DATA-CENTRIC OPTICAL NETWORKS AND THEIR SURVIVABILITY

Didier Colle, Sophie De Maesschalck, Chris Develder, Pim Van Heuven,

Adelbert Groebbens, Jan Cheyns, Ilse Lievens, Mario Pickavet, Paul Lagasse, Piet Demeester

Ghent University - IMEC, Department of Information Technology

Sint-Pietersnieuwstraat 41, 9000 Gent (Belgium)

***tel. no**. +32 9 267 35 93*

***fax. no**. +32 9 267 35 99*

***e-mail** {didier.colle, sophie.demaesschalck, chris.develder, pim.vanheuven, adelbert.groebbens, jan.cheyns, ilse.lievens, mario.pickavet, lagasse, demeester}@intec.rug.ac.be*

Abstract. The explosive growth of data-traffic, for example due to the popularity of the Internet, poses important emerging network requirements on today's telecommunication networks. This paper describes how core networks will evolve to Optical Transport Networks (OTNs), which are optimised for the transport of data-traffic, resulting in a IP-directly-over-OTN paradigm.

Special attention is paid to the survivability of such data-centric optical networks. This becomes more and more crucial, since more and more traffic is multiplexed onto a single fiber (e.g, 160*10Gbps), implying that a single cable cut can affect incredible large traffic volumes. More in particular, this paper is tackling multi-layer survivability problems, since a data-centric optical network consists of at least an IP and optical layer. In practice, this means that the questions "in which layer or layers to provide survivability" and "if multiple layers are chosen for this purpose, then how to coordinate this functionality in these layers" have to be answered.

In addition to a theoretical study, some case studies are presented in order to illustrate the relevance of the described issues and to help in strategic planning decisions. Two case studies are studying the problem from a capacity viewpoint. Another case study presents simulations from a timing/throughput performance viewpoint.

Keywords: multi-layer survivability, MPλS, MPLS, IP-over-OTN, recovery, capacity dimensioning.

Submission for IEEE Journal on Selected Areas in Communications,
special issue on WDM-BASED NETWORK ARCHITECTURES

1. Introduction: from IP/ATM/SDH/WDM to IP-MPLS directly over OTN-MPλS

The popularity of the Internet [1], [2] has lead in recent years to an explosive growth of the traffic to be carried by telecommunication networks. Since a few years data traffic even dominates voice traffic [3]. and recent forecasts don't seem to predict a quick slowdown of this greediness [4] [3].

It is obvious that this will have a major impact on today's telecommunication networks. These networks will be more and more optimized for the dominant data (mainly IP) traffic. Today, a typical (core of a) telecommunication network consists of a transport network carrying the traffic of several parallel services: e.g., Plain Old-switched Telephone Service (POTS), leased-line services, etc. Such a Transport Network (TN) may e.g. consist of an ATM network (functioning as service integration layer) on top of an SDH network. Fiber exhaust is currently solved by multiplying the capacity of a fiber ten – or even hundred – times by means of point-to-point Wavelength Division Multiplexing (WDM) systems. Recently, WDM-systems of 160 10Gbps wavelengths have been announced [5]. This multiplexing technique has proven to be very cost-efficient due to the economy-of-scale [6].

It is obvious that incumbent operators also want to profit from the new Internet Service Provider (ISP) market fragment. They are at a more comfortable position, since they still have their important revenue-generating voice [3] business and other services, in contrast to new-comers. However, they are of course not willing to immediately replace their current infrastructure and thus they start their ISP business by running their IP network in parallel with their currently existing network services, on top of the same transport network. This means they typically are in (or have just left) an IP/ATM/SDH/WDM multi-layer scenario [7]. The practical meaning of this scenario is explained in Figure 1.

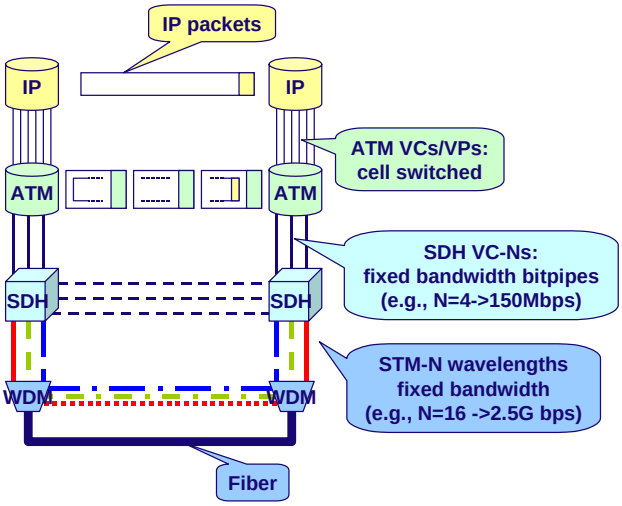


Figure 1: illustrates the IP/ATM/SDH/WDM technology mapping. IP routers exchange IP packets, by sending them through ATM connections, which requires encapsulation of an IP packet in many ATM cells. ATM nodes are interconnected by fixed bandwidth bitpipes (VC-Ns) through the SDH network. The capacity on the fibers interconnecting the SDH DXCs is increased by multiplexing multiple wavelengths onto a single fiber.

The transport of IP packets through ATM has some major drawbacks. First of all, there is the important cell tax: approximately 10% overhead (5 bytes header per 48 bytes payload). Secondly, an IP packet has a typical length of 500 or 1500 bytes [8] and is thus typically encapsulated in many ATM-cells. This implies that per IP packet many ATM cells have to be handled and processed in intermediate ATM nodes. Yet another disadvantage is that there is an extra layer to maintain and manage. Of course, ATM also has its benefits: its connection-orientation, opening opportunities for Traffic Engineering (TE), due to the decoupling of routing (control plane) and forwarding (data plane).

However, the steady and ongoing progress and research in optimizing IP router designs [9] implies that IP doesn't have to take the drawbacks of ATM for granted, if it would be able to overcome its lack in TE-capability. The MultiProtocol Label Switching (MPLS) concept, grown within the IETF, has proven to be suitable for this purpose [10], [11], [12], [13]. Thus, in the end, we may expect that an MPLS-empowered IP-network absorbs the TE-feature of ATM and bypasses the ATM-layer, by coding the MPLS-labels in a shim-header in front of the IP-packet. Similar to ATM, a Label Switched Router (LSR) will label-switch the packets (i.e., look up the incoming <interface, label>-pair in the Label Information Base (LIB), in order to know along which interface to forward the packet with which label). This bypasses the legacy cumbersome lookup

operation of the destination address in the routing table. To populate the LIB with appropriate mapping information, a protocol (either the Label Distribution Protocol (LDP [14]) or the Resource reSerVation Protocol (RSVP [15])) in the MPLS control plane will be used, allowing to setup and tear down so-called Label Switched Paths (LSPs) through the MPLS network. (Note that in the remainder of this paper we will use the following terminology, to refer to an "IP" network: *IP-network* refers to an MPLS-incapable network, *MPLS-network* is short for an MPLS-capable IP network, and *IP-MPLS network* will be used when it can be either an IP-network or an MPLS-network. It also may happen that we call an MPLS-network an MPLS-empowered/capable IP-network (to stress the MPLS capability). The services and traffic (demand) carried by an IP-MPLS network are always indicated by *IP-services* and *IP traffic* respectively.)

Even more, the steady growth of the IP traffic (will soon) allow(s) bypassing the ATM-layer, simply because the SDH switching granularity (will) match(es) the required line-speeds for the direct interconnection of IP-MPLS routers. IP-MPLS-router interface-cards of up to 622Mbps or even 2.5Gbps are currently commercially available and deployed [9], [16]. As traffic won't stop growing, in no time SDH Digital Cross-Connects (SDH DXCs) won't be able anymore to catch up with the required switching granularity (a coarse granularity of the underlying layer is beneficial for the IP-MPLS network from a scalability point of view). At that moment, SDH will be bypassed as well and the cross-connect functionality will be pushed into the optical domain, resulting in a so-called Optical Transport Network (OTN). Optical Network Elements (ONEs) with limited flexibility are already commercially available and full-flexible large Optical Cross-Connects (OXCs) are ready for massive commercialization [5], [17].

A final consideration in our roadmap for next generation networks is the fact that transport networks tend to be rather static, due to the fact that an operator has to setup each connection manually through the Network Management System (NMS). This doesn't match with the exponentially growing and highly dynamic IP traffic pattern, requiring frequent changes of the wavelength bandwidth pipes provisioned by the OTN-network to carry the IP-MPLS network traffic. Therefore, a current hot research topic is to investigate how this provisioning process can be automated. As in all switched networks, the control plane will serve this need, as illustrated in Figure 2. Signaling through the control channel of the User-Network Interface (UNI) – thus between the IP-MPLS and OTN network – (e.g., OIF UNI spec 1.0 [18]) makes it possible for the client to automatically request the setup of a new lightpath through the OTN. The control channel through the Network-Network Interface (NNI) allows the exchange of signaling messages for routing protocol information exchange

(e.g., Link-State Advertisements (LSAs) being used in the Open Shortest Path First (OSPF) routing protocol), setup of a lightpath, etc.

