the GI/M/1 system the Laplace transform $A(s)$ of the interarrival time distribution is required, based on which a parameter $s$ is introduced with:

\[ s = A(\mu - \mu s) \]  \hspace{1cm} (A.28)

The queueing system parameters are calculated as:

\[ \bar{K} = \frac{\rho}{1-\sigma} \text{ with } \sigma_k^2 = \frac{\rho \cdot (1 + \sigma - \rho)}{(1-\sigma)^2} \]  \hspace{1cm} (A.29)

\[ \bar{Q} = \frac{\sigma \cdot \rho}{1-\sigma} \text{ with } \sigma_Q^2 = \frac{\rho \cdot \sigma \cdot (1 + \sigma \cdot (1-\rho))}{(1-\sigma)^2} \]  \hspace{1cm} (A.30)
\[ \bar{T} = \frac{1}{\mu} \cdot \frac{1}{1 - \sigma} \]  \hspace{1cm} (A.31)

\[ \bar{W} = \frac{1}{\mu} \cdot \frac{\sigma}{1 - \sigma} \]  \hspace{1cm} (A.32)

with distribution function \( F_w(x) = \begin{cases} 1 - \sigma, & x = 0 \\ 1 - \sigma \cdot e^{-\mu(1-\sigma)x}, & x > 0 \end{cases} \)  \hspace{1cm} (A.33)

### A.2.7 GI/G/1 system

The GI/G/1 system assumes that tasks arrive with generally distributed independent interarrival times, and are served with general service times. This approach is one of the most general, so there exist only approximations or bounds for different values of \( c_A \) and \( c_B \). The systems M/G/1 and GI/M/1 can be regarded as upper and lower bounds for the GI/G/1 system as a function of \( c_A \) and \( c_B \), as listed in table A.3.

<table>
<thead>
<tr>
<th>( c_A^2 )</th>
<th>( c_B^2 )</th>
<th>M/G/1</th>
<th>GI/M/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1</td>
<td>&gt;1</td>
<td>LB</td>
<td>LB</td>
</tr>
<tr>
<td>&gt;1</td>
<td>&lt;1</td>
<td>LB</td>
<td>UB</td>
</tr>
<tr>
<td>&lt;1</td>
<td>&gt;1</td>
<td>UB</td>
<td>LB</td>
</tr>
<tr>
<td>&lt;1</td>
<td>&lt;1</td>
<td>UB</td>
<td>UB</td>
</tr>
</tbody>
</table>

*Table A.3: Upper and lower bounds (UB and LB, respectively) for the GI/G/1 queueing system [Bot98]*

The following upper bound exists:

\[ \bar{W} < \frac{\sigma_A^2 + \sigma_B^2 \cdot \lambda}{2 \cdot (1 - \rho)} \]  \hspace{1cm} (A.34)

A more refined version is exact for the M/G/1 system and also a good approximation of the GI/M/1 as well as the GI/G/1 system if \( \rho \) is not too small and \( c_A^2 \) and \( c_B^2 \) are not too big:

\[ \bar{W} < \frac{1 + c_B^2}{\left(1/\rho^2\right) + c_B^2} \cdot \frac{\sigma_A^2 + \sigma_B^2 \cdot \lambda}{2 \cdot (1 - \rho)} \]  \hspace{1cm} (A.35)

A lower bound is given by:

\[ \bar{W} > \frac{\rho^2 \cdot \sigma_B^2 + \rho \cdot (\rho - 2)}{2 \cdot \lambda \cdot (1 - \rho)} \]  \hspace{1cm} (A.36)

In the following four approximations are given which are either straightforward or accurate, but not computational complex:
• Allen-Cunnee approximation, exact for M/G/1:

\[
\frac{\rho}{\bar{W}} = \frac{\mu}{1 - \rho} \cdot \frac{c_A^2 + c_B^2}{2}
\]  
(A.37)

• Krämer/Langenbach-Belz approximation

\[
\frac{\rho}{\bar{W}} = \frac{\mu}{1 - \rho} \cdot \frac{c_A^2 + c_B^2}{2} \cdot G_{KLB}
\]  
(A.38)

with the correction factor \(G_{KLB}\):

\[
G_{KLB} = \begin{cases} 
\frac{2 (1 - \rho)}{\rho} \left( \frac{c_A}{c_A + c_B} \right), & 0 \leq c_A \leq 1 \\
\frac{1 - (1 - \rho) \left( c_A \right)}{e^\left( c_A \right) c_A + c_B}, & c_A > 1
\end{cases}
\]  
(A.39)

• Kulbatzki

\[
\frac{\rho}{\bar{W}} = \frac{\mu}{1 - \rho} \cdot \frac{c_A f(c_A, c_B, \rho) + c_B^2}{2}
\]  
(A.40)

where the function \(f(.)\) is defined as:

\[
f(c_A, c_B, \rho) = \begin{cases} 
1, & c_A = 0 \text{ or } c_A = 1 \\
\rho \cdot (14.1 \cdot c_A - 5.9) + (-13.7 \cdot c_A + 4.1) \cdot c_B^2 & 0 \leq c_A \leq 1 \\
+ \rho \cdot (-59.7 \cdot c_A + 21.1) + (54.9 \cdot c_A - 16.3) \cdot c_B & c_A > 1
\end{cases}
\]  
(A.41)

• Kimura: The following approximation holds particularly for \(c_A^2 < 1\):

\[
\bar{W} = \frac{c_A^2 + c_B^2}{2} \cdot \bar{W}_{\text{M/M/m}} \cdot \left[ (1 - c_A^2) \cdot e^{\frac{2(1-\rho)}{\rho} c_A^2} \right]^{-1}
\]  
(A.42)
This approach is limited, however, to all systems with arrival and service processes which have a finite variance and, hence, finite values for the coefficients of variation \( c_A \) and \( c_B \). The formulae only hold when the higher-order moments of the arrival and service processes exist, which is, for instance, not the case in the widely used Pareto distribution. The results are not applicable to heavy-tailed traffic in general; however, they could present a lower analytical bound for the delay and number of tasks/packets in the queue for traffic composed out of several heavy-tailed traffic streams.

Under conditions as set out in section 3.3 and Appendices C and E, however, the models discussed in this appendix can be applicable to problems involving LRD traffic inputs.

A.3 References


Appendix B - Traffic statistics

A key parameter for the analysis of future network architectures is the input traffic statistics. The stochastic characteristics of incoming traffic and the degree of statistical multiplexing determine the performance of queues and switches in the network.

In this appendix a number of discrete and continuous probability density functions (PDF) are listed. Discrete models are most useful for traffic generation on the packet level (as used in chapter 3), as each packet always contains an integer number of bits. Continuous models are used for fluid traffic sources, i.e. when modeling of on the individual bit-level is not required anymore, e.g. in high bit-rate systems (2.5, 10 Gb/s and beyond).

B.1 Discrete PDF

Table B.1 lists four important discrete PDFs, the Bernoulli, binomial, geometric and Poisson distribution with respective parameters, mean, variance and coefficient of variation. The PDFs for the RV X are given as:

Bernoulli: \( X \in \{0, 1\} : P(X = k) = p \) with \( 0 < p < 1 \)

Binomial: \( P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \) (B.1)

Geometric: \( P(X = k) = p \cdot (1-p)^{k-1} \) (B.2)

Poisson: \( P(X = k) = \frac{\alpha^k}{k!} \cdot e^{-\alpha} \), \( k \in \{0, 1, 2, \ldots\}, \alpha > 0 \) (B.3)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter</th>
<th>( \text{E}{X} )</th>
<th>( \text{Var}(X) )</th>
<th>( c_x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>( p )</td>
<td>( p )</td>
<td>( p \cdot (1-p) )</td>
<td>( \frac{1-p}{p} )</td>
</tr>
<tr>
<td>Binomial</td>
<td>( n, p )</td>
<td>( n \cdot p )</td>
<td>( n \cdot p \cdot (1-p) )</td>
<td>( \frac{1-p}{n \cdot p} )</td>
</tr>
<tr>
<td>Geometric</td>
<td>( p )</td>
<td>( \frac{1}{p} )</td>
<td>( \frac{1-p}{p^2} )</td>
<td>( 1-p )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
<td>( \frac{1}{\alpha} )</td>
</tr>
</tbody>
</table>

Table B.1: Discrete PDFs and their parameters, mean, variance, and coefficient of variation
B.2 Continuous PDF

Table B.2 shows a series of continuous PDFs which are used in queueing theory and the simulation of communication systems.
<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter</th>
<th>Mean $E(X)$</th>
<th>Variance $\sigma^2(X)$</th>
<th>Coeff. of variation $c_X = \frac{\sigma(X)}{E(X)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\mu$</td>
<td>$\frac{1}{\mu}$</td>
<td>$\frac{1}{\mu^2}$</td>
<td>$1$</td>
</tr>
<tr>
<td>Erlang</td>
<td>$\mu, k$</td>
<td>$\frac{1}{\mu}$</td>
<td>$\frac{1}{k \cdot \mu}$</td>
<td>$\frac{1}{\sqrt{k}} \leq 1$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\mu, \alpha$</td>
<td>$\frac{\alpha}{\mu}$</td>
<td>$\frac{\alpha}{\mu^2}$</td>
<td>$0 &lt; \frac{1}{\sqrt{\alpha}} &lt; \infty$</td>
</tr>
<tr>
<td>Hypoexponential</td>
<td>$\mu_1, \mu_2$</td>
<td>$\frac{1}{\mu_1} + \frac{1}{\mu_2}$</td>
<td>$\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2}$</td>
<td>$\frac{\sqrt{\mu_1^2 + \mu_2^2}}{\mu_1 + \mu_2} &lt; 1$</td>
</tr>
<tr>
<td>Hyperexponential</td>
<td>$k, \mu, q_i$</td>
<td>$\sum_{i=1}^{k} \frac{q_i}{\mu_i} = \frac{1}{\mu}$</td>
<td>$2 \cdot \sum_{i=1}^{k} \frac{q_i}{\mu_i^2} - \frac{1}{\mu^2}$</td>
<td>$\sqrt{2 \cdot \mu^2 \cdot \sum_{i=1}^{k} \frac{q_i}{\mu_i^2} - 1} &gt; 1$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\lambda, \alpha$</td>
<td>$\frac{1}{\lambda} \cdot \Gamma \left(1 + \frac{1}{\alpha}\right)$</td>
<td>$\frac{1}{\lambda^2} \cdot \frac{1}{\Gamma \left(1 + \frac{2}{\alpha}\right)} - \left(\frac{\Gamma \left(1 + \frac{1}{\alpha}\right)}{\Gamma \left(1 + \frac{2}{\alpha}\right)}\right)^2$</td>
<td>$\sqrt{\frac{\Gamma \left(1 + \frac{2}{\alpha}\right)}{\Gamma \left(1 + \frac{1}{\alpha}\right)^2} - 1}$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$\lambda, \alpha$</td>
<td>$e^{\frac{\lambda \cdot \alpha^2}{2}}$</td>
<td>$e^{2 \cdot \lambda \cdot \alpha^2} \cdot \left(e^{\alpha^2} - 1\right)$</td>
<td>$e^{\alpha^2} - 1$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$\alpha, A$</td>
<td>$\frac{A \cdot \alpha}{\alpha - 1}$</td>
<td>$8$</td>
<td>$8$</td>
</tr>
<tr>
<td>Cox ($c_X &lt; 1$)</td>
<td>$\mu_j = \mu, j = 1, \ldots, k$</td>
<td>$\frac{b_1 + k \cdot (1 - b_j)}{\mu}$</td>
<td>$\frac{b_1 + k \cdot (1 - b_1) \cdot (b_1 (1 - k) + k - 2)}{\mu^2}$</td>
<td>$\frac{b_1 + k \cdot (1 - b_1) \cdot (b_1 (1 - k) + k - 2)}{(b_1 + k \cdot (1 - b_1))}$</td>
</tr>
<tr>
<td>Cox ($c_X &gt; 1$)</td>
<td>$\mu_1, \mu_2, a$</td>
<td>$\frac{1}{\mu_1} + \frac{1}{\mu_2}$</td>
<td>$\frac{\mu_1^2 + a \cdot \mu_1 \cdot (2 - a)}{\mu_1^2 - \mu_2^2}$</td>
<td>$\frac{\mu_1^2 + a \cdot \mu_1 \cdot (2 - a)}{(\mu_2 + a \cdot \mu_1)^2}$</td>
</tr>
</tbody>
</table>

Table B.2: Continuous PDFs and their parameters, mean, variance, and coefficient of variation
Appendix C - Application of queueing theory for heavy-tailed distributions and finite buffer size

This appendix introduces some of the terminology used in chapter 3, and their mathematical definitions, especially with respect to burstiness, long-range dependence (LRD), and self-similarity. Since these termini sometimes are not strictly used, we start with a definition of the notations and emphasize essential differences in the terminology [Boo02]. The results presented in chapter 3 are those supposed to be the most significant ones with respect to their applicability to the design of OPS and OBS network architectures. There are two approaches to address queueing problems, which are the heavy-traffic theory, governed by the central-limit theorem (CLT), most applicable for systems with high load, and the large deviations (LD) theory, which is applicable to systems with low load and bursty traffic input, when the buffer overflow is most likely the result of a rare event. A third theory, known as moderate deviations (MD) theory, was introduced to combine the advantages of both the LD and the CLT approach and to simplify the asymptotics that govern both regimes, especially with respect to scaling.

We start this appendix by defining the terminology before providing the mathematical background for the analysis of queueing systems with large and short buffers.

