OSA Journal of Optical Networking, vol. 6, no. 5, May 2007, pp535-549.

Protection of Long-Reach PON Traffic through Router Database Synchronization

David K. Hunter and Zheng Lu

University of Essex, Department of Electronic Systems Engineering, Colchester CO4 3SQ, UK

dkhunter@essex.ac.uk, zlu@essex.ac.uk

Tim H. Gilfedder

British Telecom, OP7/2 Antares Building, Adastral Park, Martlesham Heath, Ipswich IP3 5RE, UK

tim.gilfedder@bt.com

Abstract

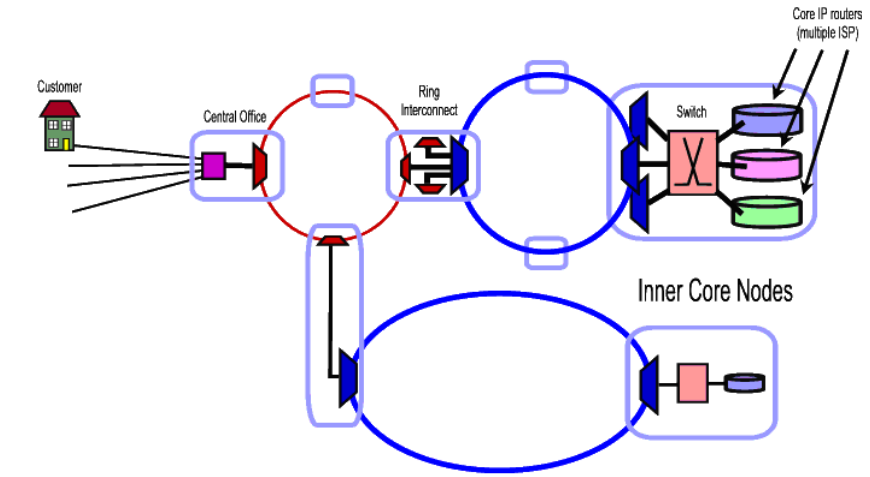
A resilience strategy is introduced for networks implementing dual homing (dual parenting) of customers, specifically those employing Long-Reach PONs (LR-PONs). Assuming that some mechanism exists to detect network element failures, the discussion concentrates on protocols that propagate information about customer reachability and how this information reroutes traffic in the event of a fault. Each router holds a database indicating which other routers, and which LR-PONs are available within the network, and these are synchronized between routers using IP. This information is used to reroute traffic in the event of failure. Simulation and analysis show that signalling time lies well within 50 ms, leaving sufficient time for redirection of user traffic.

1. Introduction

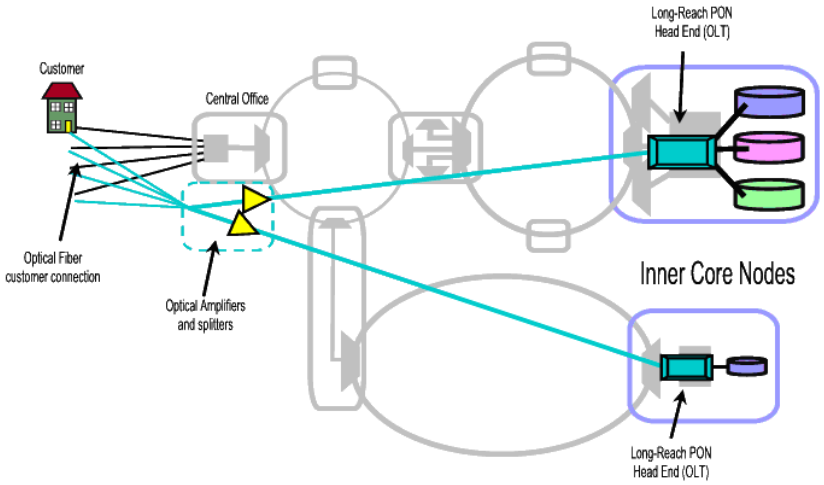
Enhanced customer experience of telecommunications services is founded on the underlying reliability of the network over which these services are provided. At the lowest (physical) layer of the network, customers are connected to the telecommunications provider via a copper (twisted pair) connection to the local exchange, also referred to as the Central Office (CO). Such connections are dedicated to the customer and are predominantly unprotected. The connections between the local exchange and the inner core network tend to be provided by fibre optic systems that aggregate the customer traffic onto larger transmission circuits and transport these circuits in protected fashion, normally via ring architectures, to inner core nodes. In this way, although a failure of the link between the customer and exchange would result in a loss of connectivity until such time the failure had been rectified, a single failure of the circuits between the exchange and the inner core network would result in a fast switchover (under 50 ms) to a standby route. Such architectures, common to many incumbent telecommunication network operators, provide customers with good overall customer service reliability. Indeed, some, usually business, customers are served by two independent local exchanges, such that single failures in the access network do not result in complete loss of traffic.

The use of fibre-optic technology in the access network is increasingly being adopted by network operators in order to provide high-speed services to customers. One of

the favoured architectures used is the passive optical network (PON) in which a number of customers (usually 32) are connected via passive optical splitters/couplers onto a single optical line termination (OLT) in the local exchange. Clearly, the backhaul network between the exchange and the inner core nodes remains largely unchanged from that of the copper-based access network architecture. However, it has been reported elsewhere [1] that the deployment of long-reach PONs could offer significant capital and operational savings by effectively merging backhaul and access technologies into a single transport solution. Extensions and enhancements to existing PON technology so as to offer higher-speed services to customers using radically reduced and simplified network architectures have been the subject of investigation by both operators and vendors [1, 2, 3], and are the subject of investigations within international standards bodies [4].



(a)



(b)

Fig. 1. Schematic highlighting the pertinent differences between conventional backhaul methodologies and those of Long-Reach PON technologies. (a) Existing backhaul methodologies consist of interconnected ring architectures (usually SDH or SONET) between the CO and inner core nodes for interconnect to alternative platforms (such as the core IP backbone) and alternative service providers (Internet or otherwise). (b) LR-PONs provide two direct links from the local area to the inner core.

