Routing and fast protection in networks of long-reach PONs

D Hunter and T Gilfedder

This paper reports on proposed solutions to the recovery from faults in a generic class of networks, where customer access is achieved via long-reach passive optical networks (LR-PONs), with the majority of customers enjoying protection in the backhaul regime to two separate metro nodes. Initial modelling studies suggest that in the event of a cable failure or single equipment element failure, redirected data will almost always leave the transmitting node in under 50 ms. For more catastrophic failures (such as router failure or loss of a metro node), recovery might take between 100—200 ms. Reachability information for each LR-PON is discovered over each area of the network, and used to inform the redirection of traffic via tunnels. The scheme uses IP signalling to enable traffic re-routing, although the underlying services may be of any type (e.g. private line) — making the scheme separate from the service, customer or provider.

1. Introduction

Enhanced customer experience for telecommunications services is founded on the underlying reliability of the network over which these services are provided. At the lowest (physical) layer of the network, connections between the local exchange and inner core nodes (termed metro nodes in BT's 21C network) are invariably protected via one or more rings or protected chains. These protected links provide sub-50 ms protection in the event of equipment or fibre failure and hence allow all customers to enjoy good overall service availability. Indeed, some customers are served by two independent local exchanges, thus even failures in the access network can be alleviated.

It has been reported elsewhere [1] that the deployment of long-reach PONs (see Fig 1) could offer significant capital and operational savings, by effectively merging backhaul and access technologies into a single transport solution. However, in so doing, this approach must meet (or preferably exceed) all the inherent resilience capabilities and options currently available to customers. At the very least, any solution must be able to offer the following:

- sub-50 ms protection for all customers from failures in the backhaul network (e.g. cable being dug up),
- fast and automatic restoration (hundreds of milliseconds to a few seconds) of all services in the event of a major failure of a metro node,
- complete end-to-end service separation for selected customers.

Being able to redirect traffic from the customer to a back-up metro node via a separate path represents only half of the problem in re-establishing services on an end-to-end basis. All traffic originating at other parts of the network must be made aware of the failure within the PON system (or even the whole metro node in which it is located), re-address the traffic appropriately and re-route this traffic correctly. This process takes a significant time with existing protocols and methodologies. Therefore, it is this latter issue that is addressed here, permitting significantly faster switching times for all customers using dual-parented long-reach PONs.

The ITU-T [2] specifies resilience provisions which are perfectly adequate for gigabit passive optical networks (GPONs), permitting recovery from broken cables or equipment faults. The objective of the work described here is to extend the resilience options of GPON to the point where they can be used to implement dual-parented LR-PONs. In Fig 1, each customer can access only one exchange — in this study, networking configurations are evaluated which permit a customer to access another metro node if the current one fails, thus providing enhanced fault tolerance and superior service to the customer. The solution achieves fast recovery from a range of faults, and does not carry out restoration switching in the physical layer.