C.1 Heavy-tailed distribution

A distribution $F$ is heavy-tailed (denoted by $F \in H$) if and only if for all $y \in \mathbb{R}$,

$$\lim_{x \to \infty} \frac{\bar{F}(x)}{\bar{F}(x - y)} = 1$$

(C.1)

with $\bar{F}(x) = 1 - F(x)$. Other common terms for heavy-tailed distributions are fat-tailed or long-tailed distributions. A property of all heavy-tailed distributions is that for all $\varepsilon > 0$:

$$\lim_{x \to \infty} \sup e^{\varepsilon x} \bar{F}(x) = \infty$$

(C.2)

A stochastic process is called light-tailed if its moment generating function (MGF) is finite for some $\theta > 0$:

$$M(\theta) = \int e^{\theta x} \cdot dF(x) < \infty$$

(C.3)
C.2 **Short- and long-range dependence**

A stochastic process is called *long-range dependent* if its autocorrelation decays slowly, i.e. shows a long memory. The mathematical definition is: Let \( \{X(t), t > 0\} \) be some stationary stochastic process with autocovariance function \( c(t) = \text{Cov}[X(0), X(t)], t \geq 0. \) If \( c \) is integrable then \( \{X(t), t \geq 0\} \) is called *short-range dependent* (SRD) and if \( c \) is not integrable then \( \{X(t), t \geq 0\} \) is called *long-range dependent* (LRD).

C.3 **Self-similarity**

A stochastic process \( \{X(t), t \geq 0\} \) is called (exactly) *self-similar* with Hurst parameter \( H \) if \( \{X(t), t \geq 0\} \) and \( \{\gamma^H X(\gamma t), t \geq 0\} \) have the same finite-dimensional distributions for all values of \( \gamma > 0. \) A stochastic process \( \{X(t), t \geq 0\} \) is called asymptotically self-similar if it converges, when suitably centered and normalized, in distribution to a self-similar process as \( \gamma \to \infty. \) Self-similarity and long-range dependence are related: For \( H > 0.5 \) a self-similar process is LRD.

When using a superposition of ON-OFF sources whose ON and OFF periods are Pareto distributed (App. B), the Hurst parameter \( H \) of the resulting traffic trace is defined as function of parameter \( \alpha \) with \( H = (3-\alpha)/2, \) where \( 0.5 < H < 1 \) for \( 1 < \alpha < 2. \)

C.4 **Classification of queueing theory**

Queueing theory is mainly about the scaling of processes, i.e. to investigate the queueing performance when scaling the buffer size. This scaling process is implemented either by aggregating a number of \( N \) flows with the same statistical properties (*many-flows* version), or by increasing/decreasing the duration over which the process is observed by factor \( N \) (*fast-time* version). Depending on the characteristics of the input traffic process, the queueing discipline is either governed by heavy traffic theory, leading to the central limit theorem, or the large deviations theory. The different assumptions of these two queueing theories differ significantly, it is difficult to combine both theories in one coherent traffic model. The moderate deviations theory was, therefore, introduced to provide a link between both theories. All three theories are introduced here briefly without the entire theoretical background, which can be found in [Wis01]. [Boo2] mainly covers results achieved by applying the LDP, which were used in chapter 3 to estimate which impact traffic statistics have on queueing systems with constant depletion.
**Scaling** is an inherent and important feature, distinguishing all three existing queueing theories described later in this appendix. When increasing the number of sources feeding into a single queue by a factor $n$, then the values of the original buffer length $b$ and the original depletion rate $c$ are assumed to be increased by the same factor. The depletion rate would be equivalent to the line rate of a communications channel. So the new, scaled buffer size becomes $B = n \cdot b$, and the scaled depletion rate is $C = n \cdot c$. This type of scaling was extensively used to explain recent findings of queueing theory for heavy- and light-tailed input statistics in OPS networks, section 3.4.

### C.5 Buffer scaling for heavy- and light-tailed traffic inputs in short and large buffers

This section provides the background for the discussion of the scaling of buffer size in OPS networks for short and large buffers (explained below) as discussed in section 3.3, and particularly the results in Table 3.1.

In the following we summarize results of recent analysis [Boo02], which for the aggregate of many sources — in brief — state that for short buffers the distribution used is irrelevant (and therefore Poisson), and that for large buffers the overflow probability follows the distribution of the ON periods. Although the queueing process considered for the WR-OBS architecture is different in the sense that buffers are emptied periodically, but not continuously, the results are thought to be relevant to be mentioned here. A short buffer could be regarded in this context as a buffer whose size is considerably smaller than the bandwidth-delay product, i.e. the product of the link bit rate (at the bottleneck) and the round-trip time. A large buffer would then be defined as a buffer with a size which is equal to or even in excess of the bandwidth-delay product.

The assumptions are as follows: The aggregate of $N$ identical ON-OFF sources is fed into a FIFO queue which is emptied at constant rate $C$. Both the ON and the OFF periods are identical and independently distributed (i.i.d.), and mutually independent RVs, called $A$ and $S$, with $A, S > 0$. The total amount of traffic generated by a single source in steady state over period $t$ is $A(t)$. Further: The RVs $A$ and $S$ are such that $E\{A_{\xi}^{\xi}\} < \infty$ (for $\xi > 0$) and $E\{S\} < \infty$ (finite means). The distribution of $A + S$ is non-lattice. Accordingly, the fraction of time for which the source is busy (ON) is

$$p = \frac{E\{A\}}{E\{A\} + E\{S\}}$$  \hspace{1cm} (C.4)
The residual activity period $A^*$ is defined as follows: conditioned that the process is in the ON state, $A^*$ has the following distribution:

$$F_{A^*}(x) = P(A^* \leq x) = \frac{1}{E\{A\}} \cdot \int_0^x P(A > y) \cdot dy$$  \hspace{1cm} (C.5)

Accordingly:

$$F_{S^*}(x) = P(S^* \leq x) = \frac{1}{E\{S\}} \cdot \int_0^x P(S > z) \cdot dz$$  \hspace{1cm} (C.6)

The performance metric is the probability of buffer overflow, $p(B, C)$, with buffer size $B$ and depletion rate $C$. The asymptotics stated in the following refer to the many sources estimate, i.e. the case when number of sources $N$ is large. The resources are rescaled such that $C = N^{-c}$ and $B = N^{-b}$. The system is non-trivial with $p < c < 1$.

In the scaled model,

$$p_N(b, c) = \text{steady-state probability that the buffer content exceeds level } N^{-b}. $$

Of particular interest is the decay rate or loss curve $I(b)$ as a function of $b$ for constant $c$:

$$I(b) = -\lim_{N \to \infty} \frac{1}{N} \cdot \ln(p_N(b, c)),$$  \hspace{1cm} (C.7)

with 

$$I(b) = \inf_{t > 0} \sup_{\theta} [\theta \cdot (b + ct) - \ln(E[e^{\theta A_0}])],$$  \hspace{1cm} (C.8)

and  

$$p_N(b, c) \approx e^{-N b} \text{ (exponential approximation).}$$  \hspace{1cm} (C.9)

### C.5.1 Queueing systems with small buffers – insensitivity to traffic characteristic

The small buffer regime was defined in [Boo02] to be valid when the state of an individual source is not likely to change during the trajectory to overflow, mainly because the time to overflow is small. Intuitively this would correspond with the previously stated assumption that this would be applicable to queueing systems whose size is smaller than the bandwidth-delay product. A more formal definition is that the probability that the state (ON or OFF) of an individual source makes two or more transitions in an interval of length $t$ is $O(t^2)$ with $t \to 0$.

For the small buffer regime the loss curve $I(b)$ is then as follows:

$$I(b) = \lim_{N \to \infty} \frac{1}{N} \cdot \ln(p_N(b, c)) = -\alpha(c) - \beta(c) \cdot \sqrt{b} + O(b)$$  \hspace{1cm} (C.10)

where $\alpha$ and $\beta$ are constants depending only on $p$ and $c$:

$$\alpha(c) = c \cdot \ln \left( \frac{c}{p} \right) + (1-c) \cdot \ln \left( \frac{1-c}{1-p} \right)$$  \hspace{1cm} (C.11)
Application of queueing theory for heavy-tailed distributions and finite buffer size

\[ \beta(c) = 2 \cdot \left[ \frac{c}{E(A)} + \frac{1-c}{E(S)} \right] \cdot \ln \left( \frac{c}{1-c} \cdot \frac{E(A)}{E(S)} \right) - 2 \cdot \left( \frac{c}{E(A)} - \frac{1-c}{E(S)} \right) \]  
\[ \text{(C.12)} \]

Interpretation of results:

- For a large number of sources, the loss curve decreases proportional to \( \sqrt{b} \), implying that even small buffers are useful since they absorb traffic fluctuations.
- The loss curve depends on the distributions of \( A \) and \( S \) only through the mean values, not their variances. Therefore, the results are independent on the distributions used for the ON and OFF periods, and the results obtained for the traditional models (Poisson arrivals, exponential holding times) are valid for heavy-tailed distributions also.

C.5.2 Queueing systems with large buffers

For the many sources, large buffer regime the results for I(b) depend on the ON time distribution \( A \). Results for light-tailed or short-range dependent (SRD) ON times differ from those for heavy-tailed or long-range dependent (LRD) as the loss curve follows the distribution of the ON times. The heavy-tailed distributions considered here belong to the class \( S \) of subexponential distributions and the class \( V \) of subexponentially varying distributions defined as follows [Boo02]:

\[ \text{Suppose the function } v_x(.) = -\ln P(x > t) \text{ is regularly varying of index } h \text{ (at infinity)} \]
\[ \text{for all } y > 0 \text{ such that} \]
\[ \lim_{t \to \infty} \frac{v_x(y \cdot t)}{v_x(t)} = y^h \]  
\[ \text{(C.13)} \]

In this case the RV \( x \) has a subexponentially varying distribution \( F_x(.) \in V \).

The classes \( S \) and \( V \) are not subsets of each other; the most important implementations of heavy-tailed distributions such as Pareto, Weibull and lognormal distribution belong to both classes.
- Light-tailed ON times:
  - For the case of light-tailed ON times the most likely path to overflow occurs when the source is ON at time \( t = 0 \), and stays ON during the entire period \([0, t]\). This assumption does not apply to heavy-tailed distributions. The loss curve I(b) is in this case:
  \[ I(b) - \theta^* \cdot b = \nu \]  
\[ \text{(C.14)} \]
  where \( \theta^* = \sup \{ \theta : \lim_{b \to \infty} t^{-1} \cdot \ln E[e^{\theta (A_0 - \xi)}] \leq 0 \} \]  
\[ \text{(C.15)} \]
  and \( \nu = -\lim_{t \to \infty} \ln E[e^{\theta^* (A_0 - \xi)}] \)  
\[ \text{(C.16)} \]
The result of equations (C.14-16) implies that only $A(t)$ and, therefore, only the ON time distributions affect the loss curve, irrespective of the distribution of $S$ (OFF times).

- Heavy-tailed ON times:

If $F_{A^*}(\cdot) \in S \cap V$, and with $v(y) = -\ln P(A^* > y)$, then

$$v(b) = -\ln P(A^* > b) = -\ln (1 - P(A^* \leq b)) = -\ln \left(1 - \frac{A^*}{b}\right) = -\ln \left(1 - \frac{A^*}{b}\right)^{\alpha - 1}$$

(C.18)

In the limit ($b \to \infty$) $I(b)$ is proportional to $v(b)$ in all three cases, and will follow the distribution of the residual activity $A^*$. If the ON times $A$ are subexponentially distributed, so $I(b)$ is sublinear (increases slower than linear) and the overflow probability decays subexponentially as well. Figure C.1 shows the plot of $v(b)$ against buffer size when ON times are Pareto distributed with parameters $A = 2$ and $\alpha = \{1.1, 1.5, 1.9\}$. For arbitrary units $v(b)$ becomes:

$$v(b) = -\ln P(A^* > t) = -\ln (1 - P(A^* \leq t)) = -\ln \left(1 - \frac{1}{\alpha} \cdot \left(1 - \left(\frac{A^*}{b}\right)^{\alpha - 1}\right)\right)$$

(C.17)

![Figure C.1: $I(b) \propto v(b)$ plotted for Pareto distributed ON times with parameters $A = 2$ and $\alpha = \{1.1, 1.5, 1.9\}$, showing sublinear increase for large values of $b$ for large buffer size. Absolute values are not relevant since the function $v(b)$ requires appropriate scaling](image-url)
The sublinear increase of $I(b)$ as a function of $b$ is evident, especially once the buffer size $b$ exceeds the mean value $E\{A\}$. The tail behaviour of the loss curve is completely governed by the distribution of the ON times; for the OFF times only the mean $E\{S\}$ is considered through parameter $p$ in equation (C.17).

### C.6 References


Appendix D - Power spectral density for discrete sources

This Appendix provides the mathematical derivation of the power spectral density of a discrete ON-OFF source, as it was used for traffic generation in chapter 3. The results presented in this appendix would allow to determine analytically the power spectral density of a single ON-OFF source. As the power spectral density is related to the autocorrelation function of the process through a Fourier transform, it would be possible to obtain the ACF even for sources with heavy-tailed ON and OFF periods, for which the ACF can be difficult (or impossible) to be calculated.

The results of this appendix could be used in combination with equation (3.11) to analytically determine the burst size variance for the burst aggregation process at the network edge.