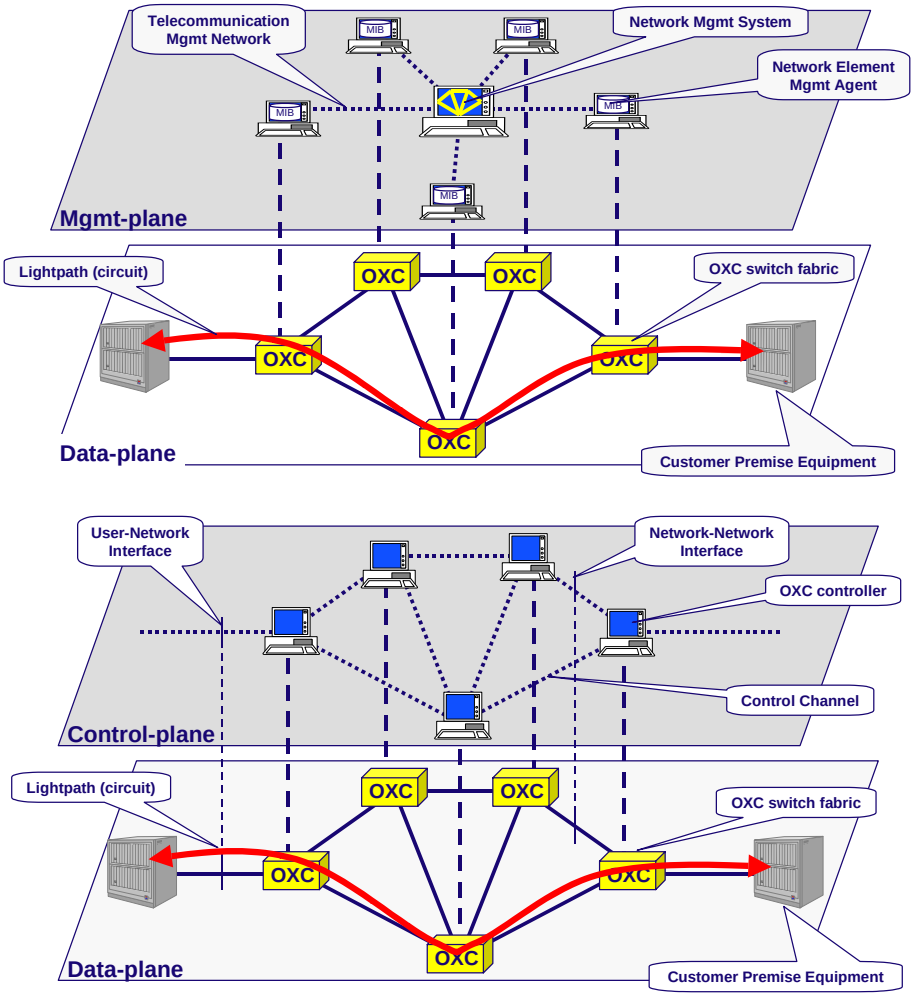


Figure 2: shows the difference between a static Optical Transport Network (OTN) at the top and an Automatically Switched Optical Transport Network (ASON) at the bottom of the figure. An ASON is an OTN, empowered with a (distributed) control plane (taking over a large part of the crucial functionality of the management plane), allowing signalling with the client through the UNI, in order to realize a switched optical channel service.

Generally speaking, there exist two main (extreme) models for an automatic switched optical network. ITU-T G.astn [19] targets an overlay model for an Automatically Switched (Optical) Transport Network (ASTN is a generalization of ASON). In the overlay-model, both the transport and its client networks have a separated and independent control plane. The IETF targets more a peer-model with the Generalized-MPLS (G-MPLS) concept. This concept originated from MPλS, where the idea was that a wavelength (Lambda) is a label as any

other label and therefore the MPLS concept can be adopted in the optical domain to serve the need for fast automatic lightpath (or Optical LSP) provisioning [7], [20]. G-MPLS is generic in the sense that it considers any type of label: a header-bitstring for a Packet-Switch Capable LSR (PSC-LSRs), a time slot for a TDM-Switch Capable LSR (TSC-LSR: e.g., SDH-DXC), a wavelength for a Lambda-Switch Capable LSR (LSC-LSR: e.g., OXC) or even a fiber in a Fiber-Switch Capable LSR (FSC-LSR) [21]. A similar terminology as the one for IP-MPLS networks will be used for optical networks: *OTN* refers to an optical network not controlled by MPλS, an *MPλS network* to an optical network controlled by an MPλS control plane, and an *OTN-MPλS network* to an optical network, regardless of the type of the control plane.

Although both client and transport network may have their own separate and independent (G-)MPLS control planes, an integration of those control planes into a single one (covering both layers) seems obvious, resulting in the so-called peer-model. The difference between overlay and peer model is illustrated in Figure 3. The peer-model may have some advantages: avoiding duplication of control plane functionality in distinct layers, and avoiding the requirement of standardization of an UNI between IP-MPLS routers and OXCs (since the single integrated control plane controls both layers). However, it suffers from the fact that integration and compatibility amongst multiple client (type) networks seems to be hard and that all information (including confidential information like the TN-topology) is freely accessible in the client domain.

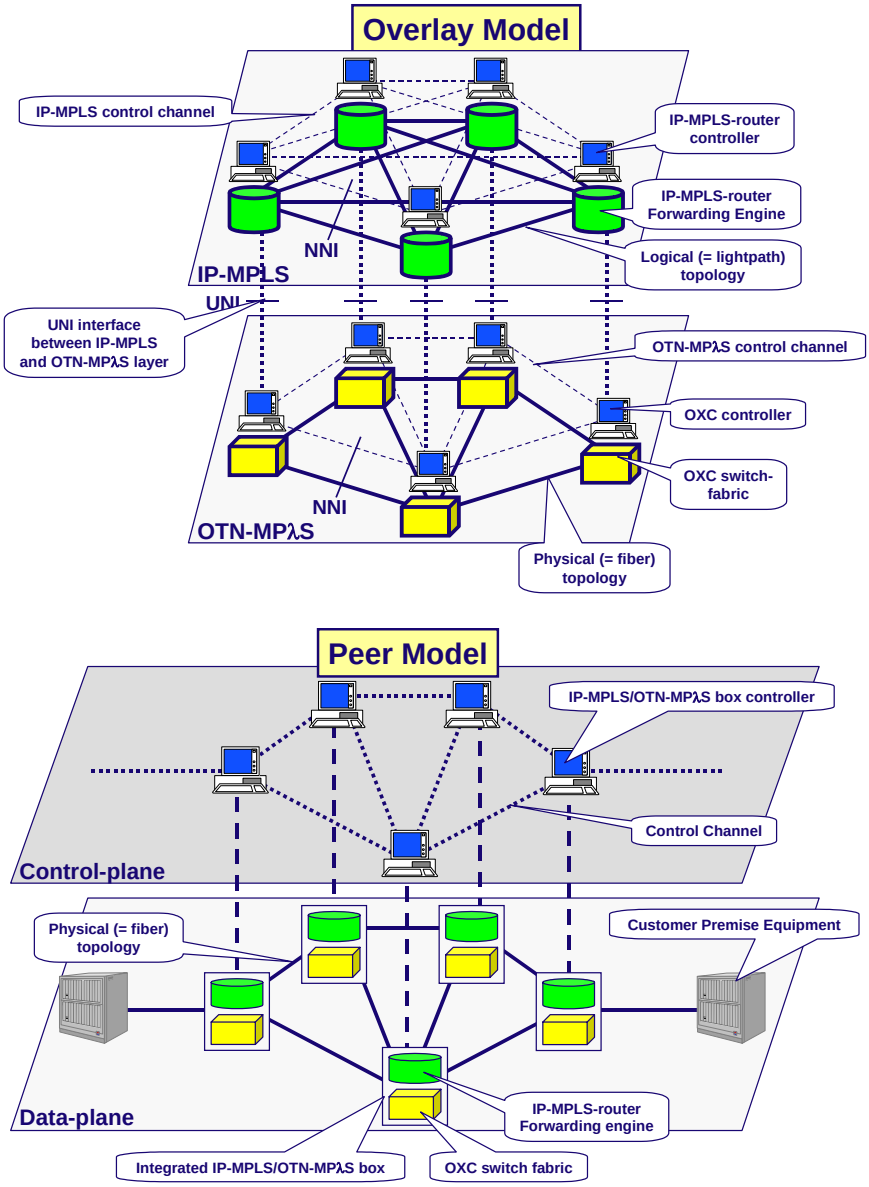


Figure 3: illustrates that in the overlay model (top) the client network is controlled by a separate control plane, independent from the control plane of the transport network. This is in contrast to the peer-model (bottom) where the control plane of the client network is integrated into the control plane of the transport network: thus, colocated client and transport network equipment are seen as a single entity.

Considering the expectation that in the long term the peer-model will become mature enough and eventually overtake the overlay-model (when IP-MPLS becomes the service integration layer), we propose as horizon for our roadmap a peer-modeled IP-MPLS/OTN-MPλS network. Note that this is the horizon of our roadmap, not the end of network evolution. There are already ideas to drive the switching granularity even higher (waveband switching or even fiber switching) and intensive research is going on in the field of Optical Packet Switching.

2. Enhancing survivability features of the G-MPLS technology for IP and OTN networks

It was already mentioned that the decoupling in MPLS of routing and forwarding opens opportunities for Traffic Engineering (TE). This is in particular true for the resilience aspects in TE. The goal of this section is to give a brief summary of the current proposals for network recovery in MPLS networks. The impact of G-MPLS is also studied. The reader is referred to [22], [23], [24], [25], [26], [27], [28], [13], [29], [7], [30] for more detailed information (terminology is not fixed yet and therefore we use our own terminology in this paper). Note that this section is focussing on resilience in a single layer (thus MPLS or MPλS): multi-layer issues are presented in a later section. The section is divided in protection and restoration, referring to the fact whether an alternative path is pre-established or not.

2.1 Restoration in MPLS

Restoration typically means that connections affected by a failure are routed along an alternative path that is calculated and set-up at the time of the failure: a big advantage of restoration is its flexibility. **MPLS rerouting** is an example of restoration. MPLS rerouting relies on the dynamic IP routing protocols. Failures are detected by adjacent routers (e.g. endpoints of a failing link) and advertised/flooded over the network, in order to allow other routers to take this topology change into account. After updating its routing tables, a router somewhere in the network may notice that it has LSPs leaving along another interface than indicated by the routing table entries corresponding to the destination of these LSPs. This will trigger the setup of LSPs along the correct (as indicated by the routing table) path.