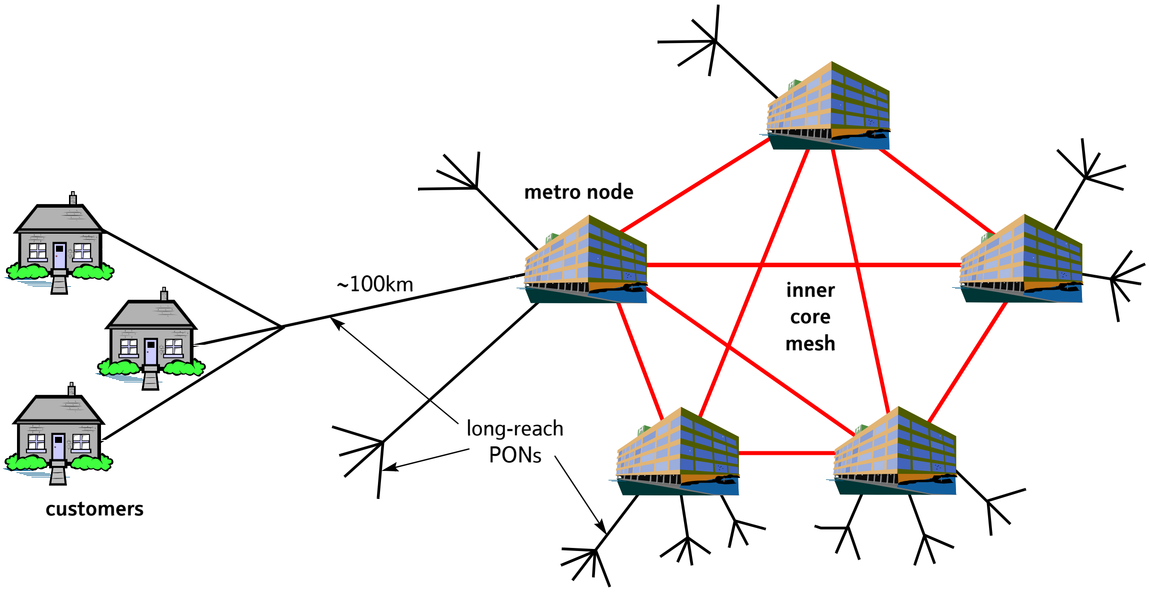

Fig 1 LR-PON networking concept.

To address these issues, this paper introduces two new concepts [3]. Firstly, a re-routing technique called LATTE (label and address tunnelling via tables at the edge) redirects traffic to a back-up LR-PON optical line termination (OLT) in a second metro node via a tunnel if the primary LR-PON OLT cannot be reached (either due to link failure within the PON itself, equipment failure of the PON OLT, or metro node failure due to fire or flood, etc). Secondly, a simple LR-PON OLT discovery protocol, optimised for speed, called FROTH (fast recovery for OLTs via transmission of hellos) collects LR-PON OLT reachability information which is fed into LATTE.

Figure 2 shows dual parenting of two distinct customer types — those that require complete circuit separation between their premises and the metro node (termed commercial customers in this paper), and those that require protection only in the backhaul section of the network (termed residential customers). Consider the effect of link **L1** failing. The commercial customer merely redirects traffic on to the secondary link (load balancing may also have been employed by the customer using the lower LR-PON) which leads to the lower LR-PON OLT ('OLT(S2)'), whereas the residential customers now must have their traffic redirected to the secondary metro node via a back-up LR-PON OLT ('OLT(S1)'). During normal operation, all primary LR-PONs ('OLT(P)') are permanently connected with ranging having taken place as normal in both cases. However, in the event of a failure in **L1**, the protection OLT that serves residential customers ('OLT(S1)') must activate and range prior to traffic being restored (although full re-ranging may not be required as some ranging parameters may already be available or previously stored). The protection OLT ('OLT(S1)') cannot be fully activated as its ranging and status messages would interfere with the primary PON, although in principle it could monitor upstream traffic from customer located optical network units (ONUs). Multiple operators collocated at metro nodes can deploy their own LATTE and FROTH protection capabilities independently from each other, thus preserving their own network integrity. Similar capabilities would be required at network-to-network interfaces such as peering points.

2. Label and address tunnelling with tables at the edge (LATTE)

When re-routing IP traffic, LATTE places IP datagrams within tunnels (IP-in-IP) as appropriate when a failure is reported by FROTH, so that they are redirected to the appropriate back-up LR-PON OLT. LATTE carries out a similar operation when necessary on paths for MPLS and ATM, and on other technologies. Redirection via tunnels takes place if necessary whenever a packet enters or leaves either the core network, or an 'area' of the network.

2.1 Allocation of IP addresses to customers

LATTE requires information about IP addresses, and the availability of LR-PONs throughout the network, which is provided by FROTH. Each LR-PON is assigned two prefixes, or IP address ranges, which are each of a size defined by the service provider, and indicate all permissible IP addresses for either commercial customers or residential customers. This is because these two customer groups are treated differently in the event of a fault. With commercial customers, an addressing scheme is adopted whereby the address on the back-up ONU interface can be deduced from the address on the interface of the primary ONU.

