Distributed Router Architecture for Packet-Routed Optical Networks

Michael Düser, Eugene Kozlovski, Robert I. Killey and Polina Bayvel

Optical Networks Group, Department of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1E 7JE, UK , Tel.: +44-20-7679-3843, Fax:+44-20-7388-9325, e-mail: mdueser@ee.ucl.ac.uk

Abstract: A new proposal for an optical packet-routed network based on a distributed router architecture in a WDM network is described. Buffering, scheduling and wavelength assignment functions are carried out in routers located at the edge of the network. The destination address is encoded by a given wavelength and allows wavelength-routing in optical core routers. This approach avoids processing and additional buffering delays in core routers. A low packet loss rate and a low latency are critical parameters to guarantee QoS for a particular class of packets and must be carefully considered when IP traffic is directly transmitted over a photonic network. This network architecture requires new hardware elements and these are discussed including a low-crosstalk, large port-count free-space grating router.

1. INTRODUCTION

With the projected annual increase of 60% in data traffic [1], solutions are required to combine the advantages of the network packet layer with the high-bandwidth optical transmission in the physical layer. To date the proposed solutions concentrate on current protocols and transmission standards namely ATM and SDH, such as in proposals of IP-over-ATM [2] and packet-over-SDH transmission [3]. Whilst these simplify the protocol stack, limitations in scaling to a large number of routers remain. The advantage of connection-oriented protocols such as ATM is their provision

of quality of service (QoS) awareness, usually not considered in connection-less protocols like IPv4.

These problems can be overcome using a new protocol proposed by the Internet Engineering Task Force (IETF) in 1997 [4] known as multiprotocol label switching (MPLS) in combination with optical routing. MPLS, like ATM lies between the network and the link layer but, in contrast to ATM, can operate with variable packet lengths, and directly with IP, avoiding the necessary adaptation of packets to ATM cells.

In this paper we describe an MPLS-based optical network architecture that does not rely on processing and buffering in the core nodes as recently suggested [5]. It is proposed to distribute the routers to the edge of the network, equipped with electronic processing, scheduling and buffering. With packets aggregated to flows, virtual lightpaths are established via MPLS between the ingress and the egress router of the network. MPLS can be additionally used to interconnect the core networks of network operators.

2.   MULTI PROTOCOL LABEL SWITCHING (MPLS)

Although MPLS is not a defined standard, it is widely discussed in the Internet Engineering Task Force (IETF) [4] with the following main functions:

* *Short label*

Since MPLS is designed to allow a hop-by-hop routing, it is sufficient to use a short label instead of a full header. This makes it particularly suitable for WDM because the MPLS label with a length of 32 bits [6] (Fig. 2) gives a reduction of a factor of 10 compared to the 320 bits header envisaged for IPv6. With this type of header it is possible to distinguish more than 1 million destination addresses, 8 classes of service, and to identify the bottom label of a label stack. The time-to-live field (TTL) allows a maximum number of 256 hops before the flow is discarded. As labels can be attached to any packet, MPLS can deal with variable length packets such as IP packets (max. 64 Kbytes payload). This is different from the ATM approach with short, fixed size ATM cells (53 byte).

Address label (20 bit)	Class of Service (3 bit)	Bottom of stack (1 bit)	Time to live (8 bit)

Figure 1. Example for a 32-bit MPLS header [8]

- *Hierarchical label stacking*

MPLS allows to stack several labels on top of each other before sending the data flow from the first edge router. Labels are organised as a stack (last-in-first-out [LIFO]) so that the top label determines the next hop. This allows for temporary labels on a MPLS flow, important in case of re-routing.
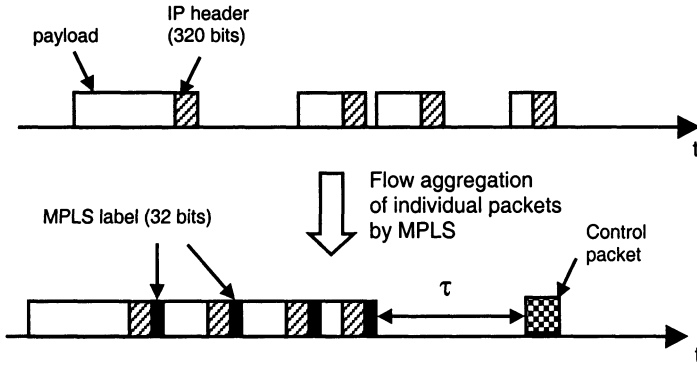

Figure 2. Aggregation of a MPLS flow from single IP packets by adding MPLS labels (32 bits) plus an additional control packet with an offset time τ proceeding the flow

- *Virtual circuit emulation using label-switched paths (LSPs)*

    Like ATM, MPLS sets up virtual paths between edge-routers, which are called label-switched paths (LSPs), reserving a particular wavelength for the transmission.