Figure 1 provides a schematic view of the salient differences between a conventional ring-based backhaul solution and a LR-PON architecture. The conventional backhaul of Fig. 1(a) uses SONET/SDH rings to aggregate the traffic from multiple COs back to the inner core, where the traffic is switched and directed to the appropriate service provider and/or platform (such as the IP backbone network). Protection can be provided in a variety of ways, for example sub-network circuit protection (SNCP) or bi-directional line switched rings (BLSR), but ultimately result in sub-50 ms switching to a standby path in the event of a fibre cut or intermediate equipment failure. In general, the protection scheme is not able to provide such fast protection in the event of a major failure of the inner core node that requires the redirection and reconfiguration of traffic to an alternative inner core node for re-connection into the inner core. Furthermore, given that a single CO may serve thousands of customers and the traffic of many such COs may be aggregated onto a single OC-192/STM-64 link to the inner core node, such a methodology significantly constrains the throughput of the inherent services. For example, four COs serving 5000 customers each on a 10 Gbit/s ring can only offer an uncontended rate of 500 kbits/s per end-customer despite individual DSL connections being able to support many megabits per second. Clearly the situation becomes more acute as the number of customers increases and/or the bandwidth that the end customers require increases. Upgrading the current structures with increasingly powerful systems may alleviate the problem to a degree, but the increasing cost of such systems is not expected to match the level of revenue that would be generated from higher-speed services leading to reduced earnings for operators [1].

LR-PON architectures, as shown in Fig. 1(b), bypass all intermediate nodes and structures and provide services directly to end customers. Optical amplifiers or regenerators overcome losses inherent in large splits and long backhaul distances. Due to their reduced size and power requirements they may even be located outside the CO, thus offering the potential of building evacuation and possible closure, resulting in significant operational savings. Such a solution can cater to both residential and businesses (provided that the businesses do not warrant a direct fibre due to the amount of bandwidth required). Businesses that require dual connections can be accessed via two LR-PONs, each terminating in separate inner core nodes.

LR-PONs, although not yet commercially available, do appear technically viable. Current Gigabit-capable PON (GPON) protocols have been shown to be supported (via regeneration) over 100 km [2] and up to 60 km via optical amplification [3]. As a result it is possible to conceive LR-PONs being available; offering the ability to serve as many as 1000 customers over a distance of 100 km. However, such a system would suffer more failures (especially cable and fibre failures) than those of the short-reach access GPON systems, and would also affect many more customers. As a result, in order to be widely adopted, the LR-PON solution must provide protection capability that meets (or preferably exceeds) all the inherent resilience capabilities and options currently available to customers between the customer premises and the inner core node. 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 dig-ups between the exchange and inner core node).
- Fast and automatic restoration of all services in the event of a major failure at the inner core node (e.g., fire, flood, extended power loss etc.).
- Complete end-to-end service separation for selected customers.

Being able to reroute traffic from the customer to a secondary inner core node over a separate path is only half of the solution to re-establishing services on an end-to-end basis. All traffic originating elsewhere in the network must be made aware of the failure within the LR-PON system, or the inner core node, and then readdressed and rerouted correctly. This process takes a significant time with existing protocols and methodologies. Therefore, it is this critical issue that is addressed here, with a possible solution proposed that permits significantly faster recovery times for all customers using dual-homed LR-PONs.

Although LR-PONs are assumed throughout, this analysis is applicable to any solution that aims to connect customers directly to inner core nodes, thus merging the distinction between access and backhaul regimes. This paper considers protection of a number of protocols, but concentrates on IP (Internet Protocol), which can carry user voice, video or data. The core network nodes employ standardized protocols which cannot easily be changed; to address this issue, our new protocols run on separate computers, with one adjacent to each inner core node. Hence it is not necessary to modify software running on existing hardware in the core. For the sake of brevity and because many of the new high-speed services are expected to be served by inner core backbone network of IP routers, hereafter the term "router" will be used to refer to the equipment located in the inner core node.

The ITU-T [5, 6, 7, 8] specifies resilience provisions which are perfectly adequate for GPONs, permitting recovery from broken cables and certain equipment faults. This work extends GPON's resilience capabilities to the point where dual-homed LR-PONs are possible. Networking configurations are evaluated which permit a customer to access another router 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.

To address these issues, this paper employs two concepts. Firstly, traffic can be rerouted to a secondary LR-PON Optical Line Termination (OLT) in a second router if the primary LR-PON OLT is not available (either due to link failure within the LR-PON itself, equipment failure of the OLT, or router failure due to fire or flood etc.). Also, a protocol, optimized for speed, has been developed from existing concepts, to synchronize information held in router databases, which describe the status of LR-PONs and other routers.

Figure 2 illustrates the general principle, showing how diversity occurs. Resilience is implemented by diverting traffic entering the core network to an alternative (or backup) LR-PON whenever necessary. The core network may be any network implementing IP or any other protocol capable of carrying signalling information. In the example shown in the diagram, there are six access networks (U-Z), which may be implemented via PONs or LANs. Each customer is dual homed, i.e., connected to two access networks, each of which is interfaced to the core network via a different router. Although shown as independent links, most residential customers would share a common link in the "access" regime whereas the "backhaul" network would be diversely routed (see connectivity indicated in Fig. 3 as explained below), thus existing cable and duct topologies would remain largely unchanged. Consider how customers A and B communicate. Normally, customer B uses access network W, however, if it becomes unavailable due to a fault, for example a cable cut, access network X is used instead. All routers are informed which access networks are

functioning and accessible from the core network. Normally, A and B communicate via the "working" path through router 3, however, if W fails, then router 1 is informed of this, and reroutes traffic for B over the "protection" path to router 4, thus implementing recovery from the failure. The routes of the working and protection paths are defined by the usual IP forwarding mechanism. Re-routing traffic in this way at the edge simplifies its implementation, by minimizing the number of points in the network at which rerouting is required.