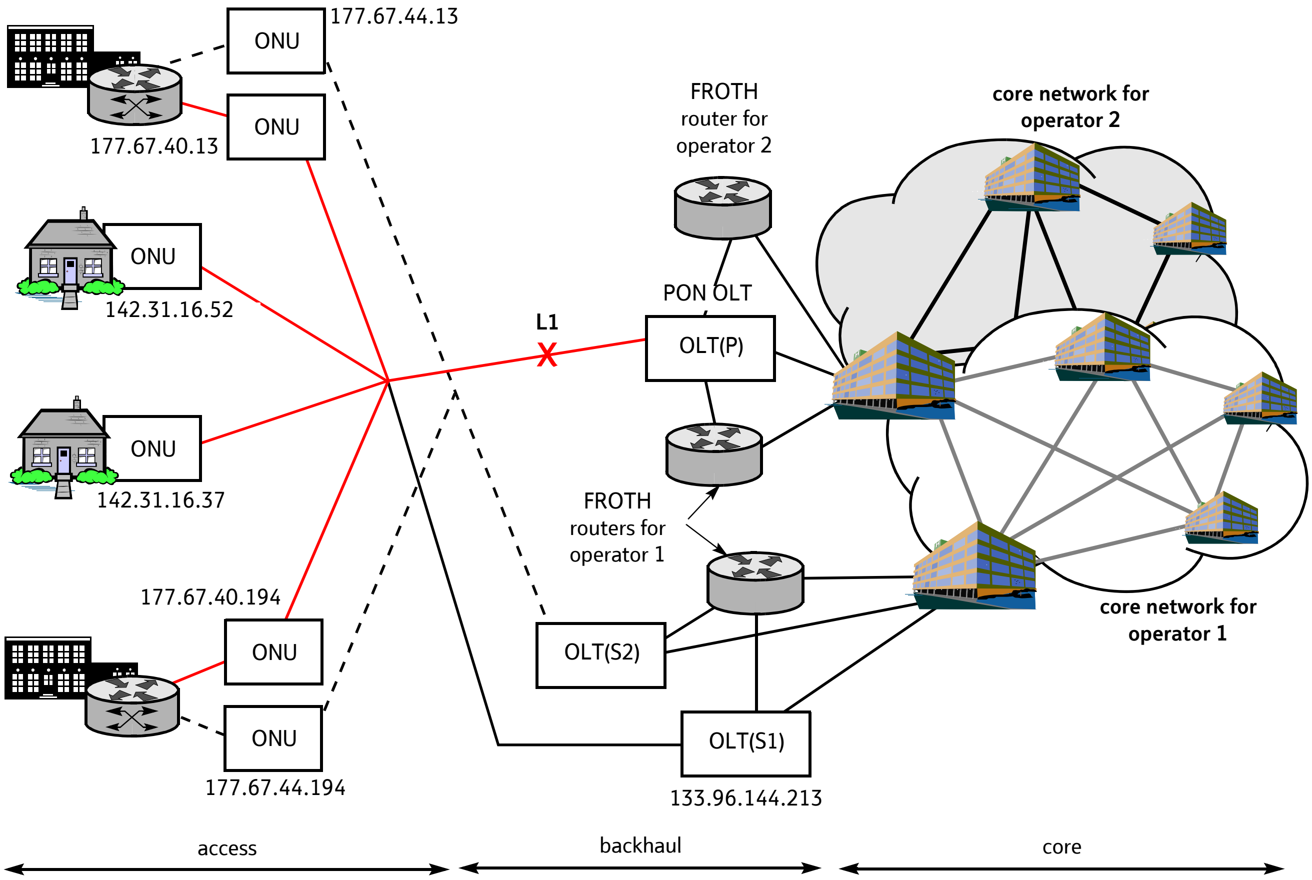

Fig 2 Example network configuration to illustrate operation of LATTE. All customers are protected from cable failures in the backhaul regime with some (predominantly commercial) customers enjoying full end-to-end protection. The first two rows of Table 1 correspond to the IP addressing scheme applied. Note that multiple operators can deploy their own LATTE and FROTH capabiliites or implement different protection schemes independently from each other.