D.1 Power spectral density of discrete ON-OFF source

The power spectral density of a discrete ON-OFF source is calculated based on the PDF and the probability generating function (PGF). The PGF is defined as the z-transform of the PDF, where the state-probability becomes the coefficient for the z-transform:

ON periods, RV X: PDF \(a(n) = P(X = n)\)  
PGF \(A(z) = \sum_{n=1} \cdot a(n) \cdot z^n\)  

OFF periods, RV Y: PDF \(b(n) = P(Y = n)\)  
PGF \(B(z) = \sum_{n=1} \cdot b(n) \cdot z^n\)

From the PGF the higher-order moments of the distribution can easily be calculated:

ON periods: \[ E\{X\} = \frac{dA(1)}{dz} \]  
OFF periods: \[ E\{Y\} = \frac{dB(1)}{dz} \]

so that the mean of the output load \(\rho\) of the source is:

\[ \rho = \frac{\frac{dA(1)}{dz}}{\frac{dA(1)}{dz} + \frac{dB(1)}{dz}} \]
The autocorrelation $R[n_1, n_2]$ of a discrete process $X[n]$ is defined as:

$$R[n_1, n_2] = \mathbb{E}\{X[n_1] \cdot X^*[n_2]\} \quad (D.8)$$

The corresponding autocovariance $C[n_1, n_2]$ differs by a constant offset $[\mathbb{E}\{X\}]^2$ only:

$$C[n_1, n_2] = R[n_1, n_2] - [\mathbb{E}\{X\}]^2 \quad (D.9)$$

For a stationary process the autocorrelation $R$ depends on the difference $m = |n_1 - n_2|$ only: $R[m] = R[n+m, n]$. \hfill (D.10)

The discrete power spectral density $S(\omega)$ and the $z$-transform of a stationary process are defined as:

$$S(z) = \sum_{m=-\infty}^{\infty} R[m] \cdot z^{-m} \quad (D.11)$$

$$S(e^{j\omega}) = S(e^{j\omega}) = \sum_{m=-\infty}^{\infty} R[m] \cdot e^{-jm\omega} \quad (D.12)$$

For stationary processes the autocorrelation function approaches a constant offset for large values of difference $m$:

$$\lim_{T \to \infty} R(\tau) = \rho^2 \quad (D.13)$$

Hence, the spectrum $S(\omega)$ contains a DC part (dirac impulse at $\omega = 0$) and an AC part (continuous spectrum for $\omega > 0$). To avoid the DC part, the spectrum can be expressed as function of the covariance spectrum $S_C(\omega)$ (denoted as $C(t_1, t_2)$ in chapter 3.4.1):

$$S(\omega) = S_C(\omega) + 2\pi\rho^2 \delta(\omega) \quad (D.14)$$

For this reason, often only the covariance spectrum is considered – as in the derivation of the power spectral density in the following paragraph.

It is possible to express the power spectral density $S(f)$ of a discrete source as a function of the PGF $A$ and $B$ [Lae97] for all processes for which the PGFs $A$ and $B$ and their first higher-order moments exist:

$$S(f) = \sigma^2 \cdot \left[1 + Q(e^{j2\pi f}) + Q(e^{-j2\pi f})\right] \quad (D.15)$$

where:

$$Q(z) = \sum_{m=1}^{\infty} C(m) \cdot z^m = \sigma^2 \cdot z \cdot \frac{P(z) - 1}{z - 1} \quad (D.16)$$

$$\sigma^2 = \text{Var}[q_k] = \rho(1-\rho), \text{calculated for a process which changes between the binary states 0 and 1 only}$$

$$P(z) = \frac{A(z) - 1}{A'(1) \cdot (z - 1)} \cdot \frac{B(z) - 1}{B'(1) \cdot (z - 1)} \cdot \frac{[A'(1) + B'(1)A(z)] \cdot (z - 1)}{A(z) \cdot B(z) - 1} \quad (D.17)$$
Finally, if the inverse Fourier transform $S^{-1}(t)$ exists, the ACF $R(\tau)$ (continuous)/$R[m]$ (discrete) of the source can be calculated from the PGFs of the ON and the OFF process. In the case of a discrete function, $R[m]$ becomes:

$$R[m] = \frac{1}{2\cdot\pi} \int_{-\pi}^{\pi} S(e^{i2\pi f}) \cdot e^{im2\pi f} \cdot df$$  \hspace{1cm} (D.18)

For a stationary process $\text{Im}\{S(f)\} = 0$, so that:

$$R[m] = \frac{1}{2\cdot\pi} \int_{-\pi}^{\pi} S(f) \cdot e^{im2\pi f} \cdot df$$  \hspace{1cm} (D.19)

**Example:**

The purpose of this example is to provide some insight how to use the autocorrelation function of a discrete source process to calculate the mean and the variance of the bursts during the burst aggregation process.

In the case here, the ACF for discrete source with geometrically distributed ON and OFF periods is already given [Gar99].

The discrete memoryless system of the ON-OFF source can be described by a two-state Markov chain, where the probabilities for remaining in the ON state (equivalent to generation of ‘1’) is $p$ and the probability for remaining in the OFF state (equivalent to generation of ‘0’) is $q$. The distribution of the ON and OFF lengths is [Bol98]:

$$P(X = l) = p^{k-1} \cdot (1 - p)$$

$$P(Y = l) = q^{k-1} \cdot (1 - q),$$

such that

$$E\{X\} = \frac{l}{1-p}$$

$$E\{Y\} = \frac{l}{1-q}$$

Next we calculate the mean $\mu_z$ and the variance $\sigma_z^2$ of the output process, as $\sigma_z^2$ is required to determine the variation of the burst size. The mean output bit rate of the output process $Z$ is given by:

$$\mu_z = E(Z) = \frac{E\{X\}}{E\{X\} + E\{Y\}} = \frac{1-q}{2-p-q}$$  \hspace{1cm} (D.20)

For a system with two states ‘0’ and ‘1’ only, the variance $\sigma_z^2$ becomes:

$$\sigma_z^2 = \mu_z \cdot (1 - \mu_z) = \frac{(1-p) \cdot (1-q)}{(2-p-q)^2}$$  \hspace{1cm} (D.21)

The discrete autocorrelation function $R[m]$ of the process $Z$ is with $p, q > 0.5$ (generation of long bit sequences, equivalent to packets):
According to equation (3.11), the standard deviation of the burst size \( S \) is then calculated as:

\[
\sigma_s^2 = \int_0^{t_{\text{edge}}} (t_{\text{edge}} - \tau) \cdot \left[ \sigma_z^2 \cdot (p + q - 1)^\tau - \mu_z^2 \right] \cdot d\tau
\]

\[
= -\sigma_z^2 \cdot t_{\text{edge}} \cdot \left[ (p + q - 1)^{2t_{\text{edge}}} \cdot \left( \frac{1}{2} + \frac{1}{\ln(p + q - 1)} \right) + \frac{1}{\ln(p + q - 1)} \right]
\]  

In the limit for \( t_{\text{edge}} \gg 1 \) (large aggregation time), with \( \ln(p + q - 1) < 0 \):

\[
\lim_{t_{\text{edge}} \to \infty} \sigma_s^2 = -\frac{\sigma_z^2}{\ln(p + q - 1)} \cdot t_{\text{edge}}
\]  

As expected for a memoryless (light-tailed) system, the variance of the burst size increases linearly with both \( t_{\text{edge}} \) and the variance \( \sigma_z^2 \) of the generating process.

### D.2 References


Appendix E - Packet loss in queueing systems with many aggregate flows

E.1 Introduction

This appendix provides details on the Likhanov-Mazumdar estimate [Lik98] in addition to the information given in section 3.3 on the impact of long- and short-range dependent (LRD and SRD, respectively) traffic on the queueing performance in OPS networks. It was shown in chapter 3 that long-range dependent and self-similar traffic can be generated by superimposing a large number of individual ON-OFF processes with heavy-tailed ON and OFF periods; in the limit, for a large number of sources with heavy-tailed ON periods, an LRD process is generated. In terms of queueing performance, LRD traffic causes sub-exponential decay of the packet loss and overflow probability with increasing buffer depth. This could only be counteracted by applying significantly more buffer than is required for SRD input processes when trying to obtain the same PLR or overflow probability in both cases. Heavy-tailed processes, for instance the Pareto distribution, are unbounded, i.e. they are allowed to generate either packets or bit rates of any size. In real networks however, packet sizes are bound to a maximum size by various protocols (e.g. 64 KB in IP), and bit rates are practically limited by the maximum bit rate supported on a particular link, e.g. 10 or 40 Gb/s.

In the context of communication network analysis it is, therefore, a valid assumption to restrict the packet sizes and bit rates used in simulations. The resulting estimates for buffer overflow and cell loss rate under the assumption of large multiplexing (N > 100) hold for systems with buffer as well as for bufferless systems, and they are to date the most accurate available in literature. The results are applicable to OPS network design since it was assumed that a queueing system with constant depletion would be used.
E.2 Likhanov/Mazumdar estimate

In the following, we provide a summary of the assumptions required to obtain the estimates for buffer overflow and cell loss ratio [Lik98].

E.2.1 Assumptions

- A single buffer of finite size \(B\) is accessed by i.i.d. stationary and ergodic sources. The depletion rate \(C\) at which the buffer is emptied is constant.

- At any discrete moment in time \(n\), \(\lambda_n\) cells are entering the queue from the sending traffic sources. So \(\lambda_n\) can be thought of as the cell sending rate (cells/unit time), corresponding to the input bit rate.

- \(\lambda_n\) is a stationary process with the mean \(E\{\lambda_n\} < C\) ensuring long-term stability.

- Scaling parameter \(N\): It is assumed that if the number of sources increases by factor \(N\), the respective service rate is \(NC\) and the available buffer space is \(NB\). So \(C\) and \(B\) are the parameters for a queueing system with a single source.

- Cell sending rates \(\lambda_{n,i}\):
  \[n \in \{0, 1, 2, \ldots\}, \ i \in \{1, \ldots, N\}\]
  \[C > E\{\lambda_{n,i}\}\]

- The moment generating function (MGF [Pap91]) \(X_{n,1}\) exists.

- Aggregate cell sending rate \(\lambda_n\):
  The distribution of \(\lambda_n\) is bounded by \(K\), such that \(\lambda_n \in \{0, \ldots, K\}\), \(K\) being the peak rate of the system. \(K > C\) required to ensure occasional buffer overflow.

- Amount of work offered by an individual source \(I\) during time interval \([0, t)\):
  \[X_{t,i} = \sum_{j=1}^{t,i} \lambda_{j,i}\]  
  \(\text{so that the total work offered during that interval is:}\)
  \[X_{t}^{(N)} = \sum_{j=1}^{N} X_{t,i,j}\]

- The MGF of \(X_{t,1}\) is given by:
  \[\phi_t(h) = E\{e^{hX_{t,1}}\}\]

- The following auxiliary functions are needed to express the Bahadur-Bao theorem [Bah60]:
  \[P[X_{t}^{(N)} > N \cdot (C \cdot t + u)] = \frac{e^{-N\varphi_t(Ct,u)}}{\tau_t \cdot \sqrt{2 \cdot \pi \cdot \sigma_t^2 \cdot N}} \left(1 + O\left(\frac{1}{N}\right)\right)\]
where
\[ I_t(C,u) = (C + u) \cdot \tau_t - \ln(\phi(\tau_t)) \]  
(E.5)
\[ \frac{\phi_t'(\tau_t)}{\phi_t(\tau_t)} = C + u \]  
(E.6)
\[ \sigma_t^2 = \frac{\phi_t'(\tau_t)}{\phi_t(\tau_t)} - (C + u)^2 \]  
(E.7)

### E.2.2 Buffer overflow probability

Assume that there exists an unique \( t_0 < \infty \) such that
\[ I_{t_0}(C,B) = \min_{t \in [0,1,2,\ldots]} I_t(C,B) > 0 \]  
(E.8)
\[ \liminf_{t \to \infty} \frac{I_t(C,B)}{\ln t} > 0 \]  
(E.9)

These assumptions are based on the large deviations principle which requires the calculation of \( t_0 \) as the most likely timescale for buffer overflow. The disadvantage of applying the LD principle is based on the difficulty in calculating \( t_0 \) explicitly.

With these assumptions the buffer overflow probability is for \( N \to \infty \):
\[ p_{\{W^{(0)} > N \cdot B\}} = \frac{e^{-N I_{t_0}(C,B)}}{\tau_{t_0} \cdot \sqrt{2 \cdot \pi \cdot \sigma_{t_0}^2} \cdot N^{3/2} \left( 1 + O\left( \frac{1}{N} \right) \right)} \]  
(E.10)

The formula shows that the overflow probability decays exponentially with scaling factor \( N \), since the scaling with \( e^{-N} \) dominates over the term \( N^{-1/2} \).

### E.2.3 Packet loss rate

Under the same assumptions as in 1.2.3, particularly that the critical timescale \( t_0 \) is known, the packet (cell) loss rate denoted by \( P_L \) of the system is given by:
\[ P_L = \frac{e^{-N I_{t_0}(C,B)}}{\tau_{t_0}^2 \cdot C \cdot \rho \cdot \sqrt{2 \cdot \pi \cdot \sigma_{t_0}^2} \cdot N^3} \left( 1 + O\left( \frac{1}{N} \right) \right) \]  
(E.11)
where \( \rho = \frac{E\{\lambda_i\}}{C} \) is the average work fed into the buffer by a single source. The relationship between overflow probability and packet loss rate is as follows:
\[ \frac{p_{\{W^{(0)} > N \cdot B\}}}{P_L} = \tau_{t_0} \cdot C \cdot \rho \cdot N \]  
(E.12)

Although both terms are related to each other, their difference actually grows with scaling factor \( N \) and can exceed two orders of magnitude when \( N > 100 \). It also depends on the critical timescale \( t_0 \) as well as the average work brought to the queue.
In the following eqn. (E.11) is used to determine the packet loss and overflow probability of a bufferless system (B = 0).