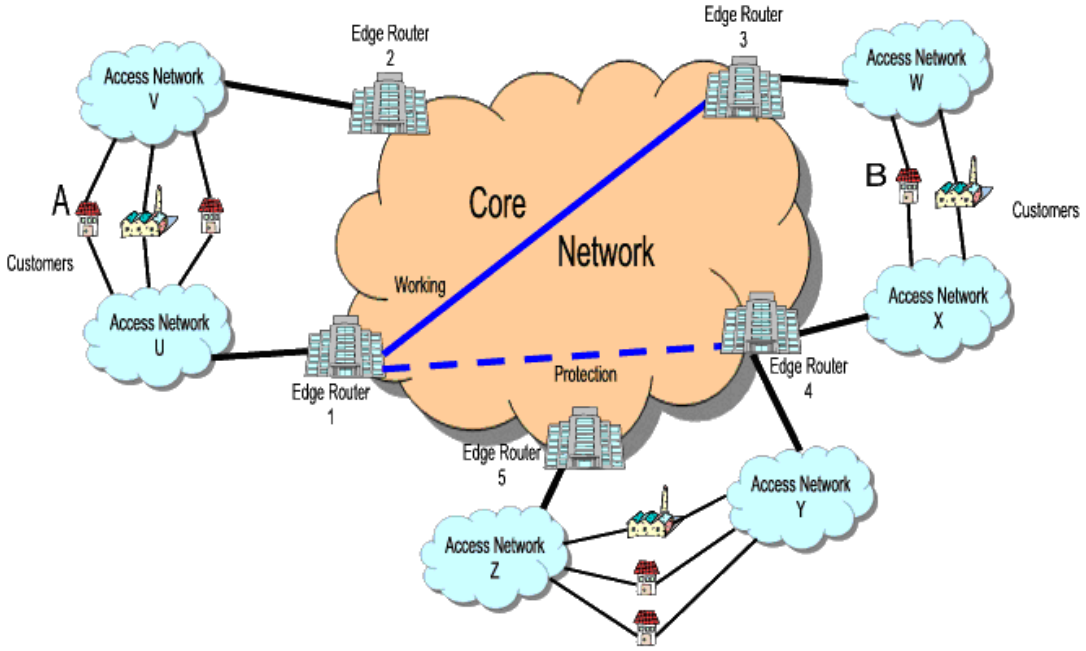


Fig. 2. A generic illustration of the dual-homing concept studied here. Routers 1-5 form an interface between the core network and the access networks U-Z. The access-backhaul networks may be represented by several technologies, with LR-PONs being of special interest. Customers A and B normally communicate via the working path.

Figure 3 shows dual homing of two distinct customer types: those that require complete circuit separation between their premises and the router (termed commercial customers in this paper) and those that require protection only in the backhaul (herein known as residential customers). The ability to cater for both customer types is a feature of this architecture. If link L1 fails in Fig. 3, a commercial customer merely reroutes traffic onto the secondary LR-PON marked "commercial protection". Before the failure, this customer may have employed load balancing by using both "working" and "customer protection" LR-PONs. The "commercial protection" LR-PON leads to OLT(S2), whereas residential customers must now reroute their traffic to the secondary router via their secondary LR-PON through OLT(S1).

During normal operation, a primary LR-PON is permanently connected through OLT(P1), with ranging having taken place as normal. However, if L1 fails, the protection OLT that serves residential customers through OLT(S1) must activate and range prior to traffic being restored. Full re-ranging may be unnecessary, as some ranging parameters may already be available, or may have been stored previously. OLT(S1) cannot be fully activated for protection as its ranging and status messages would interfere with the primary LR-PON, although in principle it could monitor upstream traffic from customer-located Optical Network Units (ONUs).

Multiple operators colocated at routers can deploy their own implementations of this scheme, with protection capabilities operating independently from one other, thus preserving their own network integrity. Similar capabilities would be required at network-network interfaces such as peering points.

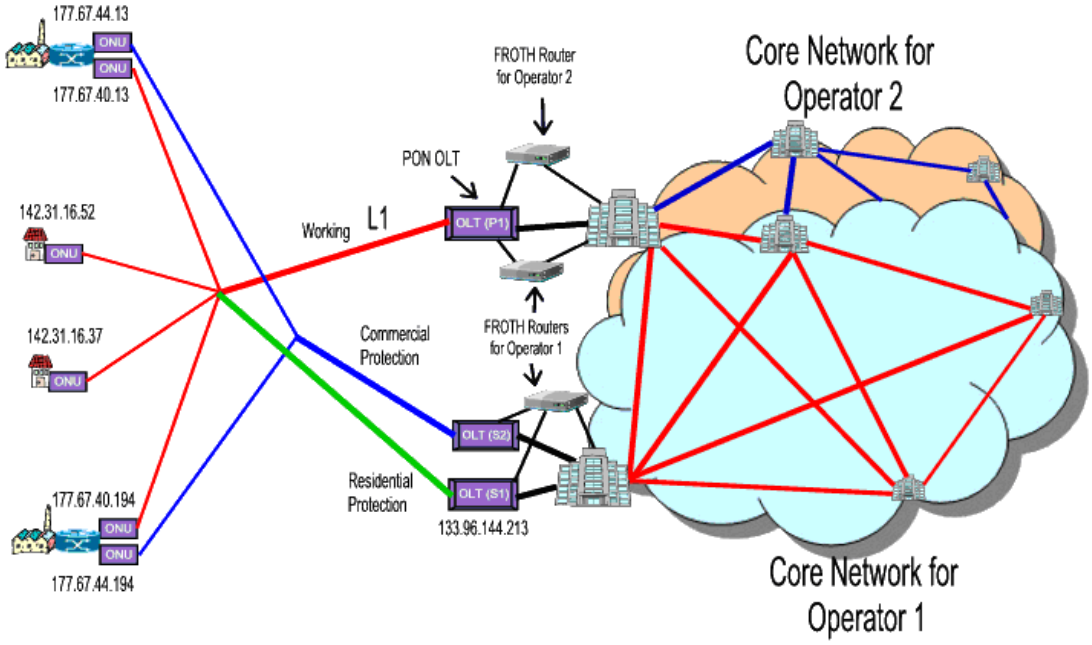


Fig. 3. Example network configuration to illustrate the operation of traffic redirection.

In the remainder of this paper, Section 2 shows how reachability information relating to LR-PONs and routers is synchronized across the network. Section 3 shows how this information may be used to protect IP traffic in the network layer, while Section 4 indicates how such dual homing could be implemented with other networking technologies. Section 5 studies the performance and scalability of the proposal via analytical modelling and simulation. Section 6 shows how multiple operators can coexist with the resilience scheme, while Section 7 concludes the paper.

2. Synchronization of router databases

Availability information about each cluster of LR-PONs, and about each router, is distributed around the network area, where a "cluster" constitutes all the LR-PONs connected to a particular router. To avoid any complications with vendor support, the protocol performing this synchronization task should run on a separate computer, which would be physically located beside the corresponding router, but would be relatively inexpensive. This would avoid the need to modify the router or its software. However, to simplify the terminology in the following discussion, it will be assumed that this functionality is implemented within each router, even although it is not in practice.

Besides distributing information about LR-PONs and routers, the IP address prefixes for the primary and secondary LR-PONs must be distributed, in order to permit derivation of a customer's secondary IP address from its primary IP address. Holding this information in prefix form reduces table size and facilitates rapid searching.

User Datagram Protocol (UDP) over IP carries out signalling for database synchronization. However, this information could in principle be carried by any

suitable means. Although transport of signalling over IP is assumed here, the databases this generates may be used to implement protection for other technologies, as discussed in Section 4.

Although the synchronization protocol borrows ideas and concepts from existing Internet routing protocols such as RIP and OSPF [9], it uses a very simple and fast signalling method, and requires simple computation at each router. It carries signalling information over the existing core network while avoiding the need for damping, in order to reduce signalling and convergence times.