- *QoS routing*

    Routing is carried out according to the class of service, i.e. packet flows originating from the same edge router for the same destination but containing different classes of are forwarded on different LSP. Transmission on different wavelengths is therefore inherently supported by MPLS. The implementation can either use a Resource Reservation Protocol (RSVP), or a Label Distribution Protocol (LDP) [7].

- *Failure detection and restoration*

    Situated between the network and the link layer, MPLS does not operate at the physical layer of the network and will therefore provide restoration after measures on the physical level are taken. Since MPLS controls the router and hence the release of flows into the network, a fast forwarding of the failure information (tens of ms) to the MPLS control is essential to start re-routing of LSPs [8].

3.   NETWORK ARCHITECTURE - A BURST-SWITCHED OPTICAL NETWORK

Currently, wavelength-division multiplexing (WDM) is used for high capacity point-to-point links connecting electronic network routers. Although electronic routers can route and forward packets, their scalability is limited beyond 1 Tbit/s [9].

The ideal infrastructure would evolve from a static wavelength-routed optical network with fixed lightpaths to an all-optical connection-less network resembling an IP network. It is assumed that switching timescales are reduced from months or days for the static model to nanoseconds in the case of a network forwarding at the packet level, in response to load changes.

The principle of burst-switching in the context of WDM networks has been investigated [10] [11], but to date has not been applied to MPLS and optical routing. The principle of this approach is to aggregate the packets to flows (Fig. 2) in the edge routers to reduce the number of wavelength-switching operations. Due to the time required to tune the laser diodes or wavelength converters to the appropriate wavelength the achievable switching speed is limited. Since MPLS operates on a flow-level, and not on a packet-level, packet aggregation is essential. MPLS has the advantage of being quasi-circuit switched. In this architecture, packets are sorted by traffic type and destination in queues in the edge routers. In this scheme we assume 3 classes of service:

- Premium: Traffic with low packet loss rate (PLR) ($< 10^{-9}$ for traffic load 0.8 as in ATM) and nearly no latency (quasi circuit-switched)
- High Quality: Packets with specified PLR and a specified maximum latency
- Best Effort: Regular IP packets with high tolerance towards packet loss and latency

The length Q of the queue is a critical parameter and must be investigated in the context of the proposed MPLS architecture. The size Q of a queue required to store incoming packets is a function of the queueing delay τ_{dq} and the bitrate b of the network:

For example, at a bit-rate of 10 Gbit/s a buffer size of 1.24 Mbyte is required to buffer the traffic passing through in 1 ms. The queues are emptied either when the size of the queue reaches a given threshold or a time-out signal is sent to guarantee a specified acceptable delay. Large memory and processing capacity are concentrated in the edge router, whilst the optical routers in the core act as a large, distributed routing network. The advantages of this scheme over other architectures are as follows:
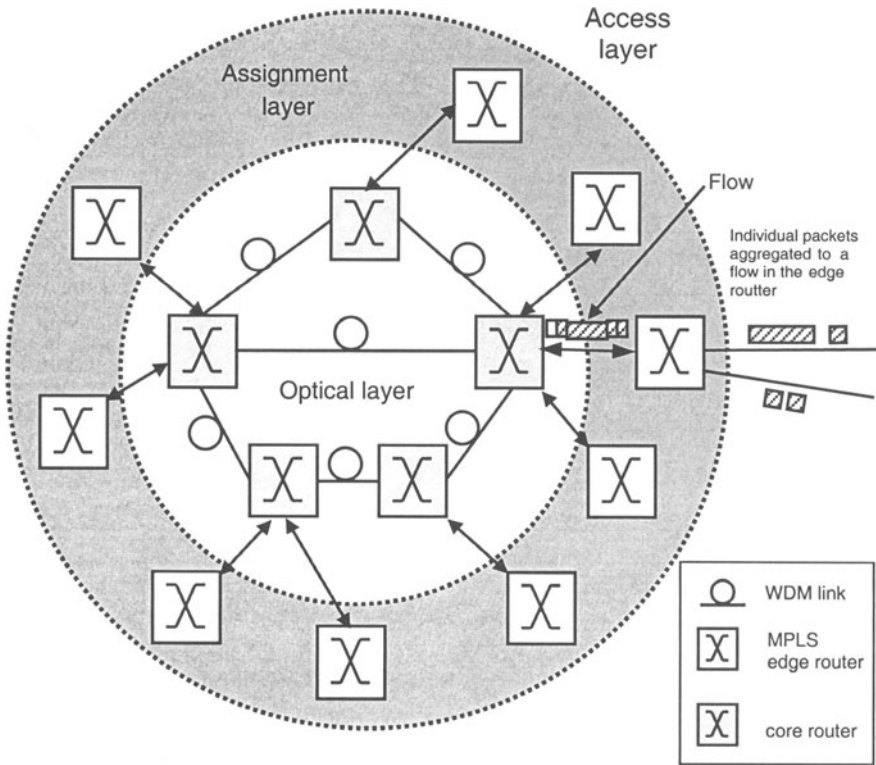
Figure 3. Principle of the proposed network architecture applying burst switching