One of the drawbacks of MPLS rerouting is that it may suffer from similar inefficiencies as the IP routing protocols on which it is relying: e.g., rather long convergence times, temporary instabilities and loops, etc. Therefore, a new MPLS restoration scheme was developed at our department: the so-called **Fast Topology-driven Constraint-based Rerouting** (FTCR): see Figure 4. It assumes that the MPLS network runs a link-state routing protocol (e.g., OSPF or IS-IS): this means that each link is advertised to all routers in the network and that each router stores all these advertised links in its link-state database (which gives an overview of the topology). A router detecting a failure immediately knows that it has to calculate an alternative route for the LSPs leaving over the dead interface and it may do this based on its current view of the network topology,

stored in its link-state database. The router simply removes the failing equipment from the link-state database and calculates a new route from itself towards the egress LSR: this implies that the part of the LSP upstream from the failure is not rerouted. Explicit routed setup of the LSP (i.e. specifying, in the label requests, each hop to be transited by the LSP) along this calculated alternative path is required (e.g., by means of Constraint Routed – LDP (CR-LDP)), since other routers may not be already aware of the failure. Later on, the IP routing protocol can continue converging/stabilizing and in the meanwhile leave the already restored LSPs alone. The principle of FTCR is illustrated with more detail in [24], [23], [22].

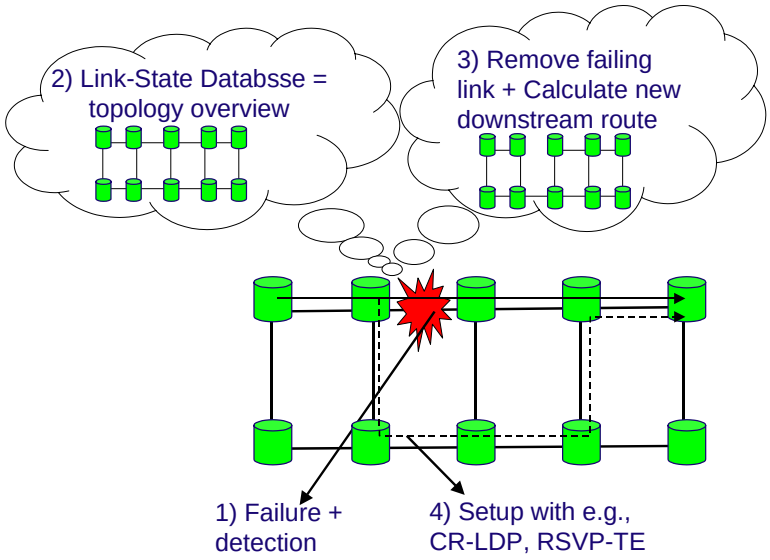


Figure 4: illustrates how an LSR detecting a failure, will reroute in FTCR outgoing LSPs which are affected by that failure. The LSR can compute an alternative route from itself towards the egress LSR based on its own link-state database and setup the LSP by means of explicit routing (e.g., CR-LDP, RSVP-TE) in order to overcome the problem that other LSRs may not be aware of the failure yet.

The fact that MPLS restoration sets up the LSP along the alternative path, at the moment that the failure occurs, requires only standard control plane functionality for the setup and tear down of connections. Even more, this remains true for MPλS (or any circuit-switched technology in G-MPLS). Restoration also allows sharing spare capacity between several failure scenarios.

2.2 Protection in MPLS

Protection in MPLS is based on a pre-established backup LSP. Such a backup LSP can span a single link or node (thus two links, in order to protect also against node failures), or a whole LSP, from ingress to egress. The

former case is called **Local Protection**, the latter **Path Protection**. The upstream LSR, where the backup LSP originates, is called a **Protection** (or Path) **Switch LSR** (PSL) and decides whether data is forwarded along the primary/working LSP or along the backup LSP. The downstream LSR, terminating the backup LSP, is called the **Protection** (or Path) **Merge LSR** (PML) and simply merges both primary and backup LSPs into a single outgoing LSP. This MPLS protection concept is illustrated in Figure 5.

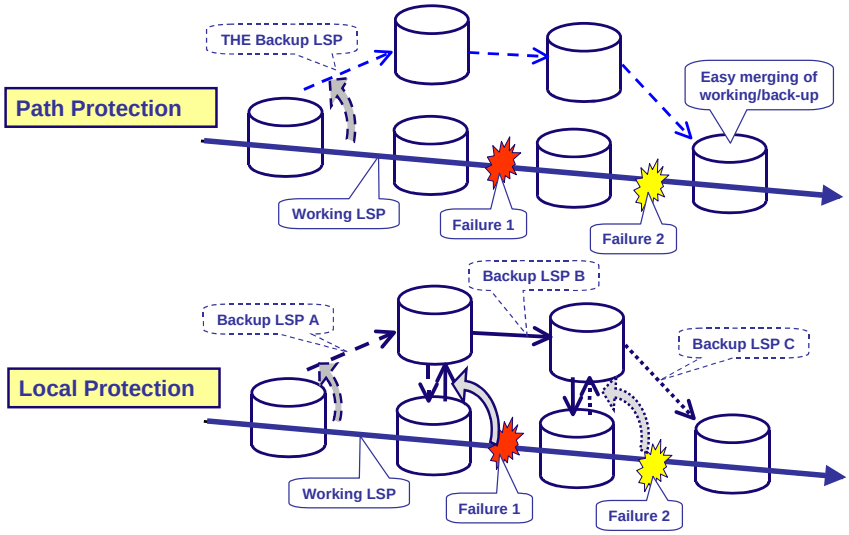
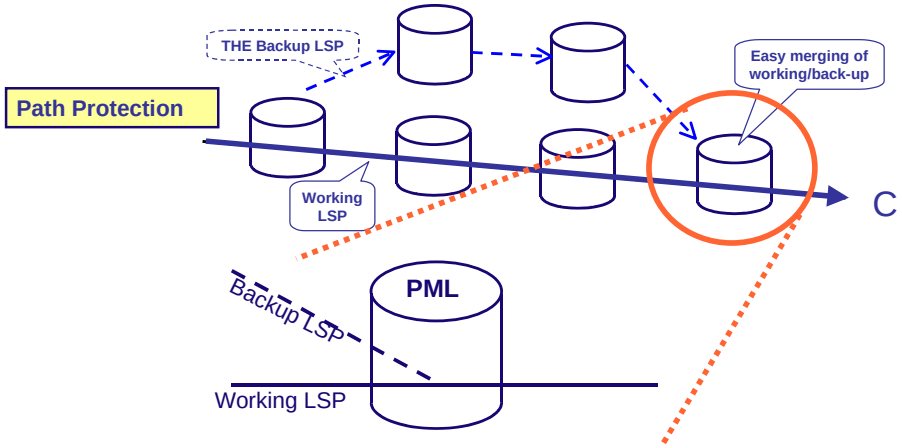


Figure 5: explains Path (top) and Local (bottom) MPLS Protection, under two different failure scenarios. Path Protection always (e.g., during failure 1 and failure 2) switches the traffic in the egress on the single backup LSP. Local Protection needs a backup LSP per link or per node being protected. In case of failure 1, traffic will be routed along backup LSP B, which is pre-established between the end-points of the link affected by failure 1. In a similar way, backup LSP C is used during failure 2.

Figure 6 explains that merging avoids the need for a protection switch in the PML, by simply forwarding any data coming in either through the working or the backup LSP along the outgoing LSP. Remember that IP is connectionless and thus does not require any in-order delivery of packets, even though Label Switched Paths are introduced in MPLS-capable IP-networks.



IN IF	IN LABEL	OUT IF	OUT LABEL
A	1	C	3
B	2	C	3

Figure 6: illustrates how merging of working and backup LSPs is realized. Both incoming LSPs have their own entry in the Label Information Base (LIB) and these entries target the same output interface and label. The router simply forwards any packet coming in through either working or backup LSP.

Local Protection typically suffers from the fact that per link/node a backup LSP is required for each primary LSP. Workarounds (resulting in a single backup LSP per link for all working LSPs over that link) [30] are proposed in case label stacking is allowed and labels have a platform-wide significance. Label stacking is used to multiplex multiple LSPs into a single aggregate LSP: this is achieved by placing an additional label (e.g., shim-header) corresponding to the aggregate LSP, in front of the label of the multiplexed LSPs. Platform-wide label significance means that a label-space exists per LSR instead of per interface. Path Protection on the other hand, suffers from the fact that it cannot perform the protection switch locally, which requires additional signaling functionality and which results in a longer interruption of the affected services or a larger amount of lost data.

The best characteristics of both protection schemes can be combined into another scheme, which we call **Local Loop-Back** (see Figure 7). The idea is that a single backup LSP in the opposite direction of the primary LSP allows performing the protection switch locally. Therefore, the backup LSP consists of two parts: a reverse part, allowing the local protection switch, and a diverse part from the ingress to the egress, in order to get the protected traffic on the backup LSP through the network.