### E.2.4 Bufferless system

A large bufferless system (B = 0) resembles a large multiplexing switch without any input buffer, and would provide a worst case result in estimating the PLR to be expected. The critical timescale considered is $t_0 = 1$, since buffer overflow happens instantaneously in each unit of time. By dropping packets that cannot be accommodated during a particular timeslot, the system is free again during the next timeslot to transmit a maximum cell rate of $C$. The PLR in this case becomes:

$$P_L = \frac{c^{-N_{I_1}(C,0)}}{\tau_1^2 \cdot C \cdot \rho \cdot \sqrt{2 \cdot \pi \cdot \sigma^2 \cdot N^3} \left(1 + O\left(\frac{1}{N}\right)\right)}$$

(E.13)

where

$$I_1(C,0) = C \cdot \tau_1 - \ln(\phi(\tau_1))$$

(E.14)

$$\frac{\phi(\tau_1)}{\phi(\tau_1)} = C$$

(E.15)

$$\sigma^2 = \frac{\phi^*(\tau_1)}{\phi(\tau_1)} - C^2$$

(E.16)

### E.2.5 Applicable traffic sources - multiplexing gain

As an LD principle is applied, it is essential that the process $X_t^{(N)}$ for the amount of work entering the queue is related with a unique timescale $t_0$ resulting in buffer overflow. Although the existence of a unique $t_0$ is difficult to prove for all distributions in general, it has been proven to hold for a number of practically relevant processes [Lik98]. These include processes $\lambda_{i,1}$ which are i.i.d., which form a Markov chain. The results still hold for sources with heavy-tailed ON periods and bounded rates, i.e. $\lambda_{i,1} \in [0, K]$, when the scaling factor $N$ is large (usually $N > 100$). But they do not hold when the buffer is fed by a single source only, in which case the tail distribution (i.e. overflow probability for large buffer size) scales subexponentially as $P(W > x) \propto x^{-\alpha}$ with $0 < \alpha < 1$ (corresponds to LRD or self-similar behaviour). The improvement in buffer overflow and cell loss behaviour shows that there is a significant multiplexing gain even in the presence of traffic sources with heavy-tailed (but bounded) ON times, reducing the memory requirements in future network routers and switches.
E.3 References


Appendix F - WR-OBS network parameter variation

F.1 Parameter variation

The deviation of the network performance parameters $t_{WHT}$, $B_{per\lambda}$, $U$ and RUF from a given PDF for the burst size distribution, $f_{t_{burst}}$ (derived from probability theory [Pap91]) was used in chapter 4. For a given random variable (RV) $X$ with PDF $f_X$, the PDF $f_Y$ of a new RV $Y$ with $Y = g(X)$ can be described by the following equation:

$$f_Y(Y) = \frac{f_X(X)}{\frac{dg}{dX}} \quad (F.1)$$

assuming $X = g^{-1}(Y)$ exists and $\frac{dg}{dX} \neq 0$. The PDFs of the network performance parameters defined in chapter 4 are derived assuming that $X = L_{burst}$. Hence, $L_{burst}$ needs to be expressed as a function of the respective performance parameter for calculation of the PDF.

Given that deviations are small compared to the mean, i.e. $\sigma << \mu$, the variance of the performance parameters is derived from the PDF of the burst size distribution. It can be approximated based on the assumption that the variation of the performance parameters can be described by a normal distribution [Her92]. This leads to the following expression for the variance of RV $Y$:

$$\sigma^2_Y = \frac{dg}{dX} \cdot \sigma^2_X \quad (F.2)$$

The 95% confidence level, $\epsilon_{95\%}$, which was used in chapter 4, is related to the standard deviation of a normal distribution by:

$$\epsilon_{95\%} = 1.96 \cdot \sigma \quad (F.3)$$

In general, the burst size and the PDF, $f_{WHT}$, are derived as a function of the wavelength holding time based on equation (F.1), as follows:

$$L_{burst}(t_{WHT}) = (t_{WHT} - t_{RTT}) \cdot b_{core} \quad (F.4)$$

$$f_{WHT} = f_{t_{burst}}[L_{burst}(t_{WHT})] \cdot b_{core} \quad (F.5)$$

The variations of $t_{WHT}$ were indicated in Figure 4.3 by error bars for a 95% confidence level. The PDF of the bandwidth-per-wavelength is:

$$L_{burst}(B_{per\lambda}) = \frac{B_{per\lambda} \cdot t_{RTT}}{1 - \frac{B_{per\lambda}}{b_{core}}} \quad (F.6)$$
In Figure 4.4 the variations of $B_{perX}$ were indicated by error bars for a 95% confidence interval. The results of for the respective PDF of the utilization are as follows:

$$L_{burst}(U) = \frac{U \cdot b_{core} \cdot t_{RTT}}{1 - U}$$  
(F.8)

$$f_U = f_{L_{burst}(U)} \left( \frac{t_{RTT} + \frac{L_{burst}(U)}{b_{core}}}{} \right)^2 \cdot b_{core}$$  
(F.9)

The PDF of the reuse factor is determined as follows:

$$L_{burst}(RUF) = \frac{t_{edge} - RUF \cdot t_{RTT} \cdot b_{core}}{RUF}$$  
(F.10)

$$f_{RUF} = f_{L_{burst}(RUF)} \left( \frac{t_{RTT} + \frac{L_{burst}(RUF)}{b_{core}}}{} \right)^2 \cdot b_{core}$$  
(F.11)

**F.2 References**


Appendix G - ITU-T frequencies for tuneable lasers under test

This appendix provides a catalogue of the ITU-T conform (G.692) frequencies for the most widely used channel spacings of $\Delta f = \{50, 100, 200\ \text{GHz}\}$. The characterization data of the tuneable lasers described in chapter 6 was checked against the ITU-T frequencies to determine whether the devices were able to operate at all frequencies.

The frequencies are defined as multiples of $\Delta f$ (equidistant frequency spacing), referring to a base frequency of 196.0 THz:

$$f_i = (196.0 \pm i \cdot \Delta f)\ \text{THz}, \text{ with } i = 0, 1, 2, \ldots$$  \hspace{1cm} (G.1)

The corresponding wavelength is calculated as:

$$\lambda_i = \frac{c_0}{f_i} = \frac{c_0}{(196.0 \ \text{THz} \pm i \cdot \Delta f)}$$  \hspace{1cm} (G.2)

where $c_0 = 299792458\ \text{m/s}$ (velocity of light in vacuum)

The wavelength spacing $\Delta \lambda_i$ is, therefore, frequency dependent:

$$\Delta \lambda_i = \frac{c_0}{f_i^2} \cdot \Delta f$$  \hspace{1cm} (G.3)

In the case of a frequency spacing of 50 GHz and the chosen frequency range of 190.8...196 THz (C-band), the respective channel spacing is 0.41...0.39 nm, justifying the common approximation of 0.4 nm. The corresponding approximations for 100 GHz and 200 GHz frequency spacing are 0.8 nm and 1.6 nm, respectively.

Below the available wavelength channels are listed for the three tuneable lasers used in the experiments described in chapter 6 (data for the fourth laser used in the wavelength locking experiments, section 6.7, were not available). The first three columns represent the wavelengths for 50, 100, and 200 GHz frequency spacing, whilst the remaining three columns show the wavelength with the highest power that could be selected subject to following criteria:

- SG-DBR #1 and SG-DBR #2: $-3\ \text{dBm} \leq P_{\text{out}} \leq 0\ \text{dBm}$ AND $\text{SMSR} \geq 40\ \text{dB}$
- GCSR: $-6\ \text{dBm} \leq P_{\text{out}} \leq -3\ \text{dBm}$ AND $\text{SMSR} \geq 30\ \text{dB}$
- Wavelengths considered for $\pm 0.1$ nm deviation from the ITU conform wavelength
<table>
<thead>
<tr>
<th>ITU grid (50 GHz) wavelength [nm]</th>
<th>ITU grid (100 GHz)</th>
<th>ITU grid (200 GHz)</th>
<th>SG-DBR #1</th>
<th>SG-DBR #2</th>
<th>GCSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1529.55</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1529.6</td>
<td>0</td>
</tr>
<tr>
<td>1529.94</td>
<td></td>
<td></td>
<td>1530</td>
<td>0</td>
<td>1530</td>
</tr>
<tr>
<td>1530.33</td>
<td>X</td>
<td></td>
<td>1530.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1530.72</td>
<td></td>
<td></td>
<td>1530.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1531.12</td>
<td>X</td>
<td>X</td>
<td>1531.1</td>
<td>0</td>
<td>1531.2</td>
</tr>
<tr>
<td>1531.51</td>
<td></td>
<td></td>
<td>1531.5</td>
<td>0</td>
<td>1531.5</td>
</tr>
<tr>
<td>1531.9</td>
<td>X</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1531.8</td>
</tr>
<tr>
<td>1532.29</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1532.68</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1532.6</td>
<td>0</td>
</tr>
<tr>
<td>1533.07</td>
<td></td>
<td></td>
<td>0</td>
<td>1533</td>
<td>1533</td>
</tr>
<tr>
<td>1533.47</td>
<td>X</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1533.4</td>
</tr>
<tr>
<td>1533.86</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1533.9</td>
</tr>
<tr>
<td>1534.25</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>0</td>
<td>1534.2</td>
</tr>
<tr>
<td>1534.64</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1535.04</td>
<td>X</td>
<td></td>
<td>0</td>
<td>1535.1</td>
<td>1535.1</td>
</tr>
<tr>
<td>1535.43</td>
<td></td>
<td></td>
<td>0</td>
<td>1535.5</td>
<td>1535.4</td>
</tr>
<tr>
<td>1535.82</td>
<td>X</td>
<td>X</td>
<td>1535.9</td>
<td>1535.9</td>
<td>0</td>
</tr>
<tr>
<td>1536.22</td>
<td></td>
<td></td>
<td>1536.3</td>
<td>1536.3</td>
<td>1536.3</td>
</tr>
<tr>
<td>1536.61</td>
<td>X</td>
<td></td>
<td>1536.7</td>
<td>1536.7</td>
<td>1536.6</td>
</tr>
<tr>
<td>1537</td>
<td></td>
<td></td>
<td>1537.1</td>
<td>1537.1</td>
<td>0</td>
</tr>
<tr>
<td>1537.4</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1537.3</td>
<td>1537.4</td>
</tr>
<tr>
<td>1537.79</td>
<td></td>
<td></td>
<td>1537.8</td>
<td>1537.7</td>
<td>1537.8</td>
</tr>
<tr>
<td>1538.19</td>
<td>X</td>
<td></td>
<td>1538.2</td>
<td>1538.1</td>
<td>1538.1</td>
</tr>
<tr>
<td>1538.58</td>
<td></td>
<td></td>
<td>0</td>
<td>1538.5</td>
<td>0</td>
</tr>
<tr>
<td>1538.98</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1538.9</td>
<td>1539</td>
</tr>
<tr>
<td>1539.37</td>
<td></td>
<td></td>
<td>0</td>
<td>1539.3</td>
<td>1539.3</td>
</tr>
<tr>
<td>1539.77</td>
<td>X</td>
<td></td>
<td>0</td>
<td>1539.7</td>
<td>1539.7</td>
</tr>
<tr>
<td>1540.16</td>
<td></td>
<td></td>
<td>0</td>
<td>1540.1</td>
<td>1540.1</td>
</tr>
<tr>
<td>1540.56</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1540.5</td>
<td>1540.6</td>
</tr>
<tr>
<td>1540.95</td>
<td></td>
<td></td>
<td>0</td>
<td>1540.9</td>
<td>1540.9</td>
</tr>
<tr>
<td>1541.35</td>
<td>X</td>
<td></td>
<td>0</td>
<td>1541.4</td>
<td>1541.4</td>
</tr>
<tr>
<td>1541.75</td>
<td></td>
<td></td>
<td>1541.7</td>
<td>1541.7</td>
<td>1541.7</td>
</tr>
<tr>
<td>1542.14</td>
<td>X</td>
<td>X</td>
<td>1542.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1542.54</td>
<td></td>
<td></td>
<td>1542.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1542.94</td>
<td>X</td>
<td></td>
<td>1542.9</td>
<td>0</td>
<td>1543</td>
</tr>
<tr>
<td>1543.33</td>
<td></td>
<td></td>
<td>1543.3</td>
<td>0</td>
<td>1543.3</td>
</tr>
<tr>
<td>1543.73</td>
<td>X</td>
<td>X</td>
<td>1543.7</td>
<td>1543.7</td>
<td>0</td>
</tr>
<tr>
<td>1544.13</td>
<td></td>
<td></td>
<td>1544.2</td>
<td>1544.1</td>
<td>1544.1</td>
</tr>
<tr>
<td>1544.53</td>
<td>X</td>
<td></td>
<td>1544.6</td>
<td>1544.5</td>
<td>1544.5</td>
</tr>
<tr>
<td>1544.92</td>
<td></td>
<td></td>
<td>0</td>
<td>1544.9</td>
<td>1544.9</td>
</tr>
<tr>
<td>1545.32</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1545.3</td>
<td>1545.3</td>
</tr>
<tr>
<td>1545.72</td>
<td></td>
<td></td>
<td>0</td>
<td>1545.7</td>
<td>1545.7</td>
</tr>
<tr>
<td>1546.12</td>
<td>X</td>
<td></td>
<td>0</td>
<td>1546.1</td>
<td>0</td>
</tr>
<tr>
<td>1546.52</td>
<td></td>
<td></td>
<td>1546.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1546.92</td>
<td>X</td>
<td>X</td>
<td>1546.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1547.32</td>
<td></td>
<td></td>
<td>1547.3</td>
<td>0</td>
<td>1547.3</td>
</tr>
<tr>
<td>1547.72</td>
<td>X</td>
<td></td>
<td>1547.7</td>
<td>1547.8</td>
<td>0</td>
</tr>
<tr>
<td>1548.11</td>
<td></td>
<td></td>
<td>1548.1</td>
<td>1548.1</td>
<td>1548.1</td>
</tr>
<tr>
<td>1548.51</td>
<td>X</td>
<td>X</td>
<td>1548.5</td>
<td>1548.5</td>
<td>0</td>
</tr>
<tr>
<td>1548.91</td>
<td></td>
<td></td>
<td>1548.9</td>
<td>1548.9</td>
<td>1549</td>
</tr>
<tr>
<td>1549.32</td>
<td>X</td>
<td></td>
<td>1549.3</td>
<td>1549.3</td>
<td>1549.4</td>
</tr>
<tr>
<td>1549.72</td>
<td></td>
<td></td>
<td>1549.7</td>
<td>1549.7</td>
<td>1549.7</td>
</tr>
<tr>
<td>1550.12</td>
<td>X</td>
<td>X</td>
<td>1550.1</td>
<td>1550.1</td>
<td>0</td>
</tr>
<tr>
<td>1550.52</td>
<td></td>
<td></td>
<td>1550.5</td>
<td>1550.5</td>
<td>1550.6</td>
</tr>
<tr>
<td>1550.92</td>
<td>X</td>
<td></td>
<td>1550.9</td>
<td>1550.9</td>
<td>1550.9</td>
</tr>
</tbody>
</table>
ITU-T frequencies for tuneable lasers under test