The millisecond timescale for sending flows through the network allows sufficient time for wavelength selection in transmitter lasers and wavelength converters. Working at the flow level and establishment of short-term connections between edge-routers allows compatibility with the MPLS protocol for the transmission over and control of the optical layer.

The packet latency is determined by the queueing delay prior to release of the packet flow into the core network and directly proportional to its length Q. Once a flow enters the core network it experiences only the propagation delay. Thus queue "threshold" and "time-out" determine the QoS for a particular type of packet.

The edge routers in the distributed router architecture (Fig. 3) connect the access with the optical core network. In addition to the processing, scheduling and buffering these must carry out the functionality of mapping

incoming packets to the optical layer of the network. This is done in two steps (Fig. 4):

(1) The packets are aggregated into MPLS flows. This is carried out by reading and processing of the entire packet header information (320 bits for IPv6), the scheduling of the packet and the storage in the correct output queue. This is carried out in the electronic domain where sufficient memory and parallel processing capacity are available to cope with optical line speeds $\geq$ 10 Gbit/s. The time a flow spends in the buffer depends on the specified maximum delay for a given class of packets and will be on the timescale of < 1 ms for High Quality type traffic.

(2) The second function needs to allocate flows and reserve bandwidth from the input edge router to output edge router through the core network. This is done by matching the address of a flow to the wavelength so that within the core network no more header lookups are required.
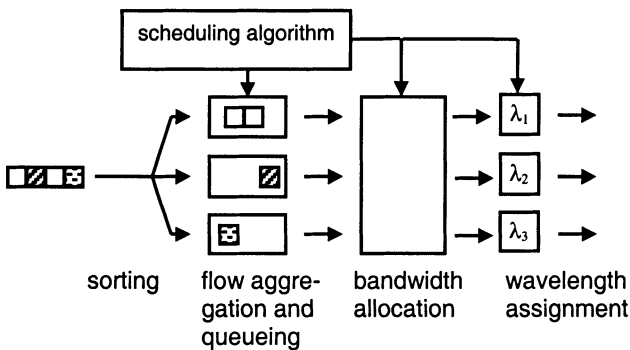

Figure 4. Treatment of packets in the proposed network architecture

The proposed architecture is based on an optical burst-switching (OBS) architecture in combination with MPLS for operation of the network and transmission between edge routers. The scheme operates on three different layers specified in table 1:

• *Access layer*

Packets with different input formats, but mainly originating from IP applications enter the edge routers. MPLS is not restricted to a particular format but is designed to operate with different protocols such as ATM, IP, etc.

• *Assignment layer*

The trade-off is between the length of the queue and the overprovisioning of bandwidth, i.e. the number of wavelength required in the network. The blocking probability or delay can be reduced by providing more wavelengths

channels, but in addition to more hardware this increases the network control and management complexity.