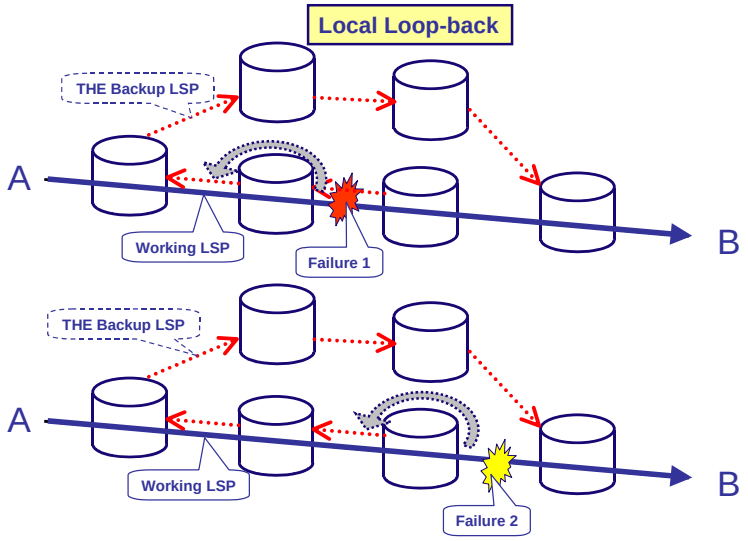


Figure 7: illustrates how the Local Loop-Back technique combines the advantage of Path Protection (single backup LSP) and Local Protection (protection switch performed locally in an LSR adjacent to the failure). The backup LSP is routed in the opposite direction of the working LSP and continues via a disjoint route to the egress LSR. The figure clearly shows that the loop-back is performed in different LSRs (although a single backup LSP is required), under distinct failure conditions: e.g., failure 1 (top) and failure 2 (bottom).

There are two main issues for protection applied to MPλS (or any circuit-switched technology in G-MPLS), as illustrated in Figure 8.

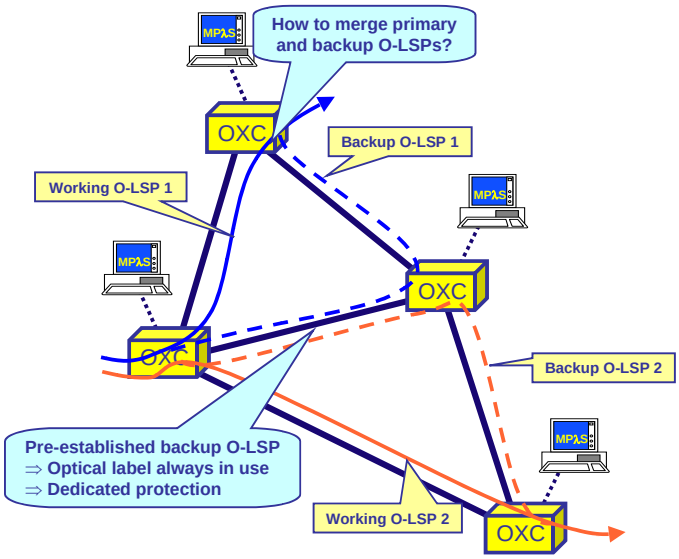


Figure 8: illustrates that there are two main issues in MPλS protection. First of all, working and backup fixed bandwidth O-LSPs have to be merged into a single outgoing O-LSP. Secondly, MPλS protection

results in dedicated protection, due to the fact that each pre-established backup O-LSP always consumes a label (or thus a wavelength), even during failure free conditions.

Merging of multiple circuits into a single outgoing circuit at the same bitrate, is in general not possible. Under certain conditions, specific equipment allows implementing a real protection merge: e.g., passively optical combining of primary and backup signals is allowed. Figure 9 clearly shows that this is only possible if one can assure that backup and primary signals never enter the passive optical combiner at the same time. Unfortunately, this is not always the case: one may opt to send unequipped signals over a link, in order to keep the power budget on that link as constant as possible. Also, signal degradation may trigger upstream a protection switch, while the degraded primary signal is still flowing through the network. To overcome this problem one may prefer to switch from one signal to the other one, as in classical 1+1 Protection. However, this switch has to be synchronized with the status in the Protection Switch LSR.

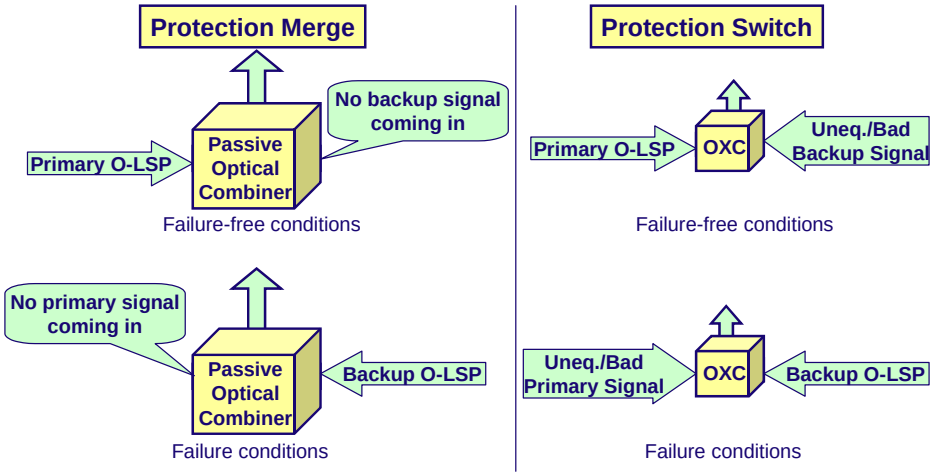


Figure 9: illustrates that a protection merge (left) can be realized by a passive optical combiner, if and only if never a backup and primary signal are received simultaneously. If this condition cannot be met, a protection switch (right) is needed instead of a protection merge.

Figure 8 also shows that pre-established backup LSPs result in dedicated protection, since no statistical multiplexing between circuits is allowed as is the case in packet-switched technologies. Or in other words, a label is always required along a backup LSP, independent whether dealing with a packet- or circuit-switched network, but only in a circuit-switched technology the occupation of a label also implies the occupation of a circuit (which is considered as the capacity in such a network). This is in contrast with packet-switched technologies that allow for statistical multiplexing between LSP routed over the same link. This dedicated

protection implies that MPLS protection in circuit-switched technologies may become far less efficient than in packet-switched technologies, from a capacity point of view.

3. Survivability Issues in Multi-layered Networks

Our roadmap in section 1 shows that data-centric optical-networks typically consist of multiple layers, even in the simplified case of IP-MPLS directly over OTN-MPλS. This section starts with a discussion on the provisioning of recovery functionality in multi-layer networks. These concepts and discussion are focussed on a two-layer network, but are generic and thus applicable to any multi-layer network. This section ends with some survivability considerations specific to IP-MPLS directly over OTN-MPλS networks.

3.1 Single layer survivability strategies and their drawbacks

Section 2 gave an overview of recovery techniques applicable to MPLS or G-MPLS (e.g., MPλS) networks. However, it did not tackle the problem in which layer to apply one of these techniques (e.g. in MPLS or in MPλS for an IP-MPLS/OTN-MPλS network). This section discusses cases where recovery is foreseen at the bottom (e.g., OTN-MPλS) or at the top (e.g., IP-MPLS) layer.

3.1.1 Survivability at the bottom layer

Recovery at the bottom layer has the advantage that a simple root failure has to be treated and that recovery actions are performed on the coarsest granularity, resulting in the lowest number of required recovery actions. Also failures do not need to propagate through multiple layers before triggering any recovery action.

However, there is no recovery scheme residing in the bottom layer that can resolve any problems due to a failure in a higher layer: any layer above or the layer where the failure occurs itself has to resolve the problem. Figure 10 shows also that in the case of a node failure in the bottom layer, this layer can only recover affected traffic transiting this failing bottom layer node. The co-located higher layer node becomes isolated and thus all traffic transiting such a higher layer node cannot be restored in the bottom layer.

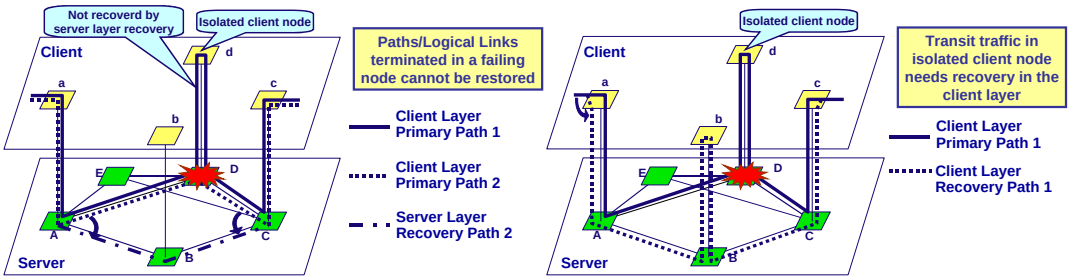


Figure 10: shows the impact of a node failure on two traffic flows between the client layer nodes a and c. The left side of the figure illustrates that the logical links a-d and d-c (both terminating in node D) cannot be restored by the server layer, resulting in the isolation of client layer node d. This implies that the first flow a-d-c (transitting this isolated node d) cannot be recovered by the server layer, but that the client layer has to recover this flow, as illustrated by the right side of the figure. The second flow is routed over a direct logical link between node a and c. This logical tranists the failing node D and thus can be restored by the server layer recovery scheme.

3.1.2 Survivability at the top layer

Another strategy is to provide the **survivability at the top layer**. The advantage of this strategy is that it can cope more easily with node or higher layer failures (see Figure 10). A main drawback of this strategy is that it needs many recovery actions, due to the finer granularity of the flow entities in the top layer. However, treating each individual flow at the top layer allows differentiating between these flows, based on their (service) importance. Or in other words, the top layer may restore critical, high priority traffic before any action is taken on low priority flows. This is not possible in lower layers, since they switch every flow in an aggregate signal with a single action. Under certain conditions, the finer granularity may also lead to a more efficient capacity usage. First of all, aggregate signals, poorly filled with working traffic, have enough capacity to transport spare resources. Secondly, the finer granularity allows distributing flows over more alternative paths. However, a trade-off exists between a better filling of the capacity of the logical links and the higher amount of higher layer equipment, when comparing this survivability at the top layer strategy with the survivability at the bottom layer strategy.