2.2 Re-routing in the event of failure

Table 1 is a sample routing table generated by FROTH, which is used by LATTE to make IP re-routing decisions. Normally, both the primary and secondary OLTs are working, so no re-routing takes place — an IP datagram addressed to the primary LR-PON OLT is indeed routed to that LR-PON OLT. The first two columns in the table indicate a prefix associated with the primary LR-PON and the number of IP addresses that exist within this prefix. The second two columns indicate whether the primary and/or the secondary LR-PONs are working and available. For commercial customers, the secondary address can be easily derived from the primary address as the final bit in the network address field is altered from ‘0’ to ‘1’ within the IP address. In this case, if the address of an incoming IP datagram falls within the range of 1024 IP addresses defined by the prefix 177.67.40.0/22 (row 1)[1], it is re-directed by an IP-in-IP tunnel [4] to the corresponding new address, which can be deduced from the old address, due the addressing scheme chosen. With IP-in-IP, an IP datagram to be re-routed is placed in the payload of another IP datagram. For a commercial customer, the address on the secondary LR-PON is its destination, without the source being aware of this. This is illustrated in Fig 2, where the customer’s primary IP address is 177.67.40.13, and the secondary IP address is 177.67.44.13 (bit 22 changed from 0 to 1). The total number of customers that require full end-to-end equipment diversity and circuit routing separation (and hence enjoying the above protection scheme) will be considerably smaller than the number of residential customers, and therefore the size of prefix for such commercial customers will be correspondingly smaller than that for residential customers.

The underlying issue of myriad traffic streams originating at multiple points in a network needing to be routed to primary and backup locations is generic to any underlying service (IP, Ethernet, TDM, etc). With this approach, although signalling within FROTH would still be carried out via IP, different platforms could be adapted so as to take advantage of this scheme. For example, another table similar to Table 1 could contain MPLS LSP identifiers, and MPLS packets could be tunnelled within another MPLS LSP, instead of using IP-

[1] Note that the /22 indicates that the first 22 bits of the IP address represent the network address and the remaining 10 bits correspond to the host id, and hence 2^{10} = 1024 IP addresses are defined by this prefix.

Table 1 An example of a routing table supplied to LATTE by FROTH for re-routing IP datagrams. The first two rows are illustrated in Fig 2. For commercial customers, the prefix for the secondary LR-PON can be easily deduced from that of the primary LR-PON by setting the rightmost bit of the network ID in the IP address to 1. The use of prefixes in this way avoids the need for individual IP addresses in the table, reducing its size and accelerating table look-up.

Primary prefix	Maximum number of IP addresses in prefix	Primary working	Secondary working	OLT address for residential customers	Manual re-route
177.67.40.0/22	1024	no	yes	0.0.0.0	no
142.31.16.0/25	128	no	yes	133.96.144.213	no
152.52.4.0/24	256	yes	yes	0.0.0.0	yes
etc	etc	etc	etc	etc	etc

in-IP tunnelling as before. Additionally, re-routing of some time division multiplexed (TDM) technologies, such as private lines, would require an interface between LATTE and the SDH management system, but could avoid duplication of bandwidth to provide protection across the inner core network.

Residential customers will not have paired (primary and secondary) IP addresses, hence it is difficult to route datagrams to it via two different OLTs. To ensure that each datagram is re-routed correctly, the tunnel must terminate at the secondary OLT, which then de-capsulates the IP datagram and forwards it to the customer. The secondary OLT is connected to a different metro node than the primary OLT (Fig 2) and, besides being capable of de-capsulating IP-in-IP, it must have its own IP address on the interface to the metro node. In Fig 2, the allocated IP address for the lower OLT is arbitrarily set at 133.96.144.213 in this example. Although performing a manual revert when the main link has been repaired would reduce the amount of traffic in the core network (no IP-in-IP encapsulation), this may have a negative customer impact and so would be avoided.

As noted earlier, the secondary OLT associated with residential customers should re-range prior to receiving the first redirected data packet from the source. It is instructed to do so either by a notification packet received from the FROTH router beside the fault, or via the usual FROTH time-out mechanism. If the diverted data packets arrive before the OLT has re-ranged, they must be buffered (or dropped if time to re-range is excessive) by the local FROTH router until the OLT is ready for operation.

The last column of the table is labelled ‘manual re-route’ and is configured manually by the network administrator. This allows re-routing to take place if necessary without any fault, e.g. for network maintenance. Although LATTE could be incorporated into the OLT, it is probably better, however, to locate it in the multiplexer feeding the OLT signals into the metro node, permitting economies of scale to be realised.

3. Fast recovery for OLTs via transmission of hellos (FROTH)

FROTH distributes reachability information about each cluster of LR-PONs around the network area, where a ‘cluster’ constitutes all the LR-PONs connected to a particular metro node. To avoid any complications with respect to vendor support, and to make this analysis as general as possible, it shall be assumed that FROTH runs on a separate computer called the FROTH router, which is physically located beside the corresponding metro node, but will be of insignificant cost by comparison. This avoids the need to alter the metro node router or its software, or to introduce further complexity into the OLTs.