Table 1. Reference model for the proposed network architecture using burst-switching

Layer	Functionality
Access layer	Incoming packet: IP, voice, video
Assignment layer	1. Synchronisation 2. Incoming packet header detection and processing 3. Data rate adaptation and convergence 4. MPLS header writing/removal 5. Scheduling 6. Packet aggregation to form flows whilst packets are queued according to packet class and destination 7. Wavelength assignment (layer 3 routing) 8. Bit-rate control
Optical layer	Switching (layer 2) Transport 3-R regeneration

The number of queues N_{queue} that need to be controlled per edge router is determined by the number of edge routers N_{edge} and the number of classes C supported by MPLS:

$$N_{queue} = N_{edge} \times C$$

Assuming a network with 20 core routers and 10 edge routers connected to each core router (N_{edge} = 200) and 8 classes of service as envisaged for MPLS (C = 8), this would result in 1600 queues to be operated by each edge router. In the case of 10 Gbit/s system this requires a capacity of approximately 2 Gbytes per ms of traffic that needs to be stored when the queues are operated independently.

The key function of MPLS is shown in Fig. 5. As MPLS is situated between layers 3 and 2 in the network, it can operate on top of SDH (backward compatibility) or digital wrappers (future component), or it can even run directly on the WDM domain. In this case network failures must be reported in the timescale of tens of milliseconds to the MPLS level to allow fast restoration and re-routing in the network.
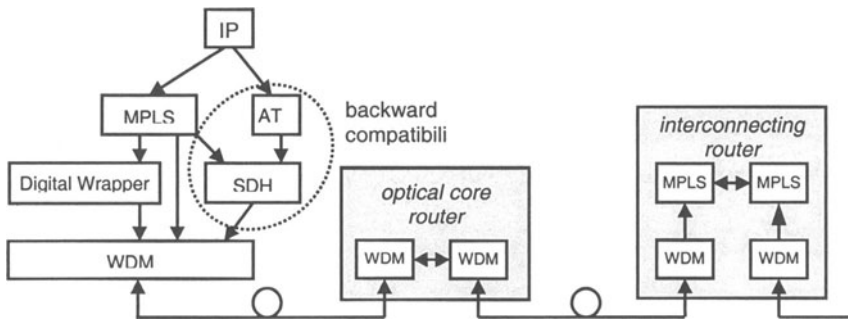
Figure 5. Protocol layers used whilst a packet is sent via different core networks

- *Optical layer*

Transmission and routing of flows is carried out in the optical domain by a core wavelength router where no optical buffering and processing functions are required. However, network operation without processing and buffering in core nodes requires a new scheduling algorithm to provide QoS for the different types of classes for packets in the network. Flows buffered in queues are assigned to a free wavelength, equivalent to the bandwidth reservation.

The most suitable reservation scheme is JET (Just-Enough-Time) [13]. A list of available wavelengths is stored in association with a central scheduling algorithm from which each router obtains free wavelengths as required. The control packet reserves a particular bandwidth for a fixed duration allowing a definite forecast when the wavelength will be available again. JET is important for QoS awareness in the network by assigning different priorities to the flows buffered in the edge router. Priorities are associated with different blocking probabilities and different queuing delays, i.e. the lower the priority, the higher is the associated blocking probability and the queuing delay.

4.   CLASS OF SERVICE DIFFERENTIATION BY LIGHTPATH AND DISTRIBUTION TREE ROUTING

Wavelength routing allows bypassing buffering and scheduling at the core routers by providing as many wavelengths as required for the complete contention resolution in each link of the core network.

In the proposed architecture, IP traffic arriving at the ingress router is buffered and scheduled electronically and each aggregated flow is attached an MPLS label and then assigned a wavelength to route the flow along the entire path to the destination. At the expense of wavelength overprovisioning, this eliminates the problem of packet header processing at each intermediate node. Depending on the QoS requirements for a particular traffic class, two methods for wavelength assignment with different amounts of overprovisioning are considered in this section.

1. *Lightpath assignment.* This type of assignment provides direct connections between source and destination and requires no processing or buffering in the optical domain [12]. Therefore, lightpath assignment is preferable for accommodating real-time traffic. Each lightpath is assigned a unique wavelength along all the links between edge routers. The wavelength can be the same in every link of the lightpath or it can vary from link to link along the path in case of wavelength conversion [13]. Let V be the set of nodes, $|V| = N$ and let the set of lightpaths be

$$P=\{p_{ij}(s_i, d_j) \mid i = 1,\ldots, N, \; j = i,\ldots, N\},$$

where s_i is source node i, d_j is destination node j, p_{ij} is the lightpath between s_i and d_j. Thus, total number of unidirectional lightpaths in the network is $|P| = N(N-1)$.