<table>
<thead>
<tr>
<th>ITU grid (50 GHz) wavelength [nm]</th>
<th>ITU grid (100 GHz)</th>
<th>ITU grid (200 GHz)</th>
<th>SG-DBR #1</th>
<th>SG-DBR #2</th>
<th>GCSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1551.32</td>
<td>X</td>
<td>X</td>
<td>1551.3</td>
<td>1551.3</td>
<td>0</td>
</tr>
<tr>
<td>1551.72</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1551.7</td>
<td>0</td>
</tr>
<tr>
<td>1552.12</td>
<td>0</td>
<td>0</td>
<td>1552.1</td>
<td>1552.2</td>
<td></td>
</tr>
<tr>
<td>1552.52</td>
<td>X</td>
<td>0</td>
<td>1552.5</td>
<td>1552.5</td>
<td></td>
</tr>
<tr>
<td>1552.93</td>
<td></td>
<td></td>
<td>1552.9</td>
<td>1552.9</td>
<td>1553</td>
</tr>
<tr>
<td>1553.33</td>
<td>X</td>
<td>X</td>
<td>1553.3</td>
<td>1553.3</td>
<td>1553.4</td>
</tr>
<tr>
<td>1553.73</td>
<td>X</td>
<td>0</td>
<td>1553.7</td>
<td>1553.7</td>
<td></td>
</tr>
<tr>
<td>1554.13</td>
<td>X</td>
<td></td>
<td>1554.1</td>
<td>1554.2</td>
<td>0</td>
</tr>
<tr>
<td>1554.54</td>
<td></td>
<td></td>
<td>1554.5</td>
<td>1554.5</td>
<td>1554.6</td>
</tr>
<tr>
<td>1554.94</td>
<td>X</td>
<td>X</td>
<td>1554.9</td>
<td>1554.9</td>
<td>1554.9</td>
</tr>
<tr>
<td>1555.34</td>
<td>X</td>
<td></td>
<td>1555.3</td>
<td>1555.3</td>
<td>1555.3</td>
</tr>
<tr>
<td>1555.75</td>
<td>X</td>
<td></td>
<td>1555.8</td>
<td>1555.8</td>
<td>0</td>
</tr>
<tr>
<td>1556.15</td>
<td></td>
<td></td>
<td>1556.2</td>
<td>1556.1</td>
<td>1556.2</td>
</tr>
<tr>
<td>1556.55</td>
<td>X</td>
<td>X</td>
<td>1556.5</td>
<td>1556.5</td>
<td>1556.6</td>
</tr>
<tr>
<td>1556.96</td>
<td></td>
<td></td>
<td>1556.9</td>
<td>1556.9</td>
<td>1557</td>
</tr>
<tr>
<td>1557.36</td>
<td>X</td>
<td></td>
<td>1557.3</td>
<td>1557.3</td>
<td>1557.4</td>
</tr>
<tr>
<td>1557.77</td>
<td>X</td>
<td></td>
<td>1557.8</td>
<td>1557.7</td>
<td>1557.7</td>
</tr>
<tr>
<td>1558.17</td>
<td>X</td>
<td>X</td>
<td>1558.2</td>
<td>1558.1</td>
<td>0</td>
</tr>
<tr>
<td>1558.58</td>
<td>X</td>
<td></td>
<td>1558.5</td>
<td>1558.6</td>
<td>1558.6</td>
</tr>
<tr>
<td>1558.98</td>
<td>X</td>
<td></td>
<td>1558.9</td>
<td>1559</td>
<td>1558.9</td>
</tr>
<tr>
<td>1559.39</td>
<td></td>
<td></td>
<td>1559.3</td>
<td>1559.4</td>
<td>1559.3</td>
</tr>
<tr>
<td>1559.79</td>
<td>X</td>
<td>X</td>
<td>1559.7</td>
<td>1559.8</td>
<td>0</td>
</tr>
<tr>
<td>1560.2</td>
<td></td>
<td></td>
<td>1560.2</td>
<td>1560.3</td>
<td>0</td>
</tr>
<tr>
<td>1560.61</td>
<td>X</td>
<td></td>
<td>1560.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1561.01</td>
<td></td>
<td></td>
<td>1561</td>
<td>1561.1</td>
<td>0</td>
</tr>
<tr>
<td>1561.42</td>
<td>X</td>
<td>X</td>
<td>1561.4</td>
<td>1561.4</td>
<td>1561.5</td>
</tr>
<tr>
<td>1561.83</td>
<td></td>
<td></td>
<td>1561.8</td>
<td>1561.9</td>
<td>0</td>
</tr>
<tr>
<td>1562.23</td>
<td>X</td>
<td></td>
<td>1562.3</td>
<td>0</td>
<td>1562.3</td>
</tr>
<tr>
<td>1562.64</td>
<td></td>
<td></td>
<td>1562.7</td>
<td>0</td>
<td>1562.7</td>
</tr>
<tr>
<td>1563.05</td>
<td>X</td>
<td>X</td>
<td>1563</td>
<td>0</td>
<td>1563</td>
</tr>
<tr>
<td>1563.45</td>
<td>X</td>
<td></td>
<td>1563.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1563.86</td>
<td>X</td>
<td></td>
<td>1563.8</td>
<td>1563.9</td>
<td>1563.9</td>
</tr>
<tr>
<td>1564.27</td>
<td></td>
<td></td>
<td>1564.3</td>
<td>1564.3</td>
<td>0</td>
</tr>
<tr>
<td>1564.68</td>
<td>X</td>
<td>X</td>
<td>0</td>
<td>1564.7</td>
<td>1564.6</td>
</tr>
<tr>
<td>1565.09</td>
<td></td>
<td></td>
<td>0</td>
<td>1565.1</td>
<td>1565</td>
</tr>
<tr>
<td>1565.5</td>
<td>X</td>
<td></td>
<td>1565.4</td>
<td>1565.5</td>
<td>1565.5</td>
</tr>
<tr>
<td>1565.9</td>
<td></td>
<td></td>
<td>1566</td>
<td>1566</td>
<td>0</td>
</tr>
<tr>
<td>1566.31</td>
<td>X</td>
<td>X</td>
<td>1566.4</td>
<td>1566.4</td>
<td>0</td>
</tr>
<tr>
<td>1566.72</td>
<td></td>
<td></td>
<td>1566.7</td>
<td>1566.8</td>
<td>0</td>
</tr>
<tr>
<td>1567.13</td>
<td>X</td>
<td></td>
<td>1567.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1567.54</td>
<td></td>
<td></td>
<td>1567.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1567.95</td>
<td>X</td>
<td>X</td>
<td>1568</td>
<td>1568</td>
<td>0</td>
</tr>
<tr>
<td>1568.36</td>
<td></td>
<td></td>
<td>1568.4</td>
<td>0</td>
<td>1568.3</td>
</tr>
<tr>
<td>1568.77</td>
<td>X</td>
<td></td>
<td>1568.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1569.18</td>
<td></td>
<td></td>
<td>1569.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1569.59</td>
<td>X</td>
<td>X</td>
<td>1569.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1570.01</td>
<td>X</td>
<td></td>
<td>1570</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1570.42</td>
<td>X</td>
<td></td>
<td>1570.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1570.83</td>
<td></td>
<td></td>
<td>1570.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1571.24</td>
<td>X</td>
<td>X</td>
<td>1571.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1571.65</td>
<td></td>
<td></td>
<td>1571.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1572.06</td>
<td>X</td>
<td></td>
<td>1572.1</td>
<td>1572</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix H - Advanced multimode fibres for high bit rate transmission in LANs and MANs

H.1 Introduction

This appendix deals with the experimental investigation carried out as part of this work, and results of high bit rate transmission over multimode fibre (MMF) as required for emerging standards such as 10 Gigabit Ethernet (10GbE). The work described in this appendix is most relevant to the design of high-speed metro and access networks, but does not form a central part of the analysis of dynamic optical network architectures.

The optical network architectures presented in this work have so far been solely focused on applications in wide area networks, with individual wavelengths operating at bit rates of 10 Gb/s and beyond. Although optical transmission systems providing in excess of 1 Tb/s capacity are already available [Big01], currently the bottleneck preventing utilization of this bandwidth is at the access layer [Ian02]. For private users, network access is nearly exclusively through copper wiring, ranging from standard twisted-pair to cable modems offering bit rates up to a few Mb/s per user (asynchronous digital subscriber loop, ADSL) [Reu01]. To sustain further network growth in wide-area networks, as discussed in chapter 1, it is essential that the local area networks (LAN) and metropolitan area networks (MAN [Gha02]) extend the access to high bandwidth to the user – leading to the concept of fiber-to-the home (FTTH [Ket00]) when optical fiber is used as the transmission medium.

Growth in network capacity has also been driven by the enterprise market, with a large number of computers interconnected to LANs. The dominating standard in this area to date has been Ethernet: following the surge in capacity, the current standards for transmission are set by Gigabit Ethernet (GbE, IEEE 802.3z) [Tha02, Lam02] and 10-Gigabit-Ethernet (10GbE, IEEE 802.3ae) [Cun02]. Due to the bandwidth-distance product of copper wiring, limited to distances < 80 m for GbE, the transmission was predominantly defined to be optical – particularly for 10GbE as shown in Table 1.1 below.
The optical medium to be considered for transmission is graded-index (GI) multimode fibre (MMF) due its large core diameter (50 or 62.5 μm compared to 9 μm for SMF). With GI-MMF installed until the 1990s showing bandwidth-distance products limited to 160...200 MHz-km, there was a real need to extend the achievable bandwidth to the user, particularly for 10GbE. As shown here in record transmission experiments, GI-MMF with advanced index profiles holds the key for future high bit rate transmission in LANs, potentially extending to campus size (< 2 km) and MAN type distances (typically 10...100 km, [Reu01]).

MMF have found an application in local area network applications due to their large core diameters, enabling simplified and therefore inexpensive connectivity in comparison to standard single mode fibre (SSMF). Current specifications of 10GbE define transmission over physical media for both MMF and SMF as defined in IEEE 802.3ae standard. Therein, the following interfaces have been defined, split into applications for local area networks (LAN) and wide area networks (WAN). The aim of the IEEE 802.3ae task force was to create two different interfaces: A LAN interface operating at 10 Gb/s for short-reach connectivity, e.g. between routers, within buildings and across campuses. The WAN interface was specifically designed to provide an interface to SDH/SONET, and accordingly operates with a payload compatible to OC-192/VC-4-64c. The physical interfaces (PHY) defined within IEEE 802.3ae are listed in Table H.1.

Short distances within this standard are covered by the 10G Base-LX4 and the 10G Base-SR physical interfaces, with maximum reach of 65 m for the serial interface, using a specifically designed restricted offset launch using vertical cavity emitting lasers (VCSEL).