2.A. Operation of database synchronization

Each router is informed of local OLT or LR-PON failures via a specially designed and optimized local area network. It also uses a protocol such as BFD (bidirectional forwarding detection [10]) or SDH/SONET to learn whether it can communicate with each of its neighbouring routers. All this information is flooded via triggered updates (carried in UDP datagrams) to all the other routers in the network (or network area, as discussed later). Triggered updates also contain any relevant IP addressing information for LR-PONs in the cluster.

In this way, each router acquires global knowledge, over the area or whole network, about which other LR-PONs are available, and what IP address prefixes are associated with commercial and residential customers in each of them. This is held in a database in each router, which also indicates which pairs of routers can communicate with one another via a direct link. If a router can communicate directly with no others, it is assumed to have failed.

2.B. Triggered update packets

Routers usually communicate reachability information via triggered update packets, which also contain overhead information. The latter includes an MD5 (or other) digest of a text password, to address security issues. Triggered updates are sent by each router to all its immediate neighbours.

A triggered update packet reports on the status (working or not working) of one OLT, or one link between two routers. Each router sends a triggered update immediately when a local OLT fails or is repaired, informing all other routers. It also sends a triggered update if BFD or SDH/SONET detects that a neighbouring router has become either reachable or unreachable via a particular direct link. A router only originates a new triggered update in this way (rather than relaying existing information) when changes occur adjacent to it.

Each router holds a current serial number for each link to neighbouring routers, and for each LR-PON in its cluster. When new information arrives about a local LR-PON or a direct link to a neighbouring router, this serial number is incremented by one and placed in the relevant field of the record in the router's database. This serial number is then carried with the LR-PON or link status information in each triggered update that is sent out. Each router has a serial number field for each record held in its database, indicating the serial number generated by the router originating the information. This provides a way of preventing out-of-date information from circulating around the network.

When a router receives a triggered update it determines whether the information is more recent than its existing records, by examining the respective serial number fields.

If so, it updates this record and immediately broadcasts another triggered update containing the new information to all neighbouring routers. This message carries the serial number of the new information. This is crucial to the robustness of the protocol, ensuring that even if a link or router in the core has failed, and triggered updates are lost, the information still reaches its destination via another route. Its impact on performance is demonstrated later by simulation. Furthermore, this arrangement expedites propagation of signalling information between routers, since the receiver need only wait for the first of several similar triggered updates to arrive via different routes.

2.C. Advertisement packets

In practice, it would be necessary for a router which is booting up to request a copy of a neighbour's database. This could be achieved by sending a "request" packet, with the response being one or more "advertisement" packets containing the contents of the database. It would also be necessary each router to send advertisements at infrequent intervals to all their neighbours, in case of corruption due to network errors or other causes. These features are not implemented in the simulation of Section 5.A, since, particularly without errors, their impact on performance is negligible.

2.D. Comparison with OSPF

It is natural to ask if the functions described above could be provided by an existing protocol, most notably OSPF [9]. Firstly, if OSPF were to implement LR-PON resilience, it would have to be implemented at every customer's premises. This is not appropriate or practical, for example, for reasons of scalability. Furthermore, a comprehensive redesign of OSPF would be necessary in order to ensure sufficiently fast recovery, as explained below.

In OSPF, flooding of link state advertisements (LSAs) throughout an area happens very quickly, in a few tens of milliseconds. However, to ensure stability, and to prevent route flapping, Dijkstra's algorithm should only be rerun once all relevant LSAs have arrived at a router. For this reason, after receiving the first LSA due to a fault, each router waits for several hundred milliseconds before Dijkstra's algorithm is run. This is particularly necessary if there has been a major fault which produced many LSAs. Furthermore, some implementations perform pre-processing on the link state database) during this time, in order to promote stability. Also, this prevents Dijkstra's algorithm from running too frequently, which could overload the router's processor and prevent it from responding to "hello" messages.

These features implement "damping" of OSPF's response to a fault. With OSPF, each fault normally produces at least two LSAs, since if a link fails, each end sees the fault and originates a changed LSA. If several routers, which were connected to each other and other routers, went down due to a fault, each router affected would then send out an LSA.

Such damping is not desirable for LR-PON resilience due to the delay it introduces, nor is it necessary. Assume that a triggered update has been received, and assume that the condition that produced it (for example, failure of a LR-PON) remains valid. Based upon this, each router makes a decision about the status of the LR-PON, namely whether it is working or failed. It is then impossible for another triggered update to reverse this decision. This is because there is only a single triggered update serial number associated with a given LR-PON or core link failure. On the other hand,

suppose that in OSPF, Dijkstra's algorithm is run immediately after receiving the first LSA, and a first-hop routing decision is made. Then another LSA may arrive shortly afterwards which would change that decision, resulting in instability.

3. Protection of IP traffic in the network layer

When a LR-PON or a router has failed, IP datagrams are rerouted to the appropriate secondary LR-PON OLT by routers elsewhere in the network. To do this, datagrams are either placed within tunnels (IP-in-IP), or their IP destination address fields are overwritten. This takes place whenever a datagram enters either the core network, or an "area" of the core network.

3.A. Allocation of IP addresses to customers

Each LR-PON is assigned several prefixes, or IP address ranges, which may be defined and dimensioned individually by the service provider, and indicate all permissible IP addresses for either commercial customers or residential customers. These two groups of customers require separate prefixes because they are treated differently in event of a fault. A customer may define its own prefix, allowing their existing IP addresses to be retained if required, furthermore DHCP (Dynamic Host Configuration Protocol [11]) can provide selected customers with static addresses. For each prefix on a primary LR-PON, a secondary prefix of the same size exists for interfaces on the secondary LR-PON. Thus a pair of LR-PONs configured for dual homing typically have more than one pair of prefixes associated with them, for use when making rerouting decisions.

3.B. Re-routing in the event of failure

Table 1 is a sample router database, which is derived from the databases described in Section 2. It is used to make IP rerouting decisions, with the first two shaded rows corresponding to the IP address in Fig. 3. Normally, both the primary and secondary LR-PONs are reachable, so no rerouting is necessary. An IP datagram addressed to the primary LR-PON is in fact routed via that LR-PON's OLT. The first three columns in the table indicate the prefixes associated with the primary and secondary LR-PONs, and the number of IP addresses defined by them. The next two columns indicate whether the primary and secondary LR-PONs are working and available.

Primary prefix	Secondary prefix	Number of IP addresses	Primary working?	Secondary working?	OLT address (residential)	Manual reroute?
177.67.40.0/22	177.67.44.0/22	1024	no	yes	0.0.0.0	no
142.31.16.0/25	142.31.16.128/25	128	no	yes	133.96.144.213	no
152.52.4.0/24	152.52.5.0/24	256	yes	yes	0.0.0.0	yes
etc.…	etc.…	etc.…	etc.…	etc.…	etc.…	etc.…

Table 1. An example of a routing table used for rerouting IP datagrams. The shaded rows are illustrated in Fig. 3.