2. *Distribution trees assignment.* This type of assignment is provided by the set of distribution trees

$$T = \{t_j(s_i, D_i) \mid i = 1,\ldots, N\},$$

where s_i is source node i, $D_i = V \setminus s_i$ is the set of destinations for the source s_i, t_j is the tree routing packets from s_i to the set D_i. The total number of trees is $|P| = N$.

Thus, a single wavelength is assigned to all the edges of each distribution tree t_j which can route traffic from source node s_i to all the destinations of the core network [14] (see Fig. 6 for an example of lightpath and tree assignment in a 4-node network).

Whilst, with the distribution tree approach, the wavelengths are shared among different destinations for each source, packet contention at the intermediate nodes is avoided, as each tree is assigned a unique wavelength along its edges. Packet buffering and scheduling is carried out in the source edge-routers, with packets being transmitted only when the required wavelength becomes available.
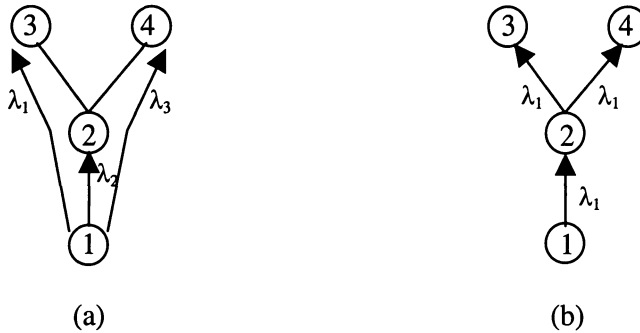
Figure 6. Difference between lightpath and distribution tree assignment in a 4-nodes network. (a) Three wavelengths are required to connect node 1 with the other nodes by lightpaths. (b) One wavelength is required to connect node 1 with the other nodes by a tree as packets routed from one source edge-router to different destination edge-routers share the same wavelength.

The algorithm used to build distribution trees is as follows.

(1) For every source node, a shortest-path tree spanning all the others nodes is calculated.

(2) Since for real network topologies there is usually more than one way to span the nodes with shortest paths, a set of all the possible shortest-path trees is built for a given source node.

(3) All the trees in the set are consecutively considered as the distribution tree for a given source node, and a shortest path tree is chosen as distribution tree if, and only if, the number of wavelengths in the most loaded link of the network is lower than using any other tree in the set.

(4) Steps (2) – (3) are repeated for all the nodes in the network. As a result, distribution tree is assigned to every node and wavelengths are reused so that packets coming from the same source through the given link share the same wavelength.

Consider any node $s_i \in V$. In order to provide unidirectional logical connections to all the other nodes, it must originate N-1 lightpaths, and therefore, requires exactly N - 1 transmitters. At the same time, in order to provide logical connections with all the other nodes by distribution trees, fewer transmitters are required because for each source node $s_i \in V$, several destinations can share the same wavelength on a given fiber, depending on the network topology. Therefore, it can be shown that the total network load, i.e. the number of wavelengths in the most congested link, can be significantly reduced by means of distribution trees compared to lightpaths.

However, the distribution trees impose longer delays on traffic delivery because packets destined to different nodes can only be carried serially along those edges, which share the same wavelength. Additionally, low-level

routing is required at each intermediate node in order to route packets to a particular destination along the tree.

Nevertheless, distribution trees represent a trade-off between the number of wavelengths required and the queuing delays at edge routers, and can be used to accommodate lower-priority traffic. Preliminary calculations show that use of distribution trees can reduce the wavelength requirement in such networks as NSFNet, ARPANet and USNet by up to 40% compared to lightpaths. The number of bits required to code packet destination along the tree can be derived from the network diameter. This number can be shown to be almost the same for real network topologies and varies from 2 to 4 bits for the above-mentioned networks.

Hence, considering that packet class-of-service (CoS) field is coded by the assignment type, the entire MPLS header can be mapped into the combination of the wavelength code and a small number of forwarding bits allowing fast switching in the core routers (Fig. 7).