FROTH uses a much simpler signalling method than, for example, OSPF or SDH [5], and implements simpler computation at each node. This is because it carries signalling information over the existing core network while avoiding complex processing in intermediate routers along the path, in order to reduce signalling time and hence convergence time. The FROTH router is informed of OLT or LR-PON failure via a local area network within the metro node, and forwards this information via IP status packets to all the other FROTH routers in the network. Status packets may also contain any relevant IP addressing information for LR-PONs in the cluster. Two classes of failure are considered — failure of the LR-PON link (or path) or LR-PON OLT, and failure of the metro node or FROTH router.

3.1 Operation of FROTH

Signalling in FROTH takes place via IP, protecting not only IP, but potentially a range of other technologies. For LATTE to function, each FROTH router must have global knowledge, over the area or whole network, about which other LR-PONs are reachable, and what IP address prefix is associated with commercial and residential customers in each of them. If another LR-PON OLT is unreachable, it could be due to the failure of, for example, an LR-PON path, its OLT, or the associated metro node or FROTH router. Based upon this information, a FROTH router updates its routing table, which is passed on in modified form to LATTE, so that datagrams destined for the first choice destination LR-PON OLT are diverted to the corresponding second choice destination LR-PON OLT, if their first choice is unavailable.

3.2 Status packets

Each status packet contains a number of records, as well as overhead information. The latter includes a text password, to address security issues. Each record relates to one OLT or a metro node and its associated FROTH router. Each FROTH router sends a status packet to every other FROTH router at fixed intervals via the core network, which reports on one OLT at a time (taken in turn) and also the metro node and FROTH router. Also, a status packet is sent out immediately when an OLT fails or is repaired, and it is hence necessary to inform all other FROTH routers. A FROTH router can only originate a new record in this way (rather than relaying existing information) when changes occur adjacent to its own metro node. Unlike more catastrophic failures, this does not require a time-out, since all other FROTH routers are informed as quickly as possible. Hence each FROTH router knows which other OLTs can be reached and which cannot — a requirement for LATTE to operate. Each FROTH router forwards old records about all other LR-PONs anywhere in the area, but only if the associated timestamp indicates that these records supersede its existing data.

When a FROTH router receives a status packet it checks to see if the information is more recent than its existing record on each device represented. If so, it updates this record and immediately broadcasts another asynchronous status message containing the new information to all other FROTH routers. This is crucial to the robustness of the protocol, ensuring that even if a link or router in the core has failed, the information will still reach its destination via another route.

The impact of this on performance is captured in the modelling work later.

FROTH must not be triggered in error by packet losses, either due to protection events or congestion within the core network. To avoid this, information is flooded around all FROTH routers, as outlined above. FROTH routers forward any new updated information they acquire in this way to all other FROTH routers. Hence, regardless of packet loss, reachability information about any particular cluster will still reach every other FROTH router, although perhaps by an indirect route, except in pathological cases of network failure. Furthermore, it will be shown that this scheme has the desirable advantage of speeding up the propagation of signalling information between FROTH routers.

4. Performance and timing evaluation of FROTH

4.1 Recovery time

The performance and the scalability of FROTH have been evaluated, with respect to both fault recovery time and signalling traffic level. The recovery time is modelled by combining a number of delay components as appropriate, to model several different recovery scenarios. These include reporting the fault, updating a FROTH routing table, signalling from FROTH to LATTE, or re-ranging operations on an LR-PON. There is also a variable delay term, modelled via statistical measurements from the Internet, representing the transit time of an IP datagram from one point in the network to another. Existing measurements of TCP round trip times (RTTs) provide crucial and useful insights into the statistical distribution of propagation delays in any IP network. Based upon reported measurements [6], the propagation delay along a path through the network is modelled as a random variable whose tail is Pareto distributed, while the remainder of the probability density function is Gamma distributed.

Four principal scenarios are modelled, each with different timing (see Fig 2). In each there can either be failure of an LR-PON, OLT or backhaul connection, or there can be failure of a metro node or FROTH router. Also, there is the choice between residential customers and commercial customers. As noted earlier a residential customer cannot have service restored if the access fibre fails.

The recovery time is evaluated by the model, being defined as the time from the primary destination LR-PON becoming unreachable, to the first packet of re-directed user data arriving at the back-up destination LR-PON OLT.

