Guiding Mice and Elephants in an Optical World

Alexander A. Kist and Richard J. Harris
RMIT University
BOX 2476V, Victoria 3001, Australia
Email: kist@ieee.org, richard@catt.rmit.edu.au

***Abstract*— Recent years have seen major efforts to converge existing circuit switch telephony and Internet services into one network, governed by the Internet Protocol suite. The rapid traffic increase in this consolidated network is accommodated by optical networking technologies. Quality of Service in such carrier grade networks has become a major concern.**

Flow-based networking can help to address these challenging issues since flows are the natural smallest unit where behavioural requirements can be applied. This paper outlines flow-based networking and introduces a method for flow-based overflow routing in an optical MPLS/GMPLS network.

I. Introduction

Traditionally, there are two principal switching/routing paradigms used in communication systems, i.e. circuit switching and packet switching. The current telephony system is circuit switched and the Internet is packet switched. Packet switching requires complex routers as the outgoing interface is determined on a packet-by-packet basis. Major advantages of packet switched networks are scalability and considerable gains in resource utilisation due to statistical multiplexing gains. Internet Protocol (IP) networks are generally packet based, but from an applications point of view, traffic appears as flows rather than packets. For many applications single packets are meaningless, the relevant information is spread over many packets. The information is naturally clustered, i.e. several packets constitute a file, an email message, a telephone call, a *Virtual Private Network* (VPN) tunnel etc. Packets that have a meaning by themselves are rare in the user domain; however, network management and signalling functions use short packets.

The aim of this research is to propose a routing scheme that automatically distributes network load in the case that paths in the network become overloaded. Currently, there is no mechanism available that allows the routing of traffic on the basis of single flows, therefore allowing for dynamic overflow routing. This principle is widely used in *Public Switched Telephone Networks* (PSTNs), examples include Dynamic non-hierarchical routing (DNHR) [1] which uses different path sets for different times of the day, *Dynamically Controlled Routing* (DCR) [2], *Dynamic Alternative Routing* (DAR) [3] and *State- and Time-Dependent Routing* (STR) [4].

Methods that allow load distribution in IP networks include *Multiprotocol Label Switching* (MPLS) [5] which introduces a connection oriented model to IP environments, and separates data and control plane functions. The generalised version of MPLS, GMPLS, extends its functionality to the management of almost any network element, even supporting non packet switched technologies. Other work proposes load distribution by *Open Shortest Path First* (OSPF) weight optimisation to spread network load more evenly and enhance the network's ability to cope with single route failures ([6] and [7]).

None of the above described methods allow load distribution on the fly. The aim of this work is to use a simple networking paradigm of flow-based routing for reliable and efficient routing strategies. *Caspian Networks*, a start-up business, promotes flow-based routers [8]; however the general notion, as such, has not received much attention in the research community. But the concept of microflows is widely used. For example, the *Equal-Cost Multipath* (ECMP) [9] mechanism used by the OSPF protocol utilises flow information to split larger flow aggregates across alternative interfaces.

Routing in optical networks has to address two problems: Firstly, the optical light path allocation has to be found, and secondly, the IP layer packet routing has to be addressed. This paper is concerned with the second problem. Discussions about routing and switching in Optical Networks centre on how different technologies, i.e. *electrical routers* and *optical cross connects*, are best used, connected and managed. *Optical Flow Switching* (OFS) proposes the use of optical end-to-end connections to bypass electronic and IP layer routing in networks and the establishment of light paths for large data transactions. Many other traffic engineering solutions are proposed for next generation optical networks. These cover areas such as optimised shortest path design in GMPLS networks [10] and combined optical IP routing and grooming [11]. Many of these schemes require complex system implementations and/or advanced knowledge of traffic demands. The proposal in this paper assumes a hybrid architecture of routers that are connected by an optical core network governed by GMPLS as the control plan/signalling protocol.

Discussions use the notion of flows to describe network traffic. *Microflows* are defined as a collection of packets with the same source and destination address, the same source and destination ports and are separated by interarrival times which are below a maximum threshold. *Flows* are the aggregation of microflows. Both are measured in bytes per second. Internet traffic typically includes very small flows that consist only of single packets, and massive flows that account for a considerable percentage of the overall traffic. These flow types are commonly referred to as *mice* and *elephants* flows, respectively.