Therefore, distribution trees allow for buffer-free packet delivery with only a small forwarding label to be processed in the optical domain. Another advantage of distribution trees is in their ability to provide direct multicast connections without queuing delays, as in the case of multicast, packets can be transmitted simultaneously [14].
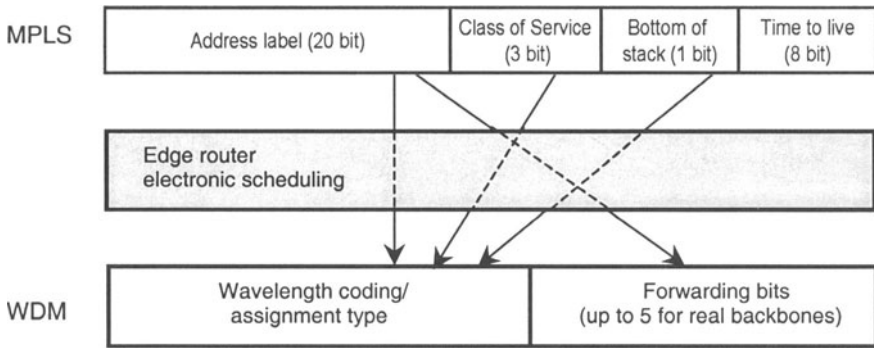

Figure 7. Mapping MPLS header to WDM domain.

5.   EXPERIMENTAL NETWORK ARCHITECTURE

The experimental investigation of the proposed network architecture requires a packet generator and analyser, a fast tuneable laser or a set of switchable transmitters and/or wavelength converters, and a large-port

optical router. Transmission of packets can also be studied by integrating the label generation and error detection in a single set-up (Fig. 8). Long distance and multi-hop transmission of packets can be studied using a recirculating fibre loop.

The generation and analysis of optical packets at line rates of 10 - 100 Gbit/s throughout is carried in a computer used for the control of the set-up and aggregation of flows. According to the destination address and class of service of the flow a free wavelength is selected and the pattern is simultaneously transferred to the pulse pattern generator (PPG), to modulate the cw wavelength carrier of the multi-wavelength transmitter. The wavelengths carrying flows are routed by the free-space grating router (expected to be at ns rates). The wavelength tuning delay could be reduced by using a multi-wavelength array rather than a tuneable laser.
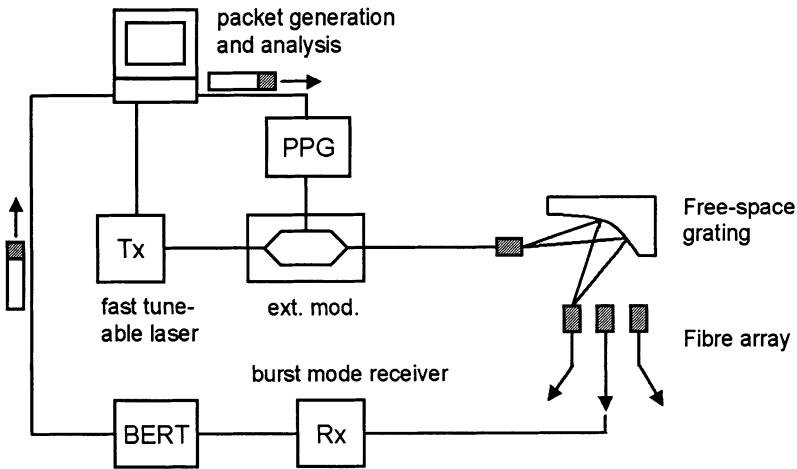

Figure 8. Free-space grating router in DEMUX configuration for back-to-back measurements for distributed switching

Optical switching matrices suffer from problems in the scalability for large numbers of ports [15] due to crosstalk and high insertion loss of the elements. Free-space routers overcome the limitations of waveguide-based routers as they inherently can achieve lower crosstalk values and a free-space grating device (Fig. 9) with < 40 dB crosstalk and better than 3.7 dB fibre-to-fibre loss designed for 100 ports [16] and operating at 10 Gbit/s was recently demonstrated [17]. This achieves the required switching matrix by distributing the switching functionality to the edges.
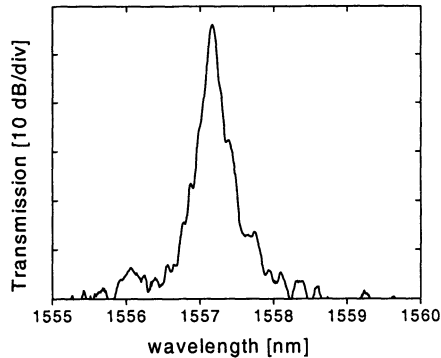
Figure 9. Measured passband of free-space grating demultiplexer [17]

Other key elements are:
* *Burst mode receiver*

Packet error detection requires the implementation of burst mode operation. This requires synchronisation with incoming optical packets at nanosecond timescale. Specifically a separate synchronising pattern is required to align the incoming data stream with the detector reference clock. Use of a burst mode packet receiver will allow the simultaneous measurement of queueing delay, PLR and traffic load. This will allow defining the design rules for a given network.