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), it is rerouted via IP-in-IP tunnelling [12] to the corresponding secondary address, which can be easily deduced from the primary address. All customers are protected from cable failures in the backhaul regime with commercial customers experiencing full end-to-end protection, since they require full end-to-end equipment diversity and circuit routing separation. A source sending datagrams to a commercial customer is unaware that the ultimate destination

address may be on the secondary LR-PON. This is illustrated in Fig. 3, where the customer's primary IP address is 177.67.40.13, and its secondary IP address is 177.67.44.13. There will be many more residential customers than commercial customers, hence the prefix size for the latter will be correspondingly smaller.

Residential customers do not have paired (primary and secondary) IP addresses, hence datagrams cannot be routed to such a customer via two different OLTs. Here, the IP tunnel terminates at the secondary OLT, which then decapsulates the IP datagram and forwards it to the customer. The secondary OLT is adjacent to a different router that the primary OLT (Fig. 3), and, besides being capable of decapsulating IP-in-IP, it has its own IP address. In Fig. 3, the IP address allocated to the lower OLT is arbitrarily set to 133.96.144.213.

The secondary OLT associated with residential customers must rerange prior to receiving the first rerouted data packet from the source. It is instructed to do so by a notification packet sent by the router adjacent to the fault. If the diverted data packets arrive before reranging is complete, they are buffered by the local router until the OLT is ready.

The last column of the table is labelled "manual reroute" and may be configured manually by the network administrator to permit rerouting without any fault, primarily for network maintenance. Traffic redirection is implemented in the multiplexer which concentrates OLT signals into the router.

3.C. Address substitution and IP-in-IP encapsulation

IP-in-IP encapsulation [12] has a disadvantage in this context, despite being a widely accepted technique. For any given value of maximum transmission unit (MTU), IP-in-IP encapsulation reduces the possible payload size by 20 bytes, the size of an IP header. This may cause problems with those applications using UDP which tend to transmit long packets, as it would cause fragmentation. The effect on TCP is more subtle. When TCP sets up a connection, it determines the MTU via path MTU discovery. If, while a TCP connection is in progress, it is diverted into an IP-in-IP tunnel, the MTU will effectively decrease by 20 bytes without TCP's knowledge. If, as is likely, TCP now transmits a packet with the size of the old MTU, it will be fragmented by IP when it passes through the tunnel.

Hence fragmentation of IP datagrams takes place with both UDP and TCP. Although the resulting degraded efficiency is generally undesirable, it may be tolerable for a short time, since it only arises temporarily after a fault. The additional headers use up transmission resources, while the transmission of additional fragments requires more computation by routers.

This problem may be avoided by substituting the new IP address into the datagram's header, without tunnelling, but this creates problems and complexities of its own. In some circumstances, only datagrams having the correct destination IP address are acceptable to a host. Hence such address substitutions would then confuse IP. This is resolved by making the secondary IP address *on a dual-homed customer's premises* equal to its primary address, thus avoiding such addressing difficulties there. However, the secondary IP addresses shown in Fig. 3 are used by the core and its routing protocols, although both customer ports have the same IP address. To permit forwarding through the core, this secondary IP address is replaced by the primary IP

address in each datagram emerging from the core, when it passes via traffic redirection to a LR-PON.

4. Protection of other networking technologies

While the signal carried in order to synchronize router databases takes place using IP, most technologies can potentially be protected, as detailed below. All traffic (IP, Ethernet, time-division multiplexed (TDM) etc.) raises the general underlying issue of many traffic streams from multiple points in the network requiring routes to primary and backup locations. Different platforms could be adapted to take advantage of the approach introduced here, with signalling for database synchronization still being carried over IP. Re-routing of some TDM technologies, such as private lines, would require an interface to the SDH management system, but could avoid bandwidth duplication when providing protection across the core network.

4.A. Protection by WDM lightpaths

Protected LR-PON traffic could in principle be rerouted via WDM wavelength links. Assume that a working LR-PON is attached to router X, and that the corresponding backup LR-PON is attached to router Y. If the working LR-PON fails, traffic for it entering X can be diverted over a wavelength channel to Y to the backup LR-PON. There are two ways of doing this. Traffic can be groomed at higher layers within router X from multiple incoming wavelength links and then sent over a single wavelength link to router Y. However, this is not really protection using WDM, since higher layers are involved. Alternatively, optical cross-connects could be placed on the backhauls of LR-PONs on routers X and Y in order to redirect the traffic; however this would imply cost, complexity and LR-PON ranging issues. This is bandwidth efficient only if the capacity of a wavelength is the same as the capacity of a LR-PON, although this is generally a reasonable assumption.

Assume X has next-hop nodes in the core that are labelled say A, B and C, and routes (not passing through X) exist from each of these to Y. If router X fails, all traffic destined for all LR-PONs connected to X is diverted to Y by redirecting it at nodes A, B and C, based upon information in each router's database. The diverted traffic is generally spread over several wavelength links upon entering A, B or C, and each of these may be shared with other traffic. Hence grooming is necessary at higher layers in each of nodes A, B, and C. For these reasons, since intervention is required from higher layers, it is not really feasible to protect LR-PON traffic purely in the WDM layer.

4.B. Protection of MPLS when it is delivered to the customer

Rarely, MPLS may be offered directly to a customer, with a label edge router (LER) at each customer's premises. The LR-PONs appear to MPLS to be exactly like any other links in the network, and existing techniques, such as fast reroute [13] can be employed. It will be necessary for events such as LR-PON backhaul failure to be signalled to MPLS, in order to facilitate encapsulation within tunnels for recovery. If necessary, information could be provided by the database of functioning routers and functioning LR-PONs, which is maintained by the synchronization protocol described above.

4.C. Protection of Ethernet

Usually, Ethernet is based on spanning trees, VLANs, learning bridges and learning hubs. To implement dual homing, each host would have one Ethernet address. If it sends a broadcast packet on either LR-PON, the self-learning process will ensure that all traffic for that address will now be directed correctly to that LR-PON. Hence little special provision is necessary to implement dual homing in this case. Protection of PBT (Provider Backbone Transport [14]) traffic is a separate topic, which requires further work.