This paper is organised as follows: Section II discusses the concept of flow-based routing in more detail. Section III introduces a proposed solution for MPLS/GMPLS networks and illustrates the operation of the scheme. The paper concludes with a discussion of future work in Section IV

II. Flow-Based Routing

Since IP transport is being used for real-time multimedia applications in carrier grade networks, QoS considerations have become an increasingly important issue. To be able to guarantee QoS on a flow level, flows have to be identified. Current standard routers have no mechanism to do this. Since packets, in the case of congestion, are randomly dropped, loss can affect any flow. In reaction to packet loss, TCP will reduce its transmission rate. Flows that use UDP have no native mechanism to reduce the rate, although many multimedia applications use UDP for its performance advantages with real-time applications. If routers are able to separate flows, new routing paradigms are possible and flows can be treated transport protocol independent. QoS functions, such as policing and shaping can be applied to flow-based routing, flows can be rejected or overflown to alternative paths. Flow-based routing also enhances scalability, since large flow aggregates can be split over several routers.

A *Scheme for Alternative Packet Overflow Routing* (SAPOR) is introduced in [12] and it proposes the use of similar methods to ECMP to identify flows and utilises principles that are used by MPLS to forward packets within a router. This scheme was also suggested for an outbound Internet service provider routing scenario [13]. SAPOR implements three principles: Firstly, it ensures that packets that belong to the same microflow are routed on the same interface. This is also guaranteed in the case of overflow. Secondly, it determines the number of additional microflows that can be accommodated by the default link before its target bandwidth is reached. And lastly, if the target bandwidth is reached, additional flows are routed on alternative interfaces.

These goals are achieved by recording a minimum amount of local state information for each active flow. The information is captured in a tuple that consists of an interface identifier and a unique hash value. The hash value represents one microflow, i.e. the five-tuple origin and destination address, origin and destination port number, and protocol ID. Flow count and utilisation are measured on all interfaces and the number of possible additional microflows is estimated by the calculation of the average flow size and the given target bandwidth. If connectivity and routing information is required by SAPOR, it is acquired from the tables of the used standard routing protocol such as OSPF. SAPOR requires only a minimal amount of state information and network activity is not required. Therefore, this approach is scalable to a large number of flows.

If network operation and routing is based on micro flows certain issues arise. One problem is the size-distribution of flows. Some flows are large and consist of a vast number of packets (elephants) others are small and may consist only of single packets (mice). The duration of flows might also be different, i.e. flows can last for milliseconds or they can last for days. Unless additional intelligence is applied, all flows are treated the same way by flow-based routing. Application of the flow-based networking paradigm in an IP networking environment has a number of advantages. These include: the ability of QoS provisioning on a flow level, e.g. guaranteed bandwidth, ATM-like and PSTN-like behaviour. Since all packets are routed on the same link, packets are strictly ordered and adverse affects to TCP are avoided. The flow base routing paradigm also allows for resource management and improved network utilisation. This includes also the possibility of fast failure recovery.

III. Flow routing in optical MPLS/GMPLS networks

Previous sections introduced flow-based routing, this section discusses a practical routing setup and suggests the application of flow base routing in an MPLS/GMPLS context. The proposal is based on the concepts of *Dynamic Alternative Routing* introduced in Section III-A and uses the network model presented in Section III-B. The scheme itself is discussed in Section III-C. The discussions assume that flow-based routing is possible and a method such as SAPOR is available.

A. Dynamic Alternative Routing

Dynamic call routing in circuit switched telephone networks has been widely used to improve performance and increase utilisation. *Dynamic Alternative Routing* (DAR) [14], proposed by British Telecom, is a call routing strategy that selects alternative path stochastically in case the original path is not available. DAR relies only on local information and is therefore robust and requires fewer network resources than centralised schemes.

DAR works as follows: A fully meshed network consists of n nodes and $n \cdot (n-1)/2$ links. Every link (i, j) has a capacity c_{ij} and a trunk reservation parameter assigned. Every *Origin-Destination* (OD) pair has an alternative tandem node k assigned, used by overflow traffic. During it's operation, traffic is routed via the directed connection, overflow traffic is sent via k. If k reached its maximum load, the call setup fails and an alternative tandem node is randomly selected out of the pool of nodes N without node i or node j. If additional overflow traffic has to be routed for this OD pair, it is routed via the new tandem node. Note, it is necessary that trunk reservation is applied to avoid instabilities in the routing. Several alternatives and advancements to this simple scheme have been proposed and are in use. To adopt DAR to IP networks, a number of steps are necessary and are discussed in the following sections.