<table>
<thead>
<tr>
<th>Type</th>
<th>Wavelength</th>
<th>SMF</th>
<th>MMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10G Base-LX4</td>
<td>1300 nm 4λWDM</td>
<td>10 km</td>
<td>65*/300 m*</td>
</tr>
<tr>
<td>10G Base-SR</td>
<td>850 nm serial</td>
<td>---</td>
<td>65/300 m</td>
</tr>
<tr>
<td>10G Base-LR</td>
<td>1310 nm serial</td>
<td>10 km</td>
<td>---</td>
</tr>
<tr>
<td>10G Base-ER</td>
<td>1550 nm serial</td>
<td>40 km</td>
<td>---</td>
</tr>
<tr>
<td>10G Base-SW</td>
<td>850 nm serial</td>
<td>---</td>
<td>65/300 m</td>
</tr>
<tr>
<td>10G Base-LW</td>
<td>1310 nm serial</td>
<td>10 km</td>
<td>---</td>
</tr>
<tr>
<td>10G Base-EW</td>
<td>1550 nm serial</td>
<td>40 km</td>
<td>---</td>
</tr>
</tbody>
</table>

Table H.1: Physical interfaces defined for use with 10 Gigabit Ethernet (10GbE) as defined by the standard IEEE 802.3ae

* Over installed 65/125 μm MMF
* Over new 50/125 μm MMF
Two sets of experiments were carried out in the work described here, to demonstrate the high bit-rate-distance product of these fibres over 8 spools of MMF, 4.4 km and 1.1 km in length each. First, BER measurements were carried out using a restricted centre launch. Secondly, a Packet-over-SONET (POS) transmission link was used operating at OC-192 line rate, also with a restricted centre launch.

H.2 Related work

For more than a decade from the mid-1970s to mid-1980s, multimode fibres (MMF) dominated commercial optical transmission systems operated at 850 nm in the first optical window [Mid00]. The advent of low-loss single-mode fibre (SMF) in combination with advanced DBR and DFB laser sources in the 1980s led to optical transmission systems of the second generation which were entirely based around the new SMF fibre [Kog00]. The benefits of MMF were re-discovered in the mid-1990s when the booming internet resulted in the interconnection of PCs on the office and campus level into local area networks (LAN). Although a number of protocols prevailed, such as ATM, FDDI, the dominant standard due to its robustness and relative simplicity, turned out to be Ethernet. Initially constructed around twisted-pair wiring, the move to Gigabit Ethernet meant that copper-based transmission media were insufficient to support the envisaged physical bit rate of 1.25 Gb/s over distances up to 300 m (IEEE 802.3z). Bandwidth-distance products of originally 100...200 MHz·km proved insufficient, the main limitation being the modal dispersion under overfilled launch (OFL) [Yab00/1]. The OFL refers to a launch condition as it is provided by an LED where all mode groups (angular as well as radial) in the MMF are excited.

Limiting the number of excited modes was the solution to this problem, known already from a measurement technique called differential mode delay (DMD) used to characterise the bandwidth of MMF by exciting selected modes only [Mar81]. The first step into this direction was [Haa93] using a centered single-mode launch and reception fibre in combination with a 62.5/125 μm MMF, enabling 1 Gb/s over 2.3 km of MMF at 1300 nm. As many installed MMFs suffer from an index dip along the fibre axis, resulting from the MCVD production process, restricted offset launches were proposed and successfully tested [Rad98, Web99]. Using SMF for excitation, 2.5 Gb/s transmission over 3 km of 50/125 μm MMF at 1.3 nm was achieved using PRBS $2^7$-1

---

7 MMF in this chapter always refers to graded-index (GI) fibre
Advanced multimode fibres for high bit rate transmission in LANs and MANs

[Rad98]. 10 Gb/s transmission at 850 nm using a VCSEL launch was demonstrated over 1.6-km of 50/125 μm MMF [Mic99]. Numerical modelling showed, however, that the highest bandwidth-distance products were achieved with a centre launch [Yab00/2, Rad97]. Improved MMF manufacturing process resulted in fibre with very low profile distortion, allowing to transmit 2.5 Gb/s over 4.4 km of 62.5/125 μm MMF at 1.3 μm [Due00]. Details of this experiment are reported in this appendix, along with an experimental packet-over-SONET (POS) showing that even higher bit-rates are possible.

High bandwidth transmission was achieved using the passband of the MMF for transmission by applying subcarrier multiplexing (SCM) [Tyl02]: [Lee99] reported 12 Gb/s over 4 km of 62.5/125 μm MMF using a 256 quadrature amplitude modulation (QAM) in combination with SCM and WDM, using 3 separate wavelength channels. A more recent demonstration showed the feasibility of 1 Tb/s on 50/125 μm MMF over 3 km, using a combination of WDM and subcarrier multiplexing (SCM) [Roc02]. All SCM techniques, however, require further microwave components at the transmitter and receiver end of the transmission fibre to achieve these capacities.

A similar technique to the one used in the set-up for the OC-192 Packet-over-SONET transmission here, called mode filtering, was used in [Haa93] to improve the bandwidth-distance product of MMF by reducing the number of modes for transmission. Experimental results show that for a centre launch technique using a SMF, the optimum eye opening is achieved when another SMF is used at the MMF output. The main effect, however, is the launch into the fundamental mode of the MMF, not the mode filtering at the output. When a MMF patchcord is used at the output instead of the SMF patchcord, the eye opening diminishes only slightly. Experimental and theoretical results in [Web98] confirm that the maximum bandwidth-distance product in 62.5/125 μm MMF is achieved with a centre launch. A more recent experimental study of the penalties introduced by connector offsets [Weg01] leads to conclusion that in the case of a centre launch, light is launched into the fundamental mode only. The resulting $1/e^2$ beam diameter† is < 16 μm in graded-index fibre with 40-μm core diameter. High-speed transmission experiments for 10 Gb/s at 850 nm were carried out using 50/125 μm MMF [Mic99], with 3 dB receiver penalty after 1.6-km. These results were confirmed on 1.5-km of 62.5/125 μm MMF [Zha02]: 10 Gb/s transmission was achieved, but only

† The beam diameter is defined such that $1/e^2 = 86.5\%$ of the beam energy is contained in the corresponding area
with using an adaptive electronic equalization technique which cancelled the dispersive effects at the receiving end of the transmission system.

Both the Gigabit Ethernet (IEEE 802.3z) and 10 Gigabit Ethernet standard (IEEE 802.3ae) now adopt a serial transmitter at 850 nm with a qualified launch into 50 and 62.5/125 µm MMF, using a vertical cavity surface-emitting laser (VCSEL). This was due to the cost-effectiveness of this scheme at the time of standardization.

**H.3 2.5 Gb/s transmission over 4.5 km of 62.5 µm multimode fibre using a centre launch technique**

The bandwidth of graded-index multimode fibre (MMF) is a limiting factor in achieving multi-gigabit transmission in local area networks (LANs). Typical overfilled launch (OFL) bandwidth-distance products for the 62.5/125 µm fibres are 200 MHz-km at 850 nm and 500 MHz-km at 1300 nm [Web99]. Although much research effort has focussed on vertical cavity surface-emitting lasers (VCSEL) at 850 nm to increase bandwidths [Aro98, Woo99] this wavelength is not suitable for distances longer than 1 km because of high attenuation (3 dB/km) and chromatic dispersion (80 ps-nm⁻¹-km⁻¹). The bandwidth-distance product could be extended into the GHz-km range at 1300 nm wavelength in combination with an offset launch technique [Web99, Rad98]. This technique excites only a small subset of modes and, to date, has been demonstrated to achieve bit-rate-distance products at 1300 nm of 2.4 Gbit/s-km for the standard 62.5 µm core fibre [Rad98] and 7.5 Gbit/s-km over smaller, 50 µm core fibre [Rad98]. However, the offset launch requires special connectors, and the simpler centre launch would allow standard connectors to be used.

In this appendix, low-penalty NRZ transmission at 2.5 Gbit/s and 1300 nm over 62.5 µm MMF using a centre launch technique with a bit-rate-distance product greater than 11 Gbit/s-km are described [Due00]. At the time of publication, this was the highest bit-rate-distance product achieved on 62.5/125 µm MMF with a centre-launch and using NRZ coding, not using SCM [Roc02] or WDM techniques [Aro98]. The centre launch technique used in that experiment [Due00] was found to be well within the tolerances for standard multimode connectors [Rad98].

**H.3.1 Multimode fibre theory**

The first generation of optical transmission systems was operated at 850 nm and used GI-MMF as a transmission medium because of the larger fibre diameter. GI-MMF were the medium of choice because with diameters of 100 and 125 µm, respectively, they
were easier to connecterize than singlemode fibres with a core diameter of 9 μm. Secondly, operating with light-emitting diodes (LED) rather than lasers, the large numerical aperture (NA) compared to SMF allowed a higher coupling efficiency [Mar81].

The bandwidth of early MMF was limited to a few tens of MHz-km [Mar81], so that the dominant limiting effects are either attenuation or dispersion, nonlinear effects can be neglected. The attenuation in MMF is higher than in standard single mode fibre due to the higher dopant concentration in the core section required to obtain a high NA, and typical values found are 0.5...1.0 dB/km at 1300 nm and 2.5...3.0 dB/km at 850 nm, depending on the dopant level in the core.

The dispersion effects in MMF are dominated by the difference in the group velocities of different mode groups, resulting in modal dispersion. This is particularly strong in step-index (SI) MMF, but can be minimized (though usually not completely eliminated) by a particular class of fibres with a graded-index (GI) profile, also called power-law profiles defined as:

\[
n^2(r) = \begin{cases} 
  n_2^2 \left( 1 - 2 \cdot \Delta \cdot \left( \frac{r}{a} \right) \right) & r \leq a \quad \text{(core)} \\
  n_2^2 \left( 1 - 2 \cdot \Delta \right) & r > a \quad \text{(cladding)}
\end{cases}
\]

where \( n_1 \) is the maximum refractive index in the centre of the core, \( n_2 \) is the refractive index of the cladding, \( r \) is the radius, \( a = 31.25 \) μm is the outer radius of the core, and \( \Delta \) is the refractive index difference, with \( \alpha \) typically in the range of 2 for maximum bandwidth-distance product, i.e. is typically \( 1.5 < \alpha < 2.5 \).

Additional dispersive effects are the chromatic dispersion with its components of waveguide and material dispersion (negligible for 1300 nm), and the profile distortion. The contributions of these effects are negligible for bit rates of a few hundred Mbit/s, but affect the achievable bandwidth when higher bandwidths are required. Figure H.1 shows the achievable bandwidth-distance product for an overfilled launch (OFL), derived from an analytical model [Mar81, Ohl76], as a function of the profile parameter \( \alpha \) for 850 (H.1-a) and 1300 nm (H.1-b). The bandwidth-distance product depends strongly on \( \alpha \), and the optimum value of the profile parameter is wavelength dependent, with 1.976 for 850 nm and 1.873 for 1300 nm. The maximum bandwidth-distance product decreases with increasing linewidth due to material dispersion, exemplified by three different sources:
Advanced multimode fibres for high bit rate transmission in LANs and MANs

A DFB laser with an rms linewidth < 0.1 nm, a VCSEL with 1 nm, and an LED with 15 nm. The graph clearly shows that for an operating wavelength of 850 nm, the bandwidth-distance product is generally lower than for 1300 nm when a VCSEL or LED are used. A number of conclusions can be drawn from Figure H.1-b with respect to the experimental results achieved:

- In the case of an OFL, it is paramount to use MMF with a profile parameter close to the optimum $\alpha$ to counteract modal dispersion. In the case of a restricted launch, though, only a small number or modes (or even only one) are excited, allowing a higher tolerance towards an offset from $\alpha$, but less distortion of the profile, particularly in the centre of the fibre. This is supported by the results presented below, where the MMF with the highest bandwidth-distance product was shown to have $\alpha = 1.99$ (and therefore offset from the optimum shown in Figure H.1-b).

- The maximum bandwidth is limited by the linewidth of the optical source. Although the bandwidth increases as the linewidth increases, for 1300 nm the maximum values achieved with a DFB and a VCSEL are approximately the same. The implications are twofold: As in the experiments reported here a Fabry-Perot (FP) laser diode was used with a linewidth smaller than that of a VCSEL, its impact on the bandwidth-distance product is assumed to be negligible. In fact, the linewidth of the source should be as large as possible (i.e. without affecting the bandwidth-distance product) to reduce the impact of modal noise which increases with the coherence of the source.

- Under the OFL condition the maximum bandwidth-distance product is limited to 3.3 GHz-km even for lasers with small linewidth, hardly sufficient to support a bit rate of 10 Gb/s over 1 km. The only way of exceeding such limitations is by using a restricted launch technique as detailed below.
H.3.2 Advanced multimode fibres with improved graded-index profile

The bandwidth-distance product of GI-MMF is primarily limited by modal dispersion, i.e. the different velocities of modes within the optical fibre. It can be shown that the effect of modal dispersion can theoretically be completely cancelled by using power-law profiles [Fre97]. In practice, however, due to manufacturing tolerances the optimum power law is not always obtainable, resulting in residual modal dispersion. Figure H.2 shows the typical profile of one of the multimode fibres used in the experiment, measured by a refractive-index technique. The profile shows a nearly perfect power-law profile, and the absence of an index depression typical of many MMF should be noted. This is due to an improved fibre manufacturing technique: Previously, MMF was produced using the modified chemical vapour deposition (MCVD) technique [Mac00], often resulting in an index dip or 'burn out' in the centre of the fibre when the perform was collapsed prior to drawing. This problem was avoided using plasma-enhanced chemical vapour deposition (PCVD) for the fibres used in the experiment, allowing to deposit thinner layers and control the refractive index in the core of the fibre.