5. Performance and scalability of router database synchronization

5.A. Simulation model

The speed of database synchronization was investigated by simulation, using the commercially available OPNET simulator [15]. The database synchronization delay is defined as the time between a new fault, and all routers having identical updated databases. The following parameters were used in the simulation:

- The network had 100 routers.
- Each router was connected to a minimum of three (and maximum of ten) other routers, with the probability of a router being connected to *n* neighbours being proportional to 0.54^n. These probabilities were normalized so that they summed to unity. Thus each router was directly connected to a mean of approximately 4.11 others.
- The link lengths between routers were described by a uniform probability distribution, with minimum 10 km and maximum 1000 km.
- The line rate was 10 Gbits/s.
- Light propagated at 2×10^8 m/s in each fibre.
- The UDP datagrams carrying triggered updates were 48 bytes long.
- In the simulations, 1 ms table processing time was assumed between an incoming triggered update arriving and new ones being generated. In practice, the value would be much less, so this is a worse-than-worst-case estimate.
- It was assumed that the time taken to report a fault to a router was 1 ms. This is, for example, a worst-case estimate for a LOS (Loss of Signal) message to be generated in SDH/SONET.

A number of different scenarios were modelled:

- Two traffic scenarios:
    - The triggered updates were carried as best-effort IP traffic in the presence of background traffic. This was self-similar with a Hurst parameter of 0.8. The traffic loading was ~ 0.4. This is referred to as "best effort" signalling traffic.
    - Triggered updates were carried as priority IP traffic. This was evaluated very simply by not simulating any other traffic on the links.

- The effect of multiple link failures in the core on synchronization time.
- Synchronization after two types of fault:
    - LR-PON failure, where a single triggered update is flooded around the network.
    - Edge router failure, where a triggered update is flooded around the network for each neighbour router that cannot reach the failed router.

Four options are considered in the simulations:

- Best effort transport, LR-PON failure;
- Priority transport, LR-PON failure;
- Best effort transport, router failure;
- Priority transport, router failure.

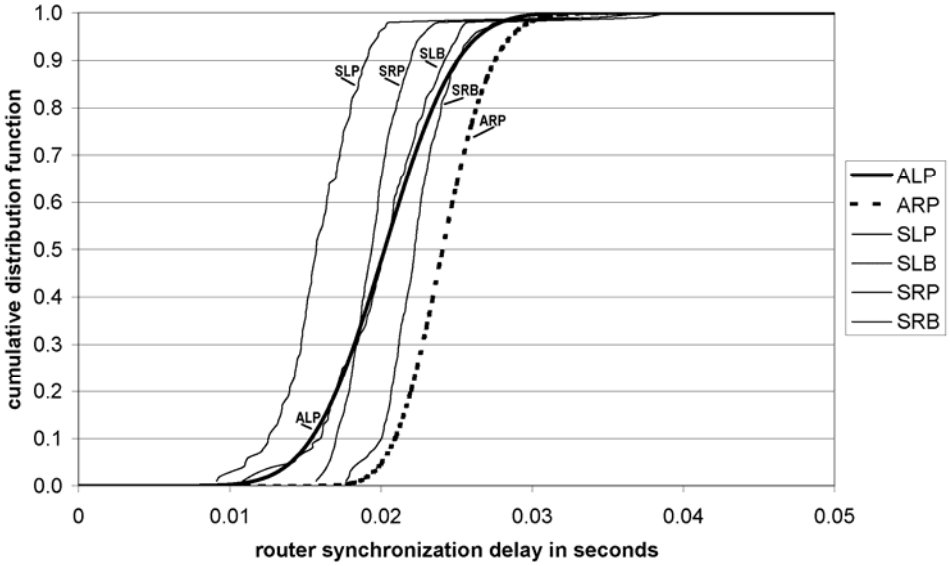


Fig. 4: Cumulative distribution function of router synchronization time, calculated analytically and by using simulation. The analytical results model priority signalling traffic, and are shown as a thick bold curve (LR-PON failure: ALP) and a dashed curve (router failure, ARP) respectively. The remainder is simulation results showing, from left to right, LR-PON failure with priority signalling traffic (SLP), router failure with priority signalling traffic (SRP), LR-PON failure with best-effort traffic (SLB), and router failure with best-effort traffic (SRB).

Figure 4 shows the cumulative probability of recovery time for these four options, assuming no core link failures. The analytical results are also shown, for priority transport of signalling of both LR-PON and router failures. As expected, the mean recovery time for best-effort transport is greater. Furthermore, when synchronizing after router failure, synchronization takes longer because there is more than one triggered update, and all have to be received before synchronization takes place.

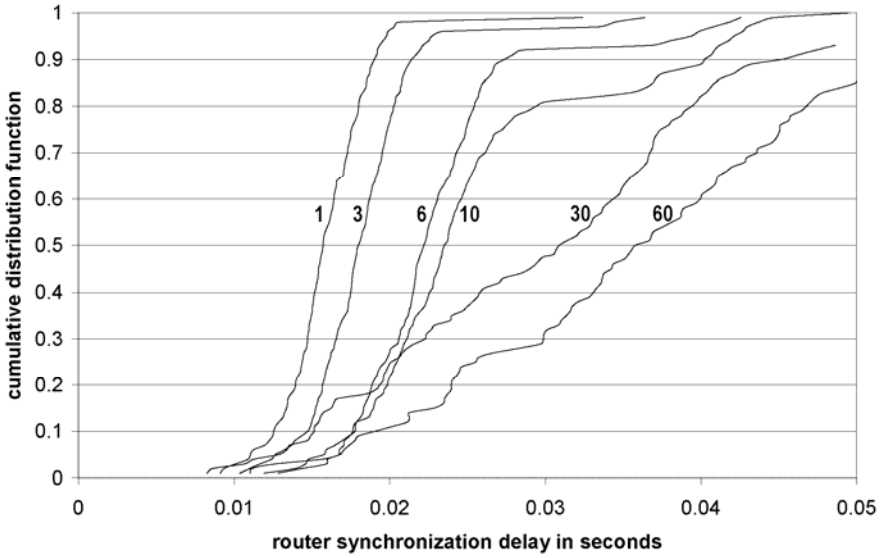


Fig. 5: Cumulative distribution function of router synchronization time, considering different numbers of failures in the core network. The LR-PON failure being modelled is counted as one of these failures, and priority signalling traffic is used throughout. The traces indicate from left to right, 1, 3, 6, 10, 30, and 60 failures. As the number of failures increases, the CDF increases more slowly with delay.

Figure 5 shows the effect on synchronization time of faults in the core network, for priority traffic. The total number of faults is shown in the graph, including the original LR-PON or router failure. As expected, the mean time taken to synchronize databases increases because fewer routes are available, as the number of failures increases. The time taken to achieve synchronization with a probability of say 90% increases steadily for 1 to 20 failures, however, for further failures, the time levels off. This indicates that the protocol degrades gracefully, even under pathological network failure conditions.