Not only the potential mismatch in granularity between the failing equipment in a lower layer and the thereby affected entities in the top layer, requiring more recovery actions, is an issue. Also the typically complex

secondary failure scenarios, as a result of a single root failure in a lower layer, can become a problem. This is illustrated in Figure 11.

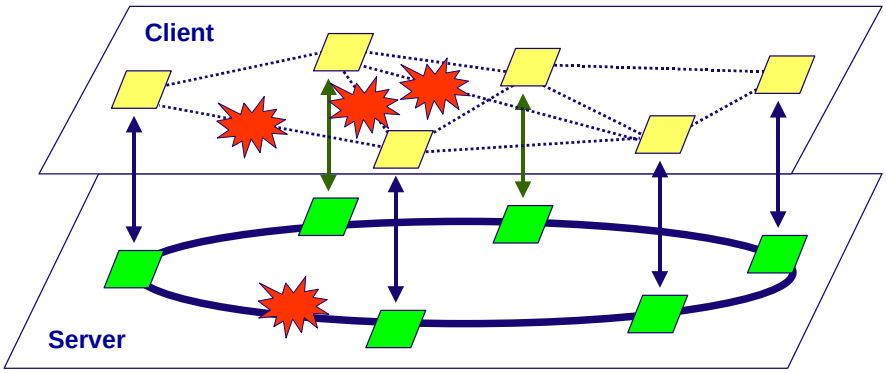


Figure 11: explains that a single root failure may propage to many so-called secondary failures.

3.1.3 Slightly different variants: survivability at the lowest detecting layer and survivability at the highest-possible layer

A slightly different variant on the survivability at the bottom layer is the **survivability at the lowest detecting layer** strategy (i.e. the lowest layer in the hierarchy able to detect the failure). This means that multiple layers deploy a recovery scheme, but that still the (single) layer detecting the **root** failure is the only layer taking any recovery actions. With this strategy, there is no problem anymore that the bottom layer recovery scheme does not detect a higher layer failure (because the higher layer which detects the failure will recover the affected traffic). However, this survivability at the lowest-detecting layer strategy can assure that traffic transiting the failing equipment is restored, but it still suffers from the fact that it cannot restore any traffic transiting higher layer equipment isolated by a node failure. The client layer in Figure 10 deploys a recovery scheme in this strategy, but the considered traffic flow is still lost, since this client layer recovery scheme is not triggered by the node failure in the server layer. This strategy is considered as single layer survivability strategy, although it considers the deployment of a recovery scheme in multiple layers. The reason is that for each failure scenario the responsibility to recover all traffic is situated in one and only one layer (the one detecting the failure).

A slightly different variant of the survivability at the top layer strategy is the **survivability at the highest possible layer** strategy. Since not all traffic has to be injected (by the customer) at the top layer, a traffic flow is recovered in the layer in which it is injected (or in other words the highest possible layer for this traffic flow). For example, a data-centric optical network may also support a leased optical channel service. This strategy is also considered as a single layer survivability strategy, although it considers a recovery scheme in multiple

layers. Indeed, survivability at the highest possible layer may lead to recovery schemes in multiple layers, but never to recover the same traffic flow. Actually, for each traffic flow a survivability at the top layer strategy is deployed (or in other words, both strategies do not differ in essence from each other).

3.2 Multi-layer survivability: concepts and solutions

The conclusion from the previous section is that both survivability at the bottom/lowest detecting and top/highest possible layer have their pros and contras. However, it is likely that a real network will combine the advantages of both approaches. Or in more general, that the choice in which layer to recover the traffic will depend on the circumstances (e.g., the occurring failure scenario). This requires a higher flexibility than the simple rules on which the single layer survivability strategies are based (always all recovery actions in the lowest (i.e., lowest detecting/bottom) layer or always in the highest (i.e., highest possible/top) layer).

3.2.1 Uncoordinated approach

A first solution is to deploy **a recovery scheme in multiple layers, without any coordination**, resulting in parallel recovery actions at distinct layers. Consider for example the link failure in Figure 12. The considered traffic flow a-c is affected and thus restored in the client layer (path a-d-c replaced by path a-b-c), while the server layer is restoring the logical link a-d (of the client layer topology), by rerouting it via node E.

The main advantage is that this solution is simple from an implementation (e.g., no standardization of coordination signals between both layers is necessary) and operational point of view. However, Figure 12 shows the drawback of this strategy. Both recovery mechanisms occupy spare resources during the failure (i.e., the server layer along A-E-D and the client layer along a-b-c, which implies occupation of spare resources on A-B and B-C in the server layer), although one scheme occupying spare resources would be sufficient. This implies that potentially more extra traffic (i.e. unprotected pre-emptable traffic) is squelched (disrupted). Or even worse, consider that the server layer reroutes the logical link a-d over the path A-B-C-D instead of A-E-D, then both recovery mechanisms need spare capacity on the links A-B and B-C. If these higher layer spare resources are supported as extra traffic in the lower layer, then there is a risk that this client layer spare resources are pre-empted by the recovery action in the server layer, resulting in “destructive interference”. Or rephrased, the two recovery actions taken were not able to restore the traffic, since the client layer reroutes the considered flow over the path a-b-c, which was disrupted by the server layer recovery. [16] illustrates that these

risks may exist in real networks: they prove that a switch-over in the optical domain (e.g., for protection purposes in the optical network) may trigger traditional SDH protection.

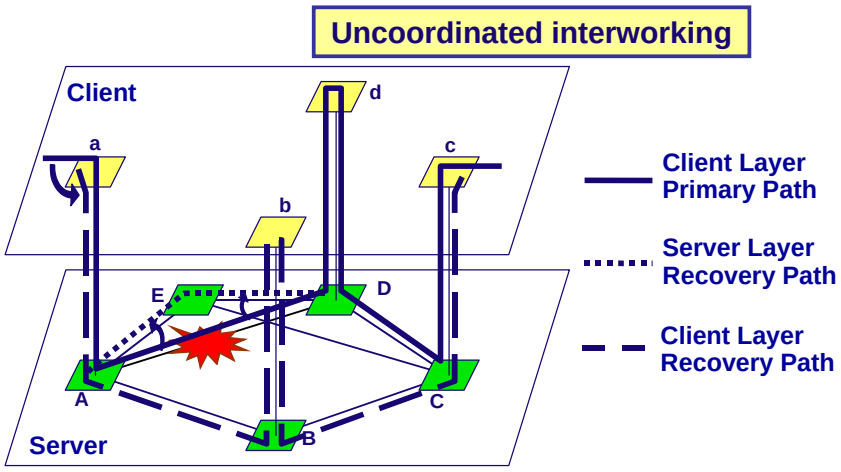


Figure 12: illustrates the uncoordinated multilayer survivability strategy. The failure of the physical link A-D in the server layer, also affects the corresponding logical link a-d in the client layer. Since recovery actions in both layers are not coordinated they will both recover the affected traffic. The server layer reroutes all traffic on the failing link A-D through node E. The client layer restores the connection end-to-end, by routing it along the path a-b-c. It is clear that in this example recovery actions in a single layer would have been sufficient.

3.2.2 Sequential approach

A more intelligent approach, compared to the uncoordinated approach, is the sequential approach, where the responsibility for recovery is handed over to the next layer when it is clear that the current layer is not able to fulfill the recovery task. There exist mainly two approaches:

1. **Bottom-up approach**: the recovery starts in the bottom/lowest detecting layer (where the failure is detected) and all traffic which cannot be restored by this layer (e.g., due to capacity shortage), will be restored by a higher layer. The advantage of this approach is that recovery actions are taken at the appropriate granularity (recovery actions on a finer granularity, in a higher layer, are only taken when necessary) and complex secondary failures are treated only when needed.
2. **Top-down approach**: is the other way around. Recovery actions are initiated in the top/highest possible layer and only if the higher layer cannot restore all traffic, lower layer actions are triggered. An advantage of this approach is that a higher layer can more easily differentiate traffic with respect to the service types

and thus it may try to restore high priority traffic first. A drawback of this approach is that a lower layer has no easy way to detect on its own, whether a higher layer was able to restore traffic or not (an explicit signal is needed for this purpose).

The remainder of the paper assumes the bottom-up approach (since this is the most intuitive one), except when explicitly referring to the top-down approach. An example of the bottom-up approach is shown in Figure 13. The server layer starts with attempting to restore the logical link a-d, but it fails since this logical link terminates on the failing node D. Therefore, the client layer recover scheme is triggered to restore the considered traffic flow a-c, by rerouting it over node b instead of node d.

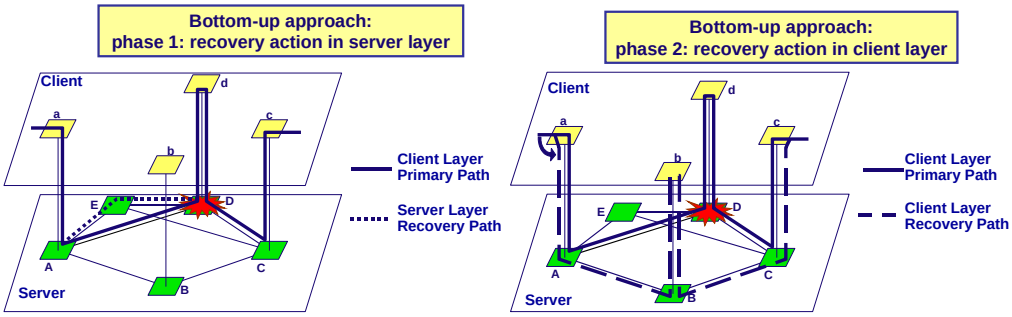


Figure 13: presents the bottom-up approach. In this approach the server layer begins trying to recover the traffic as much as possible. The logical links a-d and d-c of the client layer terminate on node D (which is failing) and thus the server layer cannot restore the traffic carried on these links. Therefore, the recovery scheme in the client layer will be triggered in one or another way. This scheme will recover the traffic transiting the isolated node d.