Suppose that traffic is destined to a customer on one LR-PON primary OLT but the main link fails, hence the traffic must be switched over to the secondary or back-up LR-PON OLT (Fig 2). Table 2 summarises the results. The first column indicates the type of failure that is under examination — 'LR-PON' is for cable failure or other failures for which the FROTH router can transmit an appropriate failure message, whereas 'metro' indicates failures for which a time-out is necessary. The second column represents the classification of customer that is under consideration. The third column indicates the length of time required to pass before a 'time-out' is detected, hence initiating fault recovery. The fourth column indicates the target recovery time for failures, and the remaining columns indicate the probabilities that traffic can be restored to a secondary node within such a time given the mean transmission delay of the network (indicated in milliseconds).

Table 2 Summary of results. The last three columns provide the probability that traffic would be restored within the target (T), for mean network path delays of 5, 10 and 15 ms.

				Proportion of traffic restored within target (T)		
Failure type	Commercial/residential	Time-out interval	Delay target (T)	Network mean delay = 5 ms	Network mean delay = 10 ms	Network mean delay = 15 ms
LR-PON	Commercial	N/A	50 ms	0.989	0.975	0.962
LR-PON	Residential	N/A	50 ms	0.988	0.974	0.961
Metro	both	50 ms	100 ms	0.998	0.994	0.992
Metro	both	50 ms	150 ms	0.9997	0.997	0.995
Metro	both	18 ms	50 ms	0.998	0.994	0.992
false time-out		18 ms	N/A	8 nines	6 nines	6 nines

The first and second rows indicate that between 96% and 99% of traffic could be restored within a target of 50 ms (dependent on size of core network) given a failure for which the FROTH router can transmit an appropriate message. Additional analysis has indicated that the time taken for traffic to start leaving the far end correctly addressed to the back-up LR-PON OLT, having received the signal to redirect the traffic, occurs extremely quickly. This implies that much of the delay in re-establishing traffic is due to the data traffic traversing the network as opposed to any signalling delay. It is possible that the statistical density function of network transmission delay chosen here exacerbates this delay, and a fuller study, considering a variety of further experimental data would be advisable to understand the sensitivity of this effect. The model of metro node failure (rows 3, 4 and 5) shows that recovery from failure is slower in this case, but, as expected, the time between transmissions of status messages may be reduced (within the limits of feasibility), with a corresponding reduction in time-out interval, to improve performance. A time-out interval of 18 ms, for example, is expected to allow sufficient time for service recovery to complete within 50 ms. Finally the last row highlights the fact that ‘false’ time-outs can occur (albeit with low probability) and that this is dependent on the time-out limit and the mean delay through the network for status messages.

4.2 Traffic level due to status packets

The level of signalling traffic on the network also influences scalability. Suppose that all packets are 496 bits long, that all records within them are 6 bytes long, status packets are transmitted every 15 ms (IP multicast), the number of OLTs per metro node is 200, the number of metro nodes is 100, and a core link carries data at 60 Gbit/s (6 × 10 Gbit/s). With these figures, it can be shown that status packets occupy 0.02% of the total available network bandwidth. If they are now transmitted every 5 ms, the signalling traffic approximately triples, and 0.06% of the total available network capacity is now used for signalling. This may be worthwhile, in order to enhance recovery time when a metro node or FROTH router fails.

4.3 Use of area

It is easily shown that FROTH as it stands does not scale to very large networks. As the number of nodes increases, the amount of signalling traffic generated by the protocol becomes impracticably large. To ensure scalability, the network may be divided into areas, although these need not correspond to OSPF areas. LATTE is implemented on each link at the boundaries between areas, or between an area and the international peering points — in fact whenever a signal enters an area.

It has been assumed that the full network of 100 metro nodes constitutes one area, the worst possible scenario. It can be shown that the amount of signalling traffic is $O(N^2)$. If the network of 100 nodes were divided into 4 areas of 25 nodes, then the amount of signalling traffic would be reduced by a factor of approximately 16. Generally speaking, the propagation delay through a network is $O(\sqrt{N})$, so while propagation delays through each area would on average be roughly half those through the whole network (with a corresponding decrease in recovery time), the principal improvement would be in reducing signalling traffic.