Table 2 shows the probability (expressed as a percentage) of synchronization taking place in less than 10, 20, 30, 40, and 50 ms. All these results, both analytical and simulated, are for commercial customers. The results for residential customers would be very similar, but only if informing the new destination LR-PON of its status, and carrying out re-ranging, took place well within 50 ms. For this to happen, the backup LR-PON would have to be provided with ranging information by the primary LR-PON. Techniques to do this are a topic for further research, which is being carried out elsewhere. If re-ranging took more than the database synchronization time, then it would become a constraint on the traffic recovery process.

Figure 6 shows, for the case of priority traffic, the amount of traffic generated by the protocol in each router as synchronization takes place. As expected, it shows a peak in the volume of traffic at the middle of the recovery time. The graph shows that the overall volume of signalling traffic generated during router database synchronization is negligible, compared to the link capacity.

	failures	Sync<10ms	Sync<20ms	Sync<30ms	Sync<40ms	Sync<50ms
Best effort transport, LR-PON failure	1	1%	47%	98%	100%	100%
	10	0%	12%	77%	87%	100%
	30	1%	17%	38%	67%	90%
Priority transport, LR-PON failure	1	2%	96%	100%	100%	100%
	10	0%	21%	81%	88%	100%
	30	2%	24%	47%	82%	100%
Best effort transport, router failure	1	0%	9%	98%	100%	100%
	10	0%	1%	60%	78%	99%
	30	0%	0%	27%	52%	83%
Priority transport, router failure	1	0%	64%	98%	100%	100%
	10	0%	2%	76%	83%	100%
	30	0%	1%	42%	56%	92%

Table 2: Probability (expressed as a percentage) that synchronization takes place in less than 10 ms, 20ms, 30 ms, 40 ms and 50 ms.

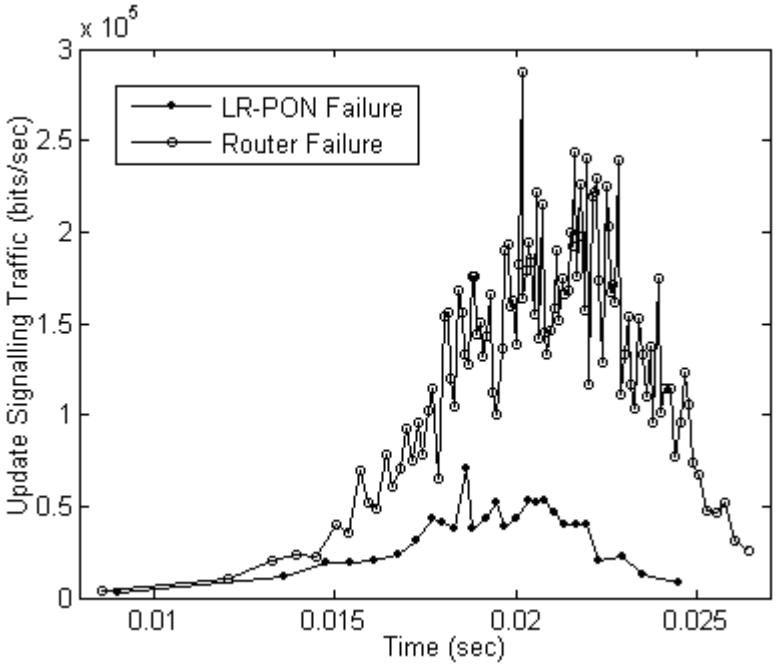


Fig. 6: Volume of signalling traffic generated during router database synchronization: the original LR-PON fault is included in this figure. The scenario is LR-PON failure with signalling carried via priority traffic.

5.B. Analytical model

The routers leading from the origin of a triggered update form a tree, where each router is connected to b other routers. This quantity is taken to have the same mean as the distribution used in the simulation, namely 4.11, and this is the value used for b in the calculations. Assuming that each router in the tree is regarded as being distinct, $b - 1$ copies of the triggered update are made at each router which is less than h hops from the source, and passed on further down the tree. Let k be the number of routers in the tree, and let $C(n, k)$ be the probability that these k routers correspond to n routers in the real network, where $k \geq n$. This is evaluated recursively, where $\lfloor x \rfloor$ is the largest integer less than or equal to x, and $\lceil x \rceil$ is the smallest integer greater than or equal to x:

$$C(n,k) = \sum_{m=\lfloor n/2 \rfloor}^{k-\lceil n/2 \rceil} \frac{k!}{m!(k-m)!} \left(\frac{\lfloor n/2 \rfloor}{n} \right)^m \left(\frac{\lceil n/2 \rceil}{n} \right)^{k-m} C(\lfloor n/2 \rfloor, m) C(\lceil n/2 \rceil, k-m),$$

$C(1, 0) = 0$; $C(1, k) = 1$ where $k > 0$. In the formula, the n real routers in the network that are under consideration are divided into two equal halves (or two approximately equal halves if n is odd.) The first part of the summation is similar to a binomial distribution. It is the probability that the first half of the n real routers has m tree routers assigned to it, and that the second half has $k - m$ tree routers assigned to it. The remainder is the probability that m tree routers are assigned to all the real routers in the first half, and that $k - m$ tree routers are assigned to all the real routers in the other half.

Throughout it is assumed that each tree router is assigned to a real router, where in each case, the choice between real nodes is made randomly and with equal probability. In reality, there is a correlation between tree nodes which are assigned to the same real node: they all have the same neighbouring real nodes. Nevertheless, in the model, each real node may correspond to zero, one, or more than one tree node, corresponding to the real case where the tree may reach the same real node more than once, or not at all.

The probability that all nodes in the network can be reached in h or fewer hops from the originating node is

$$D(h) = C\left(N, \frac{(b-1)^{h+1} - 1}{b-2} \right),$$

where N is the number of nodes in the network, taken to be 100, as in the simulation. The second argument of $C(n, k)$ is the mean number of nodes in a tree of depth h and mean degree $b - 1$, which is equal to the sum of a geometric series, i.e., $1 + (b - 1) + (b - 1)^2 + \ldots + (b - 1)^h$, with each term representing the number of nodes on successive levels of the tree. The probability that all nodes can be reached in h hops but not fewer is therefore

$$E(h) = D(h) - D(h-1).$$

As in the simulations, the link delay (including processing delay) is uniformly distributed with limits $a = 0.05 + 1 = 1.05$ ms and $b = 5 + 1 = 6$ ms. The mean and standard deviation are [16]

$$\mu = \frac{b-a}{2} = 3.525 \times 10^{-3} \text{ s},$$

$$\sigma^2 = \frac{(b-a)^2}{12} = 2.041875 \times 10^{-6} \text{ s}^2.$$

Assuming a normal distribution with h hops, which approximates the convolution of several uniform distributions, the cumulative distribution function (CDF) yielding the probability that the delay is less than t is:

$$F_h(t) = \frac{1}{\sqrt{2\pi h\sigma^2}} \int_{-\infty}^{t} \exp\left(\frac{-(\tau - h\mu)^2}{2h\sigma^2}\right) d\tau .$$