The implementation of these escalation strategies (i.e., handing-over the responsibility for recovery from one layer to the other one) is another issue. Two solutions are described here.

- The first one is based on a **hold-off timer.** A hold-off timer is set at the moment the server layer starts attempting to restore the traffic. If this hold-off timer goes off and (part of) the traffic is not restored, then the client layer will take over the recovery actions while the server layer ceases its attempts. The main drawback of a hold-off timer is that higher layer recovery actions are always delayed, independent of the failure scenario.
- To overcome this delay, another escalation strategy is the use of a **recovery token signal** between layers. This means practically that the server layer sends the recovery token (by means of an explicit signal) to the client layer from the moment that it knows that it cannot restore traffic anymore. A disadvantage, compared

to a hold-off timer interworking, is that a recovery token signal needs to be incorporated in the standardization of the interface between network layers.

A hold-off timer is probably less appropriate for a top-down approach, since the lower layer should be notified with an explicit signal whether the higher layer managed to restore the traffic or not.

3.2.3 Integrated approach

The **integrated approach** is based on a single integrated multi-layer recovery scheme. This implies that this recovery scheme has a full overview of all the network layers and that it can decide when and in which layer (or layers) to take the appropriate recovery actions. It is obvious that an integrated approach is the most flexible one. However, to profit from this high flexibility, one has to provide the necessary algorithmic intelligence/complexity. Another issue is the implementation/realization of such an integrated approach. It is unlikely to develop a single recovery scheme, controlling and having an overview of all network layers, in current overlaid networks.

3.3 Summary and conclusions

Section 3.1 discussed the shortcomings of single layer survivability strategies. Section 3.2 illustrated how to overcome these shortcomings by providing survivability at multiple layers. Table 1 gives a summary of the estimated performance, with respect to several characteristics, for some survivability strategies.

Table 1: compares and summarizes several performance parameters for some significant recovery strategies. The last column gives the typically (but not necessary) the preferred value for each parameter.

	Survivability Strategy				
Criteria	**Bottom layer**	**Bottom-up**	**Top layer**	**Integrated approach**	**Preferred value**
Switching granularity	Coarse	Coarse	Fine	Coarse	**Coarse**
Failure scenario	Simple	Simple	Complex	Simple	**Simple**
Recovery close to root	Yes	Yes	No	Yes	**Yes**
Capabilities, flexibility	Low	High	High	High	**High**
Failure coverage	Low	High	High	High	**High**
Co-ordination, mgmt	Low	High	Low	Low	**Low**
Resources	Low	High	Low	Low/High	**Low**

[31], [32] illustrates that the spare resource requirements can be reduced for the case of multi-layer survivability, by supporting higher layer spare resources as extra traffic in the lower layer spare resources (i.e., the **common pool** of spare resources). However, section 3.2.1 explained that a proper coordination of the recovery schemes becomes absolutely necessary in such a case.

3.4 Specific IP-MPLS/OTN-MPλS opportunities and drawbacks

The goal of this section is to highlight some specific survivability opportunities and drawbacks that arise in case of an IP-MPLS directly over OTN-MPλS network. Note that the previous sections, on generic multi-layer survivability strategies, remain true for IP-MPLS/OTN-MPλS multi-layer networks: this section only provides some additional considerations, which may be taken into account when designing such an IP-MPLS/OTN-MPλS network.

Section 2 illustrated that MPLS is suitable to provide fast protection switching in the IP-MPLS layer. Therefore, one could opt to promote recovery in the IP-MPLS layer (i.e., promote survivability at the top/highest possible layer (e.g., [33]) or a top-down strategy) as this has some favourable properties. First of all, less spare resources are needed in the IP-MPLS layer, since packet-switching is very suitable to **share spare capacity** amongst pre-established backup paths (while keeping the advantages of fast protection switching). Secondly, **dropping low-priority** (e.g., best-effort) traffic first is inherently incorporated in IP-MPLS networks, if for example Diffserv is deployed [34].

Another opportunity relates to the integrated approach, mentioned in section 3.2.3. As described in our roadmap in section 1, we expect that a peer-modeled data-centric optical network may become a reality in a longer-term future. If this becomes true, then a **single integrated multi-layer approach would become much more feasible than in current overlaid networks,** due to the single integrated control plane of a peer-modeled network.

Finally, the automation of the lightpath setup/tear-down process in an Automatically Switched Optical Transport Network (ASON) doesn't require anymore to stick with a fixed logical (IP-MPLS) topology and capacity. This opens opportunities for the re-optimization of the logical topology during a failure condition. Even more, an at least node bi-connected logical (IP-MPLS) topology is no absolute necessity anymore to survive any single failure. For example, if a router would fail (potentially resulting in a disconnected IP-MPLS network), an automatic reconfiguration of the logical IP-MPLS topology (instead of traditional rerouting (i.e., protection/restoration) of traffic) would restore the connectivity of the IP-MPLS network.

A main drawback of current IP-MPLS network is that failure detection is based on the periodic exchange of HELLO-messages between adjacent routers. If no HELLOs are received anymore through an interface, then the only conclusion can be that the opposite side of the interface is unreachable or in other words that each packet sent through the interface is sent into a black hole. But this detection scheme does not allow to differentiate

between a router failure (meaning that the router at the opposite side of the link is dead) in the IP-MPLS layer itself and a failing logical link in the IP-MPLS layer, as a result of a failure in the OTN-MPλS layer. This implies that the survivability at the lowest detecting layer is impossible in a IP-MPLS/OTN-MPλS network.

Another concern of this HELLO message detection scheme is the detection time. Current IP routers send a HELLO message each 10 seconds, and a defect is declared after the loss of 4 HELLO messages (resulting in a typical detection time of 40 seconds) [35]. However, driving this periodicity to the other of milliseconds becomes reasonable in IP-MPLS/OTN-MPλS, due to the huge capacity (e.g., 10 Gbps) of a logical link, resulting in a insignificant bandwidth overhead for the HELLO messages.

4. Case studies on Survivability in IP-MPLS directly over OTN-MPλS networks

The goal of this section is to present some case study results, which deal with survivability in data-centric optical networks. First typical network scenarios are described. Then two studies are presented, which may help in deciding in which layer (IP-MPLS or OTN-MPλS) to provide survivability functionality. The first study compares the cost of MPLS protection whether deployed in the electrical IP-MPLS or optical OTN-MPλS layer. The second one studies the influence of protection switching, and its timing, on TCP behaviour (which is typical for data traffic). The section ends with the design of a sample network that may or may not take into account IP-MPLS router failures.

4.1 Typical network scenarios

A typical IP-MPLS network consists of a meshed core network containing a few dozens of backbone IP-MPLS routers. Attached to those backbone routers are regional networks that concentrate the traffic from the access part of the network into the core part. While the core part of the network has a meshed structure, the structure of the access part of the network could be described as tree structures, as illustrated in Figure 14. Also attached to the IP-MPLS network are large server farms, containing the data for e.g. video-on-demand or web-based services. They are one of the reasons of the highly asymmetric character of IP traffic (e.g., video-on-demand: small customer request stream in the upstream direction, large video-data stream in the downstream direction) [36].

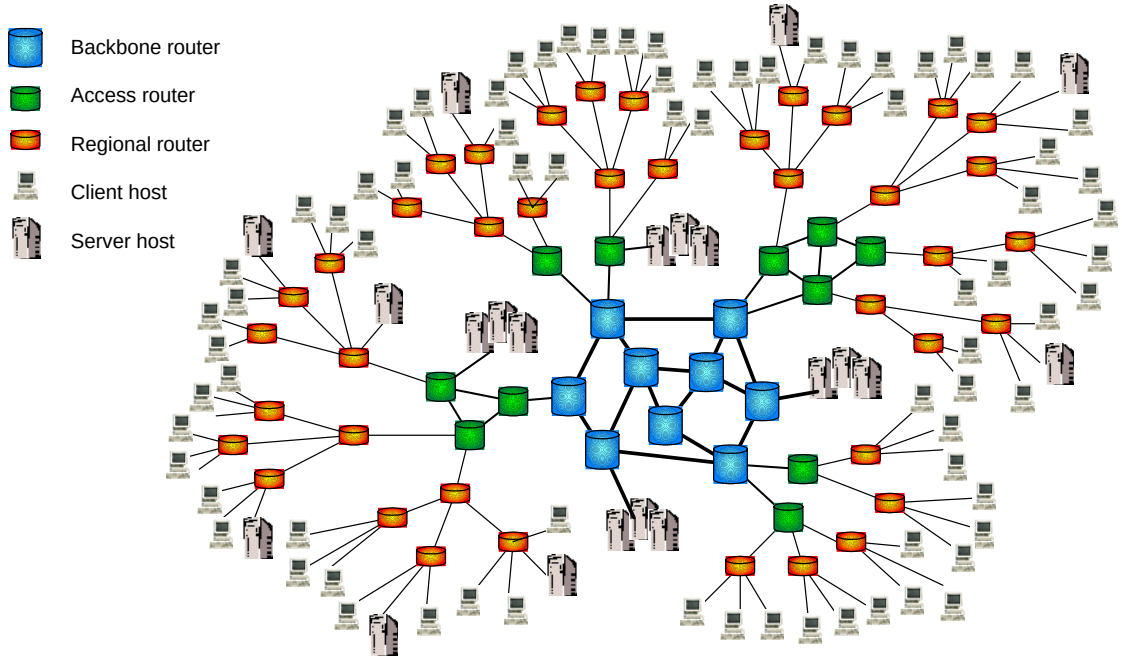


Figure 14: Typical IP-MPLS network topology (Backbone + access part)

In the IP-MPLS directly over OTN-MPλS scenario considered in this paper, the logical (backbone) IP-MPLS links are directly supported by optical paths in the OTN-MPλS layer. However, various routing options still exist, especially in the backbone part of the network. Some operators will probably have a single-hop IP-MPLS core network where traffic is routed through only two backbone routers: one through which it enters the backbone network, and one through which it leaves the backbone. This implies of course that the backbone part of the network is a full mesh on the logical IP-MPLS level. Other operators might have a multi-hop network in which the IP traffic traverses several logical links (hops) before it leaves the backbone. Since LSPs will typically start and/or terminate somewhere in the access part of the network (or even at a host), most LSPs will pass through multiple routers (even in the case of a single-hop logical core network).