Figure H.3 shows the profiles of four fibres, amongst others the same as in Figure H.2, showing that all fibres have a profile following closely a power-law profile, potentially enabling bandwidth-distance products in excess of 1 GHz-km. The profiles shown here were taken from one end of the fibre and may not represent profile variations occurring

![Figure H.2: Typical refractive index profile for GI-MMF used in the experiment, data taken by a refractive-index measurement technique](image-url)

---

8 courtesy of W.A. Reed, OFSLabs, NJ, USA
Advanced multimode fibres for high bit rate transmission in LANs and MANs

Figure H.3: Profiles of four selected GI-MMF used in the experiments. Distortions on fibre 6, right slope, are assumed to be local along the 4.5 km length of the fibre. For instance, the distortions observed in the right slope of fibre #6 are assumed to be local and not characteristic for the entire fibre spool. The profiles were fitted to an $\alpha$-profile as described in eqn. (7.1), resulting in values for $\alpha$ and the profile distortion $\chi^2$ (variance of measured profile from the fit) as listed in Table H.2.

<table>
<thead>
<tr>
<th>Fibre no.</th>
<th>Profile parameter $\alpha$</th>
<th>Profile distortion $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.9037</td>
<td>2.55·10^{-8}</td>
</tr>
<tr>
<td>2</td>
<td>1.9656</td>
<td>1.88·10^{-8}</td>
</tr>
<tr>
<td>3</td>
<td>2.0125</td>
<td>1.17·10^{-8}</td>
</tr>
<tr>
<td>4</td>
<td>1.9297</td>
<td>4.47·10^{-8}</td>
</tr>
<tr>
<td>5</td>
<td>1.8453</td>
<td>1.98·10^{-8}</td>
</tr>
<tr>
<td>6</td>
<td>1.7414</td>
<td>2.68·10^{-8}</td>
</tr>
<tr>
<td>7</td>
<td>1.9943</td>
<td>2.77·10^{-9}</td>
</tr>
<tr>
<td>8</td>
<td>1.9200</td>
<td>8.34·10^{-9}</td>
</tr>
</tbody>
</table>

Table H.2: Profile parameter $\alpha$ and distortion $\chi^2$ measured for all eight GI-MMF under test

The profile parameters vary from 1.74-2.01, with a mean of 1.9141 and a standard deviation of 0.0876. Both profile parameter and profile distortion were determined for an area of ±10 $\mu$m around the centre of the fibre since under the restricted centre launch condition, light is confined to the fundamental mode in this area of the fibre, unlike in the case of an overfilled launch, where modes across the whole diameter of the fibre are excited. This result is supported by examination of GI-MMF using a scanning...
technology which revealed that the power is mostly confined to the centre of the fibre [Whi98]. The implications of this profile distortion with respect to the transmission penalties involved are investigated in more detail in section 7.5.

**H.3.3 Experimental set-up**

In the experimental set-up (Figure H.4) a commercial 1.3 μm transceiver (HP HFCT-5402D) was used. Optimised for OC-48/STM-16 at 2.488 Gbit/s, it uses a directly-modulated Fabry-Perot (FP) laser source and an InGaAs PIN photodiode with integral data recovery circuit. The transmitter output power is −5.7 dBm, the receiver back-to-back sensitivity is −23.6 dBm for a bit error rate (BER) of 10^{-10}, with an overall link power budget of 17.9 dB. Light was launched from a singlemode fibre patchcord into the MMF, the launch optimized using a 3D-translation stage. After transmission through one spool of MMF the light is coupled into the attenuator using a bare fibre adaptor introducing additional penalty due to modal noise. To show the reproducibility of the results, the experiment was carried out using eight different 62.5 μm core MMFs, each approximately 4.5 km long and drawn from a different preform. The fibres (Yangtze Optical Fiber Corporation, China) were drawn from Plasma-activated Chemical Vapour Deposition (PCVD) manufactured preforms with negligible centre dips (the lowest measured was Δn = 8·10^{-4}) in the refractive index profile. The overfilled launch (OFL) bandwidth-distance products of all fibres exceed the IEEE 802.3z requirements with values of 920 – 1463 MHz·km at 1300 nm, while the attenuation of all fibres was measured to be in the range 0.49 – 0.51 dB/km. The fibre lengths were verified using an optical time domain reflectometry (OTDR) with a resolution of 6 metres. The FP was modulated with a 2^{23}-1 PRBS (longest obtainable), and the bit error rate (BER) against optical power was measured using an Anritsu MP1764A bit error rate test set, and the sensitivity to offset was measured for a constant BER of 10^{-9}. 

H.3.4 Results and discussion

The measured BER is shown in Figure H.5 for four out of the eight fibres (for reduced complexity). Transmission at 2.5 Gbit/s over 4.5 km was possible for all fibres with low penalties. The lowest penalty was measured to be 1.1 dB at a BER of $10^{-10}$ increasing to a maximum of 4.5 dB, with 7 out of 8 fibres exhibiting penalties of less than 2.8 dB.

Figure H.5: Bit error rate (BER) vs. received optical power for back-to-back (btb) (filled square), and for four fibres showing the best case (open square), the worst case (open circle), and two typical fibres (diamond, triangle)
The parameter critical for the deployment of MMF is the tolerance of the restricted launch to offsets from the optimum launch position. Offsets result in the excitation of additional mode groups in the MMF, reducing the bandwidth. We measured the penalty (Figure H.6) to maintain a BER of $10^{-9}$ for offsets from the centre launch position in steps of 0.5 µm using a high-precision micro-positioner. For a 3-dB penalty, the largest offset was found to be ± 7 µm from the centre of the fibre. Typical values were in the range of ± 5 µm, compatible with tolerances of most mechanical connector designs. In the worst case, the offset tolerance is only ± 2 µm. This low value was observed in fibres with an index depression in the central core region and the largest profile distortions.

![Figure H.6: Transmission penalty (for 10\(^{-9}\) BER) as a function of the input launch position (for same fibres as in Figure H.5)](image)

### H.3.5 Profile distortions

To characterise the relationship between transmission penalty and refractive index profile distortions, we define a distortion parameter $\chi^2$ as the variance of all refractive index data points from the $\alpha$-profile for the best fit to the measured refractive index profile (a typical refractive index profile was shown in Figure H.2). The variation in penalty and tolerance versus offset depends mainly on the uniformity of the refractive index profile. Figure H.7 shows the results for the measured transmission penalties and $\chi^2$. It can be seen that the fibre (fibre 4) with the largest distortion $\chi^2 = 4.47 \cdot 10^{-8}$ had the highest transmission penalty (4.5 dB) and the least tolerance to offset, whereas a nearly ideal profile with a distortion of $2.27 \cdot 10^{-9}$ resulted in only a 1.1 dB penalty and had the highest tolerance to offset (± 8 µm).

The significance of this is in confirming that the optimum $\alpha$ designed for OFL does not fully determine the transmission characteristics in the centre launch case, where only a subset of modes is excited, and with significantly reduced modal dispersion it is mainly the profile distortions which determine the transmission penalties. Hence, fibres
intended for use with centre launch should aim at a low differential mode delay to minimise the residual modal dispersion and maximise the tolerance to offset. This has not been investigated or reported in literature previously. Indeed, calculations based on the Wentzel, Kramers, Brillouin (WKB) method [Mar81] show that a bit-rate-distance product of 10 Gbit/s·km cannot be achieved in 62.5 µm fibres using a conventional OFL technique.

*Figure H.7: Transmission penalty at BER $10^{-10}$ for centre launch (○) and the variance of the profile distortions $\chi^2$ (■)*
H.4 Packet-over-SONET (POS) transmission experiment at OC-192 line rate using GI-MMF

H.4.1 Introduction

As a further test of the transmission performance of the GI-MMFs used here, transmission experiments were carried out for the POS transmission at OC-192 line rate, i.e. 9.95328 Gb/s (approx. 10 Gb/s). State-of-the-art test equipment for electronic routers was used for this purpose, which allowed TCP/IP and UDP (User Datagram Protocol [Kes97]) traffic to be mapped into SONET/SDH frames for transmission. The experiments reported here have a high degree of novelty since SDH/SONET transmission experiments have not previously been carried out over MMF.

Error-free transmission over 12.7 km of SSMF as well as over 1.1 and 4.4 km of GI-MMF (restricted centre launch) was demonstrated at 1310 nm using a directly modulated FP laser module operated at OC-192 line rate. The results not only constitute the highest bit-rate distance product reported to date, but also confirm that GI-MMF can be successfully used in the MAN area to provide high optical bandwidth directly to the WAN network.

H.4.2 Experimental set-up

Figure H.8 shows the experimental set-up. Two modules of the RouterTester are connected in duplex mode: Module 1 transmits to module 2 using a MMF link, with SMF input and outputs to match the transmitter and receiver modules. The link between module 2 and module 1 is provided by an SMF patchcord. An external clock operating at 622.08 MHz was provided as time base for module 1; this corresponds to the operating frequency of the electrical multiplexers/demultiplexers contained in each module.

The optical signal was coupled from the SMF into the MMF using a restricted centre launch technique as described in the previous section.
The RouterTester allows three different options of providing a clock to the modules:

- **Internal clock**: Each module contains an internal clock. The disadvantage of this option is that this clock is not provided to the outside to allow external measurements such as eye diagrams.

- **External clock**: An external clock at 622.08 Mb/s, maximal 0.0 dBm into 50 Ω [approx. 632 mV-pp] is provided to either one or all modules. In the case of providing it to the master module (#1), the other modules must operate in ‘recovered mode’ for best results, i.e. their clock signal is extracted from the SDH/SONET bit stream. The external clock, however, can also be provided to all modules.

- **Recovered clock**: A clock signal extracted from the received bit stream is used.

  This option provides best results for transmission over fibre when jitter impairments affect transmission.

Using an external clock from Anritsu PPG (12.5 or 3.0 Gb/s), the transmitter eye diagram as shown in Figure H.9 was measured using the Agilent Technologies 83480A Digital Communications Analyzer with 83482A optical unit containing a 30 GHz optical input. The transmitter specifications (LightLogic) were measured as:

- **Output power**: -2.68 dBm (-3 dBm nominal)
- **Ext. ratio**: 7.4 dB (lower than SDH/SONET requirement of 8.2 or 10 dB)
- **Eye width**: 76.4 ps
- **Jitter RMS**: 3.7 ps
- **Jitter peak-to-peak**: 22.1 ps
- **Centre wavelength**: 1316.38 nm

**Figure H.8: Experimental set-up for testing of the RouterTester equipped with 1310 nm, OC-192 transceiver units**

LightLogic optical transceiver module:
- 16:1 MUX, 16:1 Demux
- Directly modulated DFB laser (7.4 dB ext. ratio)
- P_{out,max} = approx. -3 dB
- Receiver pin + TIA, sensitivity approx. -14.5 dBm
- Electronic format compatible to SDH/SONET OC-192
The receiver unit consists of a pin photodiode followed by a transimpedance amplifier (TIA). Although direct BER measurement was not possible due to the lack of the appropriate software (enabling $2^{23}-1$ PRBS equivalent measurements, to be included in a later version of the RouterTester), SDH/SONET alarms (B1P1, i.e. parity check, error) were observed for a receiver input power in modules 1 and 2 for the following power levels:

- $P_{in,Rx1} = -14.26$ dBm
- $P_{in,Rx2} = -14.16$ dBm

Accordingly, the resulting power budget for the transceivers available in the experiment was approximately 11.5 dB.

**H.4.3 Transmission experiment using SMF and 62.5/125 µm MMF**

**H.4.3.1 SMF transmission experiment**

In a first experiment, the transmitter of module 1 and receiver of module 2 were connected using a SMF of 12.7 km length. Error-free transmission was demonstrated, i.e. no SDH/SONET alarms were recorded for the duration of the experiment. The receiver input power was measured as $P_{Rx2,in} = -8.0$ dBm, resulting in a total span loss of 5.32 dB including connectors. This loss could be significantly reduced when replacing the free-space connectors by splices. Eye diagrams were not measured since the external clock source was not provided for this experiment. As the dispersion of the SMF used is minimum at 1310 nm, it is believed that transmission in a real system is limited due to attenuation of the link (0.35 dB/km at 1310 nm) rather than chromatic dispersion.
H.4.3.2 MMF transmission experiment

Further experiments were carried out using 62.5/125 μm MMF. The aim of the experiments was to use the 10 Gb/s output for MMF transmission experiments at 1310 nm, applying a restricted centre launch via an SMF patchcord into the MMF, and to show the feasibility of error-free transmission using the SONET/SDH protocol at OC-192 line rate over this MMF. The quality of the optical transmission was assessed by recording eye diagrams.

The test was carried out using 62.5/125 μm graded-index MMF in 1.1-km (fibre spool #2) and 4.4-km (fibre spool #7) – all fibres unspliced. The 4.4-km length spool was the same as used in the 2.5 Gb/s transmission experiments described earlier. The differences between colour-graded and regular eye diagrams (Figs. 6.12-13 and 6.14-15, respectively) show that the system performance was strongly impaired by noise (modal noise from speckle pattern) [Pap99] and jitter (no clock recovery).

H.4.3.3 MMF transmission experiment 1.1 km, 62.5/125 μm MMF

Colour-graded eye diagrams for the transmission over 1.1-km of MMF (#2) are shown in Figs. H.10-12, for different traffic type measurements, and with slightly varied input and output launch positions (minimum resolution of micrometer 0.5 μm). The average power into the receiver for this configuration was approximately −7 dBm. For an input power of approx. −2.7 dBm into the system, this accounts for total loss of 4.3 dB in the experiment, including 0.5 dB/km attenuation and connector losses. By far the largest contribution to the total loss is expected from the MMF-SMF transition in front of the receiver. This transition will, however, not only contribute to the loss of the system, but also introduce additional modal noise as explained below.