* *Packet loss, delay and traffic load*

While the BER is the correct measure of the performance of a physical link, QoS parameters such as PLR and packet delay are additionally important for the evaluation of network performance. This requires a three-dimensional surface plot for the graphical representation of queueing delay, PLR and traffic load. The graph can be used in two ways: Either for plotting simulation results showing the performance resulting from a given network topology, or QoS parameters can be specified by a given mask whose values may not be exceeded by the real network.

* *Number of hops*

An important parameter in the design of optical transmission networks is the scalability in terms of number of physical hops. The number of hops a packet can undergo in a photonic network is limited by distortions originating both from the fibre transmission characteristics and distortions added by the optical router, such as crosstalk and insertion loss. The eye closure for a randomly connected network (Fig. 10) decreases linearly with

the number of cascaded channels. Transmission through 12 routers an additional 3 dB penalty is introduced to the system.
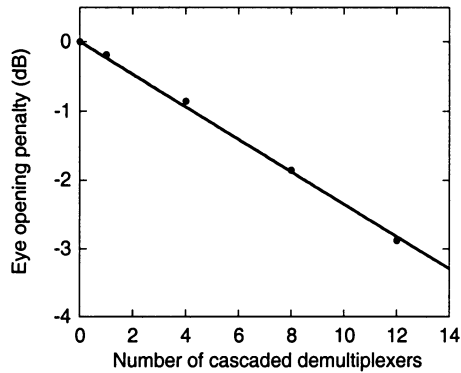
Figure 10. Calculated eye closure penalty with transmission through cascaded demultiplexers, with passband T (dB) = - 0.014 $(f - f_0)^2$

As bit-rates per channel increase, and optical power is increased to maintain adequate signal-to-noise ratios with large inter-amplifier spans, the nonlinear effects of the transmission fibre cause signal distortion, which must be minimised by optimisation of the dispersion map. One technique is undercompensation of the fibre dispersion to avoid pulse broadening due to self-phase modulation (SPM). Fig. 11 shows the amount of under-compensation, achieved with lumped anomalous dispersion at the receiver, required in 10 Gbit/s NRZ format transmission over standard fibre (D = 17 ps/nm·km), with post-compensation in each 80 km span, and 8 dBm launch power into each span. Without the residual dispersion, the eye closure penalty is > 6 dB after 10 spans, while with undercompensation, it is < 1 dB. Although this technique is very effective at extending transmission distances, the optimum amount of residual dispersion is dependent on the transmission distance, and the rapid switching of signals over different routes in burst switched networks will require careful design of the dispersion maps or the development of rapidly variable dispersion compensators.

- *Network topology - total number of edge and core routers*
  The number of hops is also related to the number of nodes present in the core network. The proposed burst-switching networks need to be carefully controlled to use bandwidth efficiently and to detect and resolve conflicts. This functionality must be embedded in the scheduling algorithm and will be restricted by the total number of nodes and the connectivity.
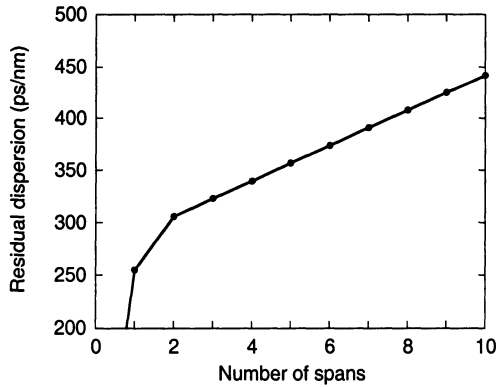
Figure 11. Calculated residual dispersion required to minimise self-phase modulation distortion in multi-span 10 Gbit/s transmission over standard fibre, with 8 dBm launch power

As signalling is required, the round-trip-time (RTT) of control packets limits the size of the network. An additional parameter in the course of the investigation is the maximum number of edge routers that can be connected to the optical core network.