4.2 Recovery at MPLS or/and at the MPλS layer?

An important issue in this whole paper is in which layer to provide a recovery scheme. The goal of this section is to present some quantitative study results, which may help to answer this question.

A first study investigates the amount of required spare resources, relative to the amount of working resources. The previous section and section 2.2 explained that MPLS protection results in shared protection when applied at the electrical MPLS layer, and in dedicated protection at the optical layer. The goal of our study is to investigate the significance of this effect, by comparing the results for both cases. Figure 15 makes such a

comparison between both relative values, for all MPLS protection schemes described in section 2.2, for two topologies. The LARGE topology contains 44 nodes, interconnected by 57 links, resulting in an average nodal degree of 2.59. The SMALL topology contains 30 nodes, interconnected by 36 links, resulting in an average nodal degree of 2.4. The values presented in the charts are an average over 10 randomly generated traffic matrices. The routing strategy is as follows. Firstly, the working route is calculated based on a Dijkstra shortest-path algorithm. Subsequently, the shortest node-disjoint route is computed for Path Protection and Local Loop-Back. It can happen that such a route is not found, which implies that traffic is lost (or not protected) during a failure by both schemes. Local Protection is based on backup paths spanning two links, in order to be able to protect also against node failures. There is only one exception: a backup path is also spanned over the last link of each connection, since it would make no sense to send the traffic one hop behind the termination node, in case the last link would fail.

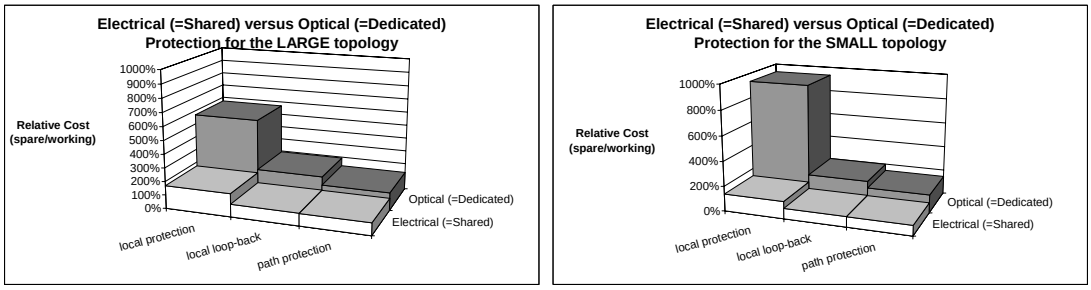


Figure 15: compares electrical and optical MPLS protection, from a capacity point of view. The presented study confirms that: 1) optical protection is more expensive (due to dedication of spare resources), 2) this is most drastic for Local Protection, which 3) is the most expensive scheme anyway.

Figure 15 indeed confirms our expectation that dedicated MPLS protection (thus in the optical layer) is more expensive than shared MPLS protection (in the electrical MPLS layer). More important is that these charts show that the difference is severe for Local Protection. This result can be sensed as follows. Sharing between two (or more) backup paths using a same resource is only possible if the two corresponding working segments (segment is a path in case of Path Protection or Local Loop-Back, one link in case of Link Protection, and two links in case of Node Protection) do not overlap. In case of Local Protection these working segments are in general shorter than for Path Protection or Local Loop-Back (one or two links versus a complete path), implying a smaller probability of working segments overlapping and hence a higher probability that sharing between the two backup paths is indeed allowed. Hence, the relative difference between dedicated and shared protection in

terms of capacity requirements will be more substantial for Local Protection than for Path Protection or Local Loop-Back.

Figure 15 also reveals that (in case of Local Protection) the topology has a significant impact and more precisely that the topology with the smallest nodal degree suffers the most from this dedication. This result can be understood intuitively as follows. If a topology becomes sparser, backup LSPs for adjacent failure scenarios (e.g., two adjacent links in case of link protection) tend to become longer and more overlapping (for instance, think about the extreme case of a ring topology to sense this). This explains why the penalty of dedication is severe in sparse networks, less in dense networks. These observations are confirmed by the study in Figure 16 investigating the impact of the (nodal degree of a) topology on the relative cost increase due to the dedication of MPLS protection in the optical domain. The conclusion is that fast MPLS protection in the electrical MPLS layer is cheaper than similar schemes in the optical transport network and that the cost increase for Local Protection in the optical layer could be very severe, due to the typical sparse topologies of transport networks.

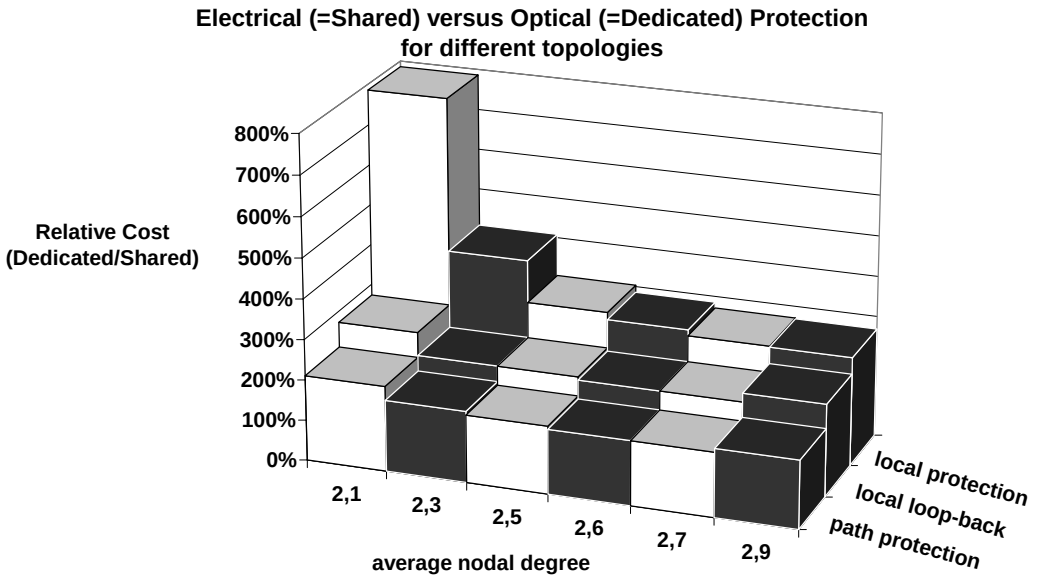


Figure 16: illustrates the increase of amount of required spare resources, due to dedication, for a set of topologies with different degrees of meshedness. The figure illustrates that this is most drastic for Local Protection, especially on sparse topologies (which are unfortunately typical for optical transport networks).

Another issue is whether the dominant data traffic (typically based on TCP) prefers fast protection switching. Assume that one wants to profit from the advantages of fast protection switching in the electrical MPLS layer.

Then, there may be a risk that switching a large amount of traffic (e.g., a complete 10Gbps line) immediately (i.e., before the TCP-mechanism gets the chance to slow down) would drastically impact other flows in the network. Indeed, as TCP is reactive in nature, not only the flows being switched to an alternative backup path will be affected, but also the other flows (already present on (parts of) the backup path). To gain a better understanding of these kind of interactions, and the role of the exact timing of the protection switch, a simulation study was carried out.

The setup of the simulation is depicted in Figure 17. We consider a backbone network of Label Switched Routers (LSRs) to which we connect access nodes via links having a bandwidth that is 90% of the backbone links. In the thus created network we set up two categories of flows. The so-called "switched flows" will follow the path crossing LSRs 4, 5, 6 and 7 when there is no link failure; upon the failure of link 5-6, a protection switch will be carried out at LSR 5 and the followed path will be 4-5-9-10-6-7, as indicated by the dotted line in Figure 17. The other category, the "fixed flows", will always use the path over LSRs 8, 9, 10 and 11. The simulation scenario consists of three periods of 5 seconds: during the first and third, all links will be up, whereas during the second period link 5-6 will fail. To investigate the influence of timing, the protection switch will be performed "manually" exactly δ seconds after the occurrence of the link failure.