When compared to the back-to-back eye diagram in Figure H.9, it can be seen that the vertical eye opening is severely distorted by noise, especially on the mark pulses. This is mainly due to modal noise at the receiver end when the optical power is coupled from the MMF into the SMF patchcord. Modal noise originates from speckle patterning at the output of the MMF [Pap99, Cun95]. For highly coherent light sources such as the FP laser used in this experiment, the output power at the end facet of the MMF is not equally distributed over the cross-section of the fibre core, but shows local minima and maxima that vary their position with time (coherence time).
Additional impairments were introduced by jitter, assumed to be mainly due to the fact that the oscilloscope was triggered from the input clock, not on the recovered clock from the received bit stream. The incoming OC-192 signal is internally demultiplexed into OC-12 tributaries without the provisioning of an external recovered clock. Jitter measurements for the eye diagrams in Figs. H.11 and H.12 showed peak-to-peak jitter of 33.4 ps and 37.2 ps, respectively. A result is additional horizontal eye closure, but clear eye openings are obtained in all cases. More detailed results on the performance under various traffic statistics for the TCP/IP packets is provided in section 7.5.
Advanced multimode fibres for high bit rate transmission in LANs and MANs

Figure H.10: Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -6.88$ dBm

Figure H.11: Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -7.2$ dBm, jitter peak-to-peak (pp) 33.4 ps

Figure H.12: Eye diagram for 1.1-km MMF, receiver input power $P_{Rx} = -7.3$ dBm, jitter (pp) 37.2 ps
**H.4.3.4 MMF transmission experiment 4.4 km, 62.5/125 μm MMF**

Eye diagrams for the transmission over 4.4-km of MMF (#7) are shown in Figs. H.13-16. The eye diagrams in Figs. H.13 and H.14 were measured for constant stream of 40 byte IP packets, whereas the colour-graded eye diagrams shown in Figs. H.15 and H.16 were recorded to test the long-term stability.

The average power into the receiver for this experiment was approximately $-9$ dBm. However, severe power fluctuations due to modal noise could be observed from the colour-graded eye-diagrams in Figs. H.14 and H.16, representing long-term eye measurements. With an input power of approx. $-2.7$ dBm into the SMF, the total link loss accumulates to approx. 6.3 dB. This value includes a fibre loss of approx. 2.2 dB (0.5 dB/km over 4.4 km) and coupling losses – which were dominant in this experiment.

A comparison between *short-term* and *long-term* (colour-graded) eye diagrams, Figs. H.13-14 and H.15-16, shows the severe degradation of the eye diagram due to modal noise and jitter. The horizontal eye opening in short-term measurements, Figs. H.13 and H.15, reveals that even at 4.4-km length the combined effects of modal and chromatic dispersion do not affect the eye opening severely.

Modal noise components are stronger than in the case of 1.1-km MMF since the spatial pulse diameter will increase with transmission distance because of mode coupling, where power is transferred from the fundamental mode into higher-order modes. Power fluctuations at 4.4-km length are distributed over a larger cross-section of the fibre than for 1.1-km MMF, causing noise on the mark pulses as can be seen in Figure H.16, for which the eye is nearly completely closed in vertical direction. Since the oscilloscope was triggered from the input clock, timing jitter is increased and contributes to horizontal eye closure as seen in Figs. H.14 and H.16. The system impairments observed for 4.4-km MMF still allow transmission of 10 Gb/s signals, but severely limit the long-term stability of the link in the current configuration. The problem with the set-up used here is the mode-filtering at the receiver end of the link, which results not only in a power loss of approximately 4 dB after 4.4 km of MMF, but in addition suffers from strong modal noise.
**Figure H.13:** Eye diagram for 4.4-km MMF, receiver input power $P_{rx} = -9.0$ dBm

**Figure H.14:** Eye diagram for 4.4-km MMF, receiver input power $P_{rx} = -9.0$ dBm, for the same measurement as shown in Figure H.13
Advanced multimode fibres for high bit rate transmission in LANs and MANs

Figure H.15: Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm

Figure H.16: Eye diagram for 4.4-km MMF, receiver input power $P_{Rx} = -9.0$ dBm, for the same measurement as shown in Figure H.15. The eye was recorded for a 60 minute period. The eye shows significant distortions due to modal noise and jitter.

H.4.4 Transmission experiments with varying traffic statistics

The RouterTester was used for a number of statistical measurements to test the reliability of the POS transmission for different packet sizes and traffic statistics over a 12.7 km SSMF as well as 1.1 and 4.4 km of GI-MMF as described in the previous section.

The RouterTester was able to provide a number of synthetic traffic streams with individually different properties:

- Traffic flows per stream: Up to $2^{16} = 65,636$ different flows (addresses) in a single 10 Gb/s stream
Packet types
- IP (Internet Protocol) packet (40 byte - 64 kB)
- TCP (Transmission Control Protocol) packet (40 byte - 64 kB)
- UDP (User Datagram Protocol) packet

Traffic statistics
- Uniform distribution for various loads
- Bursty distribution defined by ON-OFF periods with periodic, but constant pre-defined intervals as shown in Figure H.17, i.e. arrival and holding times are effectively deterministic. The load is determined by the ratio ON- and OFF-periods, $T_{ON}$ and $T_{OFF}$.

![Figure H.17: Principle of bursty traffic generation in Agilent RouterTester](image)

Data traffic load was continuously adjustable from 0% - 100%. In absolute numbers, the generated traffic is $< 9.95328$ Gb/s due to the SONET/SDH framing overhead.

Based on a variation of these traffic statistics, a number of POS transmission experiments were carried out to test the performance over a MMF link, both on 1.1-km and 4.4-km spools. The results are summarized in Table H.3.

The Agilent RouterTester offers two transmission modes to send IP/TCP/UDP packets: In uniform/continuous mode, packets are sent as a constant stream, whereas they are groomed into bursts for the bursty traffic mode. Since all packets are mapped into SONET/SDH frames and scrambled, the IP/TCP/UDP traffic statistics should have no impact on the physical transmission performance since at all times a continuous stream of bits is transmitted over the link. This is essential to keep the clock at the receiver synchronized to the incoming data. With few exceptions, IPv4 packets were used in the experiments, with lengths varying from 40...65536 byte. Each IP/TCP/UDP packet carries an individual timestamp and information in the payload to allow statistical measurements at the receiver, such as delay and packet corruption/packet loss.
Table H.3 is organized as follows:

1) The first column refers to the physical medium over which the measurement was performed, i.e. 4.4-km MMF (#7) and 1.1-km MMF (#2)
2) The packet type is specified, it was usually chosen to be IP, but in three cases TCP segments were used
3) Packet size varied from 40-byte...64kB.
4) Traffic statistics were either uniform or bursty; for bursty traffic the (constant) number of packets/burst was 200
5) Measurement duration measured in seconds (minimum resolution)
6) No. of SDH parity bit check errors (BIP-1/-2/-3) during the measurements. SDH/SONET triggers an alarm as soon as one of the parity checks (BIP 1-3) fails
7) Time for which SDH/SONET alarms were observed; in combination with 6) this can be used a metric for the burstiness of errors (plotted in Figure H.18)
8) Load is defined as a dimensionless metric and only considers IP/TCP/UDP data, not the SDH overhead
9) Packet loss rate (PLR) is based on the no. of packets which could not be recovered by the receiver and were hence marked as lost

<table>
<thead>
<tr>
<th>Fibre</th>
<th>Packet type</th>
<th>Packet size [byte]</th>
<th>Traffic type</th>
<th>Duration [s]</th>
<th>SDH errors</th>
<th>Error time(^1) [s]</th>
<th>Load</th>
<th>PLR(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4 km MMF</td>
<td>IP</td>
<td>1500</td>
<td>Uniform</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(fibre #7)</td>
<td>IP</td>
<td>1500</td>
<td>Uniform</td>
<td>300</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-64k</td>
<td>Uniform</td>
<td>300</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40</td>
<td>Uniform</td>
<td>60</td>
<td>2.0E4</td>
<td>13</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40</td>
<td>Uniform</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TCP</td>
<td>64</td>
<td>Uniform</td>
<td>60</td>
<td>34</td>
<td>8</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40</td>
<td>Uniform</td>
<td>1200</td>
<td>1.0E5</td>
<td>220</td>
<td>1.0</td>
<td>8.06E-8</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40</td>
<td>Uniform</td>
<td>3600</td>
<td>1422</td>
<td>1.0</td>
<td>1.0</td>
<td>1.49E-8</td>
</tr>
<tr>
<td>1.1 km MMF</td>
<td>IP</td>
<td>40-64k</td>
<td>Uniform</td>
<td>8</td>
<td>110</td>
<td>–</td>
<td>1.0</td>
<td>3.69E-7</td>
</tr>
<tr>
<td>(fibre #2)</td>
<td>IP</td>
<td>40-1600</td>
<td>Uniform</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-10k</td>
<td>Uniform</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-64k</td>
<td>Uniform</td>
<td>60</td>
<td>3</td>
<td>–</td>
<td>1.0</td>
<td>1.34E-6</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-64k</td>
<td>Burst*</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-1600</td>
<td>Burst*</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-1600</td>
<td>Uniform</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>40-1500</td>
<td>Uniform</td>
<td>60</td>
<td>1422</td>
<td>–</td>
<td>1.0</td>
<td>1.37E-5</td>
</tr>
</tbody>
</table>

Table H.3: Summary of packet transmission results on the 1.1-km and 4.4-km multimode fibre. No errors for transmission over 12.7-km SSMF recorded

\(^1\) Parameter for burstiness of errors; timeslots in [s] are measured for which during which at least one SDH/SONET error occurred
\(^2\) PLR = no. of packets not received/no. of transmitted packets
* Burst containing 200 packets
The following observations were made for the POS transmission experiment:

- The minimum sampling time in all measurements was 1 second. The RouterTester software does not log events on timescales smaller than 1s, so it is not possible, for instance, to monitor burst-switching experiments operating at sub-second speeds.

- The traffic statistics and packet length did not affect the PLR on both MMFs. Although the IP packet lengths and the statistics were varied, no effect could be observed. As pointed out before, this result was expected since the SDH/SONET framing and scrambling would guarantee a quasi-PRBS bit sequence on the link regardless of the IP/TCP/UDP traffic applied.

- On the 4.4-km MMF (#7) link, IP traffic was measured over a period of 5 minutes for three different packet lengths: 40-byte (IP acknowledgement), 1500-byte (max. Ethernet frame size 1536 bytes), and 40-byte...64-kB. No SDH/SONET or lost packets were observed during each of the 5-minute periods.

- There is a strong correlation between the SDH/SONET BIP errors (failure of parity check on SDH/SONET frame) and the PLR on both MMFs. Except from one case, the detection of BIP errors always resulted in the loss of IP packets.

- System stability measurement on 4.4-km MMF (#7) link: To investigate the long-term stability of the system, a continuous stream of 40-byte IP packets was sent over the link for 1 hour, resulting in an average PLR of 1.49×10⁻⁸. Figure H.18 shows the arrival statistics of BIP-1 errors as a function of the measurement time. The results confirm that errors in the link are not equally distributed, but occur in bursts. Error-free intervals of more than 10 minutes are interspersed with large numbers of parity-check errors that can reach the number of several thousands per second at the peak. This is mainly due to severe modal noise and jitter resulting in nearly complete eye closure of the transmitted signal. However, according to the RouterTester statistic, the corresponding number of lost packets was relatively low with just 876 not recoverable, resulting in an overall PLR of 1.49×10⁻⁸ for a period of 60-minutes.

- The available power budget did not allow the insertion of an additional optical attenuator to carry out PLR vs. received optical power measurements. The insertion loss of the optical attenuators (> 4.5 dB for Hewlett Packard, > 3.0 dB for manual attenuator) was so high that BIP-1 errors were observed in intervals of less than 1 minute.
Figure H.18: Arrival statistics of SDH/SONET BIP-1 errors as a function of the elapsed measurement time, showing the burstiness of errors observed in the experiment.
H.5 Summary and conclusions

Transmission results from two sets of experiments using advanced 62.5/125 μm GI-MMF demonstrated that bit-rate-distance products of up to 40 Gb/s-km were achievable when using a restricted centre-launch technique. The results were confirmed both by BER measurements and using a Packet-over-SONET (POS) test link, showing that GI-MMF with improved profile characteristics allow transmission well in excess of current 10-Gigabit-Ethernet specifications. Penalty measurements confirm that the transmission penalty is related to deviations of the profile from a perfect α-profile.

A key result of the investigation shown here is the need for MMF fibre profiles which follow the targeted α-profile closely, particularly in the centre of the fibre since this is where the optical intensity is concentrated. A potential area for future research, arising from both the experimental work presented here and modeling [Ord00], is the investigation of how index profile distortion affect the fibre bandwidth when the modal dispersion becomes negligible.

The experimental results on GI-MMF with improved index profiles show that this type of fibre has a high potential for providing a cost-effective transmission medium not only in LAN, but also campus or MAN applications. The capacity-distance product (40 (Gb/s)-km achieved in the experiments reported in this appendix indicates that advanced GI-MMF provides sufficient bandwidth to be used even for applications exceeding the specifications of current 10GbE. Advanced MMF could be used as a cost-effective transmission medium to allow users to access to the high-capacity envisaged for future optical network architectures discussed earlier in this thesis and benefit from new applications operating over such high-bandwidth connections.
H.6 References


[Cun02] D.G. Cunningham, “10 Gigabit Ethernet: from standards to application,” Proc. 28th European Conf. on Optical Commun. (ECOC 2002), vol. 1, tutorial 1, Copenhagen, Denmark, 8-12 Sep. 2002


Advanced multimode fibres for high bit rate transmission in LANs and MANs