The CDF, denoting the probability of the synchronization time for LR-PON failure being less than t, is the weighted sum of $F_h(t)$:

$$F(t) = \sum_{h=1}^{\infty} E(h) F_h(t).$$

For ease of computation, the recursive formula for $C(n, k)$ above may be approximated by $[1 - (1 - 1/n)^k]^n$, with the resultant difference in $F(t)$ always being less than 3.55% for $N = 100$, with μ and σ^2 as above. $1/n$ is the probability that a particular router in the tree is assigned to a particular real router, so $1 - 1/n$ is the probability that this does not take place. Because $(1 - 1/n)^k$ is the probability that no routers from the tree are assigned to a particular real router, $1 - (1 - 1/n)^k$ is the probability that a particular real router has one or more tree routers assigned to it. The approximating assumption of independence between the real routers is then made, hence the probability of all real routers each having one or more tree routers assigned to them is approximately $[1 - (1 - 1/n)^k]^n$.

With a mean of b neighbours for each router, the CDF for synchronization time after router failure, where a mean of b triggered updates must reach each router, is approximated by

$$G(t) \approx F(t)^b .$$

These are plotted with the corresponding simulation results in Fig. 4, denoting synchronization of LR-PON failure and router failure, with priority transmission of triggered updates. The simulation and analytical results match closely, with the analytical results yielding a slightly pessimistic estimate of synchronization time.

6. Use of areas with multiple operators

The network may be divided into areas, with redirection of IP traffic 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. Between areas, carrier grade equipment is required, due to the traffic volume.

Areas permit several telecommunications operators to coexist, even if they do not all use this type of protection. Each area may be owned by a different operator, although no operator need use this protection scheme if it does not wish to. In that case, the operator does not need to be aware of these protocols, nor does it need to make any concessions or adjustments because the other operators are using them.

In such a multioperator scenario, redirection of traffic is only necessary when rerouting traffic flowing into an area which uses the type of database synchronization described here, either from another area or from a LR-PON. The database synchronization protocol only gathers information about its own area: there is no communication between its implementations in different areas. Operators not using this protection scheme are unaware that they are being used elsewhere.

Furthermore, several operators may access one LR-PON, where traffic on the core side of an OLT is split between them. Thus this protection scheme permits flexible configuration of the network to support multiple operators.

7. Conclusions

If a simplified and low-cost network architecture, consisting of long-reach PONs linking customers directly to a core of a hundred or so switching and intelligence centres (routers) is to be realized, then levels of resilience and protection as good or better than those currently experienced are essential. Such solutions remove distinctions between access and backhaul networks, but nevertheless need to offer service guarantees that are appropriate to these regimes. Although many techniques exist to perform fast protection in the event of a cable failure, these tend to require that a single node is responsible for the protection switching prior to directing the traffic into the inner core network; making such equipment a single source of failure. Techniques that remove the need to "close" the protection at a specific point either lead to inefficiencies in the core network or take time to reconverge in the event of a failure. Clearly new approaches are required.

This paper has described two new techniques, IP traffic redirection and database synchronization, which can reroute traffic in the event of a variety of failure scenarios, thus ensuring that traffic originating from elsewhere in the network reaches the correct destination. Calculations have shown that it is realistic to expect end-to-end recovery to take place within 50 ms.

The adoption of these techniques could provide an operator of a means whereby traffic can be dual-homed onto two separate inner core nodes and fast traffic redirection can take place in the event of cable failures in the access-backhaul network or for major failures of nodes that act as gateways between the access and inner core network.

8. Acknowledgments

British Telecom supported the work of David Hunter under its Short Term Fellowship scheme. The authors acknowledge Rob Booth, Russell Davey, Alan Hill, Alan McGuire, Ben Niven-Jenkins, Albert Rafel and Peter Willis (all of British Telecom) for their constructive and helpful comments during this work. Similarly, Martin Reed of the University of Essex is thanked for his input.

9. References

1. D. B. Payne, R. P. Davey, "The Future of Fibre Access Systems?", *British Telecom Technology Journal*, vol. 20, no. 4, October 2002, pp104-114.

2. R. P. Davey, P. Healey, I. Hope, P. Watkinson, D. B. Payne, O. Marmur, J. Ruhmann and Y. Zuiderveld, "DWDM reach extension of a GPON to 135km", Post-deadline paper PDP35, *OFC 2005*, Anaheim, California, March 2005.
3. D. Nesset, D. B. Payne, R. P. Davey, T. Gilfedder, "Demonstration of Enhanced Reach and Split of a GPON System Using Semiconductor Optical Amplifiers", Paper Mo4.5.1, *ECOC 2006*, Cannes, France.
4. ITU-T Study Group 15, "Optical and other transport network infrastructures", Question 2, "Optical systems for fibre access networks".
5. "Gigabit-capable Passive Optical Networks (GPON): General Characteristics", ITU-T recommendation G.984.1, 2003.
6. "Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) Layer Specification", ITU-T recommendation G.984.2, 2003.
7. "Gigabit-capable Passive Optical Networks (GPON): Transmission Convergence Layer Specification", ITU-T recommendation G.984.3, 2004.
8. "Gigabit-capable Passive Optical Networks (GPON): ONT Management and Control Interface Specification Amendment 3", ITU-T recommendation G.984.4, 2006.
9. C. Huitema, *Routing on the Internet*, Prentice Hall, 2000.
10. R. Aggarwal, "OAM Mechanisms in MPLS Layer 2 Transport Networks", *IEEE Communications Magazine*, October 2004, pp124-130.
11. R. Droms, "Automated Configuration of TCP/IP with DHCP", *IEEE Internet Computing*, July-August 1999, pp45-53.
12. C. Perkins, "IP Encapsulation within IP", *Internet Engineering Task Force*, RFC 2003, October 1996.
13. G. Suwala, G. Swallow, "SONET/SDH-Like Resilience for IP Networks: A Survey of Traffic Protection Mechanisms", *IEEE Network*, March-April 2004, pp20-25.
14. D. Allan, N. Bragg, A. McGuire, A. Reid, "Ethernet as Carrier Transport Infrastructure", *IEEE Communications Magazine*, February 2006, pp134-140.
15. OPNET network modelling software: www.opnet.com
16. N. Hastings, B. Peacock, M. Evans, *Statistical Distributions*, Wiley, 2000.