5. Conclusions

If BT is to realise the vision of a simplified and low-cost network architecture, consisting of long-reach PONs linking customers directly to a core of 100 or so switching and intelligence centres (metro nodes), then the ability to provide levels of resilience and protection as good or better than those currently enjoyed is essential. Conventional techniques for protecting dual-parented traffic on to two separate nodes can be slow to re-converge after failures such as cable breaks, thus potentially leading to poor customer satisfaction. Clearly new approaches are required.

This paper has described two new techniques — LATTE and FROTH — that can re-route traffic in the event of a variety of failure scenarios, thus ensuring that traffic originating from elsewhere in the network is able to reach the correct destination. Calculations have shown that for LR-PON path and OLT failure, it is realistic to expect redirected data to leave the transmitting node well within 50 ms. Upon restoration, transmitting user data to the back-up LR-PON limits the overall speed of the process. For more catastrophic failures (i.e. metro node and FROTH router failure), recovery may take 100 — 200 ms, depending on how frequently each FROTH router transmits status messages. Essentially there is a trade-off. If status messages are transmitted more frequently, there is a higher network load due to signalling messages, but recovery from metro node or FROTH router failure is faster. The speed of recovery from LR-PON path or OLT failure is not affected by how frequently status messages are transmitted, since an asynchronous status message is sent immediately if an LR-PON or OLT fails.

This paper has assumed that messages are sent in-band (i.e. through the same network as the data traffic), but additional work to understand the benefits and issues surrounding the use of out-of-band signalling is also required. Benefits could include the enhanced security of signalling protocols against attacks, guarantees on recovery time, and could also effectively eliminate the possibility of false time-outs due to network congestion.

These are early proposals identifying robust and scalable mechanisms to meet the reliability requirements of future communications services. Alternative solutions may well arise during the course of further investigation into new protection mechanisms; nevertheless, this paper underlines the importance that the issue of resilience presents when considering the evolution of future communication networks towards long-reach PON deployments.

References

1 Payne D B and Davey R P: ‘The Future of Fibre Access Systems?’, BT Technol J, 20, No 4, pp 104—114 (October 2002).

2 ITU-T: ‘Gigabit-capable Passive Optical Networks (GPON)’, Recommendation G.984.1,2,3,4.

3 Hunter D: ‘Routing in Networks with Dual Parented Customers’, Internal BT Technical Report (June 2006).

4 Perkins C: ‘IP Encapsulation within IP’, RFC 2003 (October 1996).

5 Vasseur J-P, Pickavet M and Demeester P: ‘Network Recovery — Protection and Restoration of Optical, SONET-SDH, IP and MPLS’, Morgan-Kaufmann (2004).

6 Corlett A, Pullin D and Sargood S: ‘Statistics of One-way Internet Packet Delays’, IETF Internet Draft, draft-corlett-statistics-of-packet-delays-00 (2002).

David Hunter is a Reader in the Department of Electronic Systems Engineering at the University of Essex. In 1987, he obtained a first class honours BEng in Electronics and Microprocessor Engineering from the University of Strathclyde, and a PhD from the same university in 1991 for research on optical TDM switch architectures. After that, he remained at Strathclyde to research optical networking and optical packet switching, holding an EPSRC Advanced Fellowship from 1995 to 2000. After spending a year as a Senior Researcher in Marconi Labs Cambridge, he moved to the University of Essex in August 2002, where his teaching concentrates on TCP/IP, computer networks and network performance evaluation. He has authored or co-authored over 100 publications, and has acted as an external PhD examiner for the Universities of Cambridge, London and Essex. From 1999 until 2003 he was an Associate Editor for the IEEE Transactions on Communications, and he has been an Associate Editor for the IEEE/OSA Journal of Lightware Technology since 2001. He participated in editing a special issue of that journal, on Optical Networks, that was published in December 2000. He is a Senior Member of the IEEE and a Professional Member of the ACM.

Tim Gilfedder joined BT in 1997 following his PhD from the University of Strathclyde in Glasgow, UK.

He was involved with the deployment of the first commercial DWDM systems into BT’s core network, from the initial tactical planning and business case analysis through to their subsequent in-life support. He has since managed projects supporting the international backhaul and trans-border Pan-European networks, as well as providing technical support and consultancy to other parts of the business associated with optical networking. His current role is investigating disruptive optical technologies and their potential benefits for future network architectures and strategies.

He is a member of the Institute of Physics.