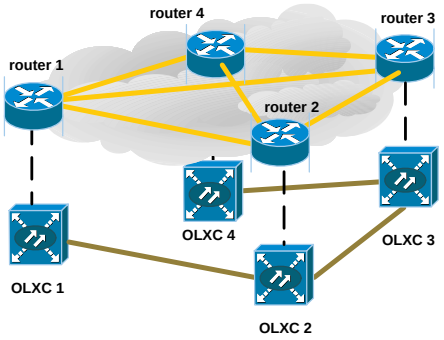

Fig. 1. Network Example

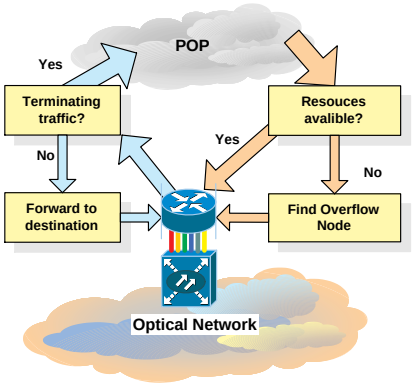

Fig. 2. Flow-based DAR - Scheme

B. Network Model

Currently, two routing and switching solutions for optical networks seem to be viable, i.e. hybrid and router only architectures. The hybrid architecture uses a network of connected *Optical Layer Cross-Connects* (OLXCs) which optically switch transit traffic (Wavelengths λ). Access traffic at the *Points of Presence* (POP) is handled by a local router. The optical layer has to manage traffic at fixed units, i.e. wavelength, since the optical layer is essentially circuit switched. In router-only architectures, all optical links are directly terminated at routers and switching is done by the router on a packet level. In the remainder of this paper, the hybrid architecture is assumed, i.e. DWDM technology and OLXCs in the core and routers at the edge of the core.

Light paths are established between the edge routers and identified by path labels. Means and methods of the path establishment are not the focus of this work. GMPLS based schemes can be used to establish the light paths/label paths. The number of wavelengths, assigned to a specific path, determines the capacity of these connections. This network is fully meshed between the edge nodes. Figure 1 depicts an example of such a network. The switching nodes symbolise WDM switches in the core and the routers symbolise edge nodes. The optical transport network consists of optical cross connects linked by backbone trunks. The routers are fully meshed by different wavelengths. Packets can be routed on these paths using the GMPLS protocol.

The resulting abstract model can be defined as follows: The logical router network has n nodes and is connected by $n \cdot (n-1)/2$ links of capacity $c_{i,j}$. The default routing scheme in such a network is straight forward: All traffic is sent on the direct, the one-hop, path. Alternative paths to the destination exist via all, but the destination node. Since this network is fully meshed, all intermediate nodes are connected to the destination. Therefore, a network with n nodes has $n-2$ alternative two hop-paths. Some of these paths will be more useful than others. The next sections discuss issues that are relevant to judge the optimality of paths and introduce a selection method.

C. Flow-based DAR

Figure 2 depicts a diagram that illustrates the operation of the Flow-based DAR scheme. It shows a single router that connects the POP to the optical network. Incoming flows that terminate at the POP are routed to the POP. Flows that do not terminate at this node belong to overflow traffic and have to be routed to their destination. Flows that exit the POP are routed to the destination, if sufficient resources are available. Otherwise, an overflow node is determined and the flows are routed to this intermediate node.

The major differences between the original circuit switched DAR network and packet switched IP networks is the networking paradigm. Since flows and calls are closely related, the scheme is formulated for a flow-based network. New flows that arrive at the router are routed via the one hop path to their destination. A function constantly monitors the packet count, flow count and utilisation for a router. If an upper threshold is reached or interfaces become unavailable and when new flows for this destination arrive, an alternative destination is requested from the overflow function. The method describing how this alternative node is selected is discussed in the next sub section.

It is assumed that every node has the means to inform an upstream node that it does not appreciate any more overflow traffic. This can be done by the use of an existing network management or signalling protocol. On receiving such a request, the node is taken out of the pool of possible overflow nodes. Trunk reservation requires that a node can distinguish between direct and overflow traffic. In packet switched networks, the packets can be distinguish if the network is fully meshed: i.e. terminating packets that are destined for another node within the same domain represent overflow traffic, as depicted in Figure 2. Trunk reservation can be implemented by a capacity margin, reserved for direct flows.

D. Two-Hop Path Classification

The selection of the overflow node is based on three factors: the current utilisation of the link connected to the overflow

node, a *Friend Factor* (FF) and the distance. The utilisation is locally available information, the friend factor reflects network feedback and the distance reproduces the network topology. A node is selected as a new *overflow node* based on the shortest distance and the condition that FF and utilisation are below the defined respective thresholds.

The node measures the load on its interfaces and calculates the utilisation. The origin node has knowledge of its next available hops. However, it does not know about the utilisation beyond the intermediate node, hence the friend factor is introduced. The FF is an integer value. It is initialised to the same value for all nodes. If a destination is rejected, its FF is set to the maximum FF of all nodes, plus one. Every time the FF is changed, the FF threshold also has to be adopted. The FF threshold influences the time that bad feedback is remembered. For example, the threshold has to be $FFmax - 1$, if rejection is to be remembered for one event only.