- *Length of burst and queuing delay*

The system performance must be investigated as a function of the length of the bursts and the queuing delay. The trade-off for a reduction of the PLR is an increasing delay because flows have to be stored in memories for longer before being released into the network. Finding the optimum balance between length of burst, queuing delay and PLR is key for the optimisation of optical packet-routed networks.

6. CONCLUSIONS

The advantages of integrating packet transport into the optical domain using the MPLS protocol have been described. Aggregation of packets into flows at edge routers allows the set-up of virtual connections between edge-routers that are passively wavelength-routed inside the network. The bandwidth offered by the large number of channels available in WDM can be efficiently exploited by MPLS, which allows to efficiently set-up different virtual connections for different classes of service whilst utilising the quasi-circuit switched nature of wavelength-routing.

ACKNOWLEDGEMENTS

REFERENCES

[1]   P. Lagasse et al., "Roadmap towards the Optical Communication Age - A European view by the HORIZON project and the ACTS Photonic Domain", Ghent, June 1998

[2]   P. Dumortier, "Toward a New IP over ATM Routing Paradigm", IEEE Comm. Mag., Jan. 1998, pp. 82 – 86

[3]   J. Carlson, E. J. Hernandez-Valencia, N. Jones, P. Langner, "PPP over Simple Data Link (SDL) using raw lightwave channels with ATM-like framing", IETF draft draft-ietf-pppext-sdl-pol-00.txt, June 1999

[4]   R. Callon, G. Swallow, N. Feldman, A. Viswanathan, P. Doolan, A. Fredette "A Framework for MPLS", IETF draft , draft-ietf-mpls-framework-05.txt, Sep. 1999

[5]   F. Callegati, K. C. Cankaya, "Design Issues of Optical IP Routers for Internet Backbone Applications", IEEE Comm. Mag., Dec. 1999, pp. 124 – 128

[6]   T. Li, "MPLS and the Evolving Internet Architecture", IEEE Comm. Mag., Dec. 1999, pp. 38 – 41

[7]   A. Viswanathan, N. Feldman, Z. Wang, R. Callon, "Evolution of Multiprotocol Label Switching", IEEE Comm. Mag., May 1998, pp. 165 – 173

[8]   T. M. Chen, T. H. Oh, "Reliable Services in MPLS", IEEE Comm. Mag., Dec. 1999, pp. 58 – 62

[9]   K. Habara, T. Matsunaga, K.-i. Yukimatsu, "Large-Scale WDM Star-Based Photonic ATM Switches", J. Lightwave Tech., vol. 16, no. 12, Dec. 1998, pp. 2191 – 2201

[10]   J. S. Turner, "Terabit burst switching", J. of High Speed Networks, vol. 8, no. 1, 1999, pp. 3 – 16

[11]   C. M. Qiao, M. S.Yoo, "Optical burst switching (OBS) - a new paradigm for an optical Internet", Journal of High Speed Networks, vol.8, no. 1, 1999, pp. 69 – 84

[12]   S. L. Danielsen, B. Mikkelsen, C. Joergensen, T. Durhuus, K. E. Stubkjaer, "WDM packet switch architecture and analysis of the influence of tuneable wavelength converters on the performance", J. Lightwave Tech., vol. 15, no. 2, Feb. 1997, pp. 219 – 227

[13]   S. Baroni, P. Bayvel, "Wavelength requirements in arbitrarily connected wavelength-routed optical networks", J. Lightwave Tech., vol. 15, no. 2, Feb. 1997, pp. 242 – 251

[14]   X. Jia, "A distributed algorithm of delay bounded multicast routing for multimedia applications in wide area networks", IEEE/ACM Trans on Networking, vol. 6, no. 6, Dec. 1998, pp. 828 – 837.

[15]   C.-S. Wu, G.-K. Ma, B.-S. P. Lin, "Extended Baseline Architecture for Nonblocking Photonic Switching", J. Lightwave Tech., vol. 15, no. 5, May 1997, pp. 771 – 778

[16]   E. G. Churin, P. Bayvel, "Design of Free-Space WDM Router Based on Holographic Concave Grating", Photon. Technol. Letters, vol. 11, no. 2, Feb. 1999, pp. 221 – 223

[17]   V. Mikhailov, P. Bayvel, "Low-crosstalk, loss and PDL grating demultiplexer for application in terabit routed networks", to be published in: Proceedings of OFC2000, 5 – 10 March 2000, Baltimore, MD