The topology on the logical layer does not correspond to the topology on the physical layer. Figure 1 depicts an example network on the physical and logical layer. To find the *shortest physical path* without knowing about the underlying network topology is a major problem. Generally, the path with the shorter round trip times are preferred over paths with longer round trip times. The distance can be estimated by the round trip time or is available as externally provided information. The list of nodes can be further narrowed, if information about the optical layer is available. If the list of paths has more than one entry, the overflow node is randomly chosen from the set of possible nodes.

E. Comments

This section outlined a flow-based dynamic alternative routing scheme. Practical applications must consider additional issues. Since microflows can have different sizes, the capacity margin for trunk reservation needs to take this into account. Research on trunk reservation in packet based context is required to avoid unstable and critical behaviour known from the circuit switched case.

All measured parameters change constantly. The polling time interval, used to measure and calculate the link utilisation should be short enough to capture relevant changes, but long enough so that processing does not cause a burden on the router's resources. The details of the protocol that informs its upstream nodes that the current node does not want any more traffic, has to be specified. Timing in this case is also relevant. Possible delays lead to inaccuracies of the traffic margin and timing at which the upstream node reacts to the message. This also needs to be included in capacity margins.

IV. Conclusion

This paper discussed issues of flow-based routing in an optical next generation network. It argued that flow-based networking is better related to consumer needs and allows differentiated treatment of flows in networks. It proposed a simple overflow routing methodology, allowing the efficient use of available resources and increases resilience to network failures.

Future work needs to address performance evaluation in relevant networks, so comparisons to existing solutions and estimates of expected savings can be provided. Other areas that need to be addressed include multi-class routing and detailed QoS considerations. Many extensions to this scheme are possible, in particular, when considering different traffic classes. Flow-based networking can enable a number of interesting network features and supports QoS related treatment of traffic, such as shaping and conditioning. A need for such features seems to emerge since IP technology begins to dominate networking technology in carrier grade and corporate networks.

V. Acknowledgements

The authors would like to thank the Australian Telecommunications Cooperative Research Centre (ATcrc) for their financial assistance of this work.

References

[1] G. R. Ash. *Dynamic Routing in Telecommunication Networks*. McGraw-Hill, 1997.

[2] J. Regnier, F. Bedard, J. Choquette, and A. Caron. Dynamically controlled routing in networks with non-DCR-compliant switches. *IEEE Communications Magazine*, pages 48 –52, July 1995.

[3] R.J. Gibbens, F.P. Kelly, and P.B. Key. Dynamic alternative routing - modelling and behaviour. *Proc 12th International Teletraffic Congress (ITC 12), Turin Italy*, 1988.

[4] K. Kawashima and A. Inoue. State- and time-dependent routing in the NTT network. *IEEE Communications Magazine*, pages 40–47, July 1995.

[5] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, and J. McManus. *Requirements for Traffic Engineering Over MPLS*. IETF, September 1999. RFC 2702.

[6] B. Fortz and M. Thorup. Optimizing OSPF/IS–IS weights in a changing world. *IEEE Journal on Selected Areas in Communications*, 20(4):756–767, May 2002.

[7] J. Murphy, R.J. Harris, and R. Nelson. Traffic engineering using OSPF weights and splitting ratios. *In Proceedings of Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002, Fremantle WA, October 13-16*, 2002.

[8] Caspian Networks. *Flow-State Routing: Rationale and Benefits*, July 2004. White Paper - www.caspiannetworks.com/files/Apeiro_Flow_State.pdf.

[9] D. Thaler and C. Hopps. *Multipath Issues in Unicast and Multicast Next-Hop Selection*. IETF, November 2000. RFC 2991.

[10] A.Elwalid, D.Mitra, I.Saniee, and I.Widjaja. Routing and protection in GMPLS networks: From shortest paths to optimized designs. *Journal of Lightwave Technology*, 21(11):2828–2838, November 2003.

[11] M. Ali, C. Assi, N. Ghani, and A. Shami. Integrated traffic grooming in converged data-optical networks. In *Proceedings of 9th IEEE Symposium on Computers and Communications (ISCC) , Alexandria, Egypt, June/July*, 2004.

[12] A.A. Kist and R.J. Harris. Scheme for alternative packet overflow routing (SAPOR). *In IEEE Workshop on High Performance Switching and Routing (HPSR 2003), Turin, Italy*, June 2003.

[13] A.A. Kist and R.J. Harris. Cost efficient overflow routing for outbound isp traffic. In *Proceedings of 9th IEEE Symposium on Computers and Communications (ISCC) , Alexandria, Egypt, June/July*, 2004.

[14] M.E. Steenstrup, editor. *Routing in Communications Networks*, chapter 2, pages 13–47. Prentice Hall, 1995.