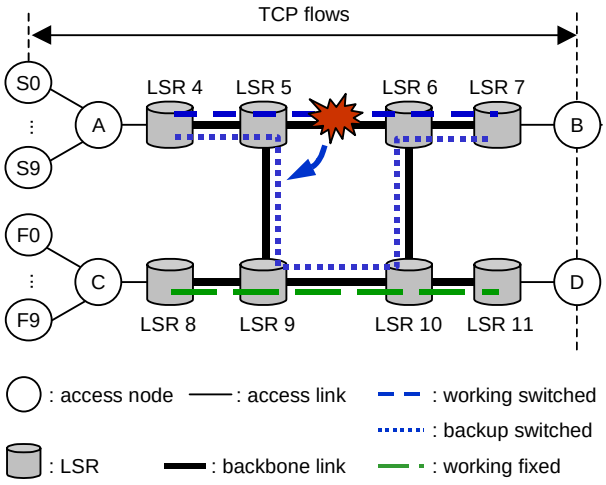


Figure 17: Simulation topology investigating effect of protection switching on TCP

From a qualitative point of view, the influence of δ can be easily predicted. If δ is set to zero, the switched flows will join the fixed ones at LSR 9 at a time when they are both sending at a quite high rate (limited only by the bandwidth of the access links). This will result in an almost immediate buffer overflow at LSR 9, causing a burst of a fairly high number of losses, afflicted on both flow categories. Introducing a small delay (δ strictly positive) will inflict losses during that period of δ on the switched flows only, thereby forcing them to back off

(cf. TCP window size reduction in response to losses) before being switched to the alternative path. As a result, the immediate buffer overflow at LSR 9 will be avoided and the fixed flows will be approached more "gently": a buffer overflow at LSR 9 will occur at a later time, and will cause fewer losses compared to the δ=0 case. In Figure 18, the evolution of goodput over time is depicted. There we clearly see the heavy impact (i.e. serious drop in goodput) of the immediate buffer overflow for δ=0 on the fixed flows.

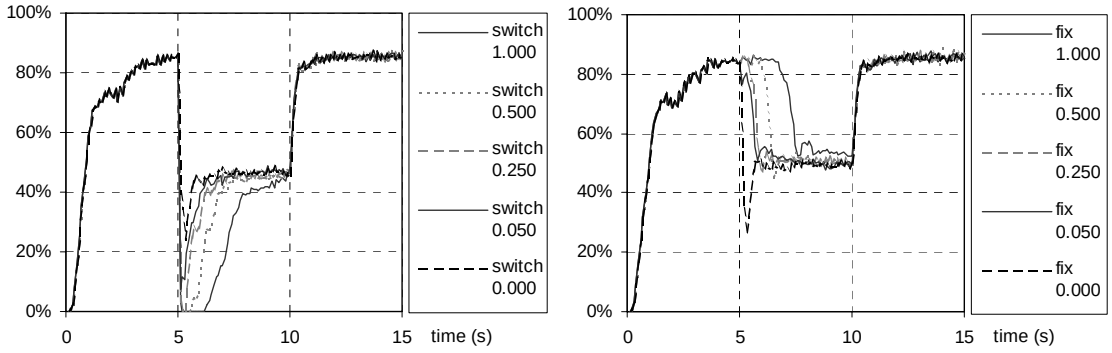


Figure 18: TCP goodput evolution over time for the different values of δ. The left part shows the goodput attained by the whole of the switched flows, whereas the right contains the evolution for the fixed flows. The goodput is expressed in % of backbone link bandwidth and was measured with a resolution of 10 ms.

To decide what delay δ results in the "best" behaviour from a quantitive point of view, we decided to use TCP goodput as a decision criterion. Indeed, goodput is what an end user cares about: it is the amount of data successfully transported end-to-end during a certain time interval (expressed in e.g. bytes/s). We ran simulations using random start times for the TCP sources, and randomly generated propagation delays for the first access links (in order to introduce diverse Round Trip Times or (RTTs) for different source-destination pairs). For each of the thus created 150 random cases, we ran simulations for five different values of δ (0, 50, 250, 500 and 1000ms) tracing TCP goodput. We compared the different values of δ by plotting the histogram of the ratio f(δ)=Good(δ)/Good(0), where Good(δ) is the total goodput – attained by the whole of fixed and switched flows – during the first 1.5 seconds after the link failure for delay δ (we chose 1.5s as we intended to focus on the smaller delays, and this is the relevant period for those cases). These histograms (and corresponding normal fits) are depicted in Figure 19. That graph shows that, on average, all cases of δ result in a better overall goodput than having no delay at all (δ=0).

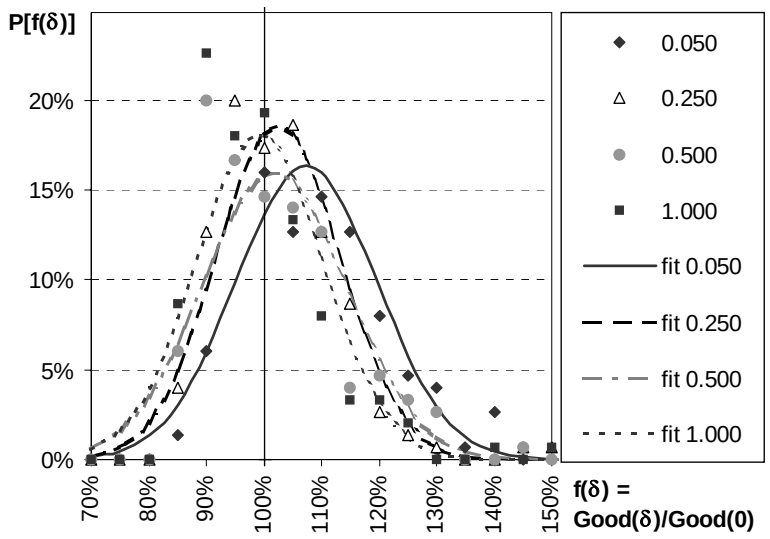


Figure 19: Histograms (with a resolution of 5%) and normal fits for relative amount of goodput. A marker at (x,y) for a particular δ means that y% of the simulation results had f(δ) within [x,x+5%[.

The measurements of Figure 19 are interpreted and summarized in Table 2.

Table 2: Comparison of different protection switch delays. The left column represents the x-value corresponding to the average of f(δ), i.e. the peak of the normal fit in Figure 19. The second column indicates the percentage of simulation results where f(δ) < 100%, whereas the rightmost column gives the number of simulation results where Good(δ) was maximal (i.e. compared to other delays).

δ	average better goodput than case δ=0	% of random cases where δ is worse than δ=0	% of random cases where delay δ is best
0.000	0.00%	0.00%	20.00%
0.050	9.85%	24.00%	64.67%
0.250	4.99%	36.67%	9.33%
0.500	4.39%	42.67%	5.33%
1.000	1.75%	49.33%	0.67%

An important observation that can be drawn from these simulation results is that the time it takes for the interacting TCP flows to stabilize after the protection switch, is in the order of a second or more (see Figure 18). It can be concluded that pushing fast protection switching to the limit (i.e. extremely fast) may not be the best thing to do. However, to decide upon the "best" time to perform the protection switch, is not easy. It depends at least on the link load (in the case presented above, when all links are up, backbone links are loaded for max. 90% due to the limits in the access part, but a protection switch results in a sudden load of almost 180%), the

RTT experienced by the TCP sources (larger RTT means slower response to topology changes), the number of concurrent TCP flows (larger number results in faster stabilisation, up to a certain limit).

However, the results presented above seem to indicate that from a practical point of view, it is not harmful to have fast protection (order of tens of milliseconds) for TCP traffic. This conclusion is probably even more true if we believe that backbone links carry a vast amount of concurrent TCP flows (cf. faster stabilisation than small number of flows, and therefore optimal delay shifts towards $\delta=0$) and/or are fairly underloaded. Indeed, when backbone links do not form the bottleneck for TCP flows, interaction between switched and fixed flows will be limited. Other simulations showed that in this latter case (e.g. for an access link bandwidth being 60% of the backbone bandwidth), the optimal protection switch delay clearly shifts to lower values (towards $\delta=0$). The simulations carried out so far, seem to indicate that only if the timescale of protection switching is well below 50ms, TCP effects may call for a stop to the efforts to minimise it. All this however does not imply that extremely fast protection switching is a must for TCP: the differences in goodput for delays in the range 0-250ms do not differ all that much, especially when the number of TCP flows is large.

The simulation discussed above considered fast protection at the MPLS layer. However, if fast protection is offered by lower layers (e.g. MPλS), we are in an altogether different situation. Indeed, in that case we will have no interaction between competing TCP flows (as we assume that the capacity for protection is reserved, and is fully available from the very instant the protection switch is carried out).; clearly, dynamic behavior of TCP in response to packet losses will still occur. In this case, the intuitively clear conclusion we have drawn from a first series of simulations is: the faster the protection switch at the optical layer is performed, the better (from a TCP goodput point of view). The simulations performed for this case had a link going down for a certain amount of time δ, without any protection actions taken at the MPLS level. For 140 random cases (random RTTs, etc., as before) and δ in {0, 5, 10, 20, 30, 40, 50, 250, 500, 1000 ms} we saw that in 94% of the cases, $\delta=0$ was the best (only packets in transit on failing link are lost); in the remaining 6% of the cases, $\delta=5$ms was the best (which is due to details in dynamic TCP behaviour in some rather peculiar cases). Thus, the avoidance of TCP interactions is an advantage of protection at the MPλS layer and means that even extremely fast protection switching at that layer does not seem to pose any problem (at least from TCP point of view).

We can conclude this section, by saying that from a capacity point of view protection in the MPLS layer is preferable compared to MPλS protection. However one has to be careful when performing fast protection

switching in the MPLS layer, since TCP may behave in such a way that its goodput slightly reduces when switching too fast. Thus this section illustrates that such a decision is far away of being straightforward.

4.3 Case Study: Design of a Multi-layer Survivable MPLS/OTN Network

The concept of survivability in a multi-layer network is illustrated here with an example. The network under study is an MPLS over OTN network [37]. Both layer networks are shown in Figure 20. The MPLS layer contains 16 routers, connected by 33 logical links. Attached to the routers of the major cities are servers that contain the application data (e.g., video data for the video-on demand service). The topology resembles a multiple star topology, with the heart of each star in a router connected to a large server (farm). The OTN layer is made up of 14 OXCs and 29 links, in a mesh topology. Both topologies are bi-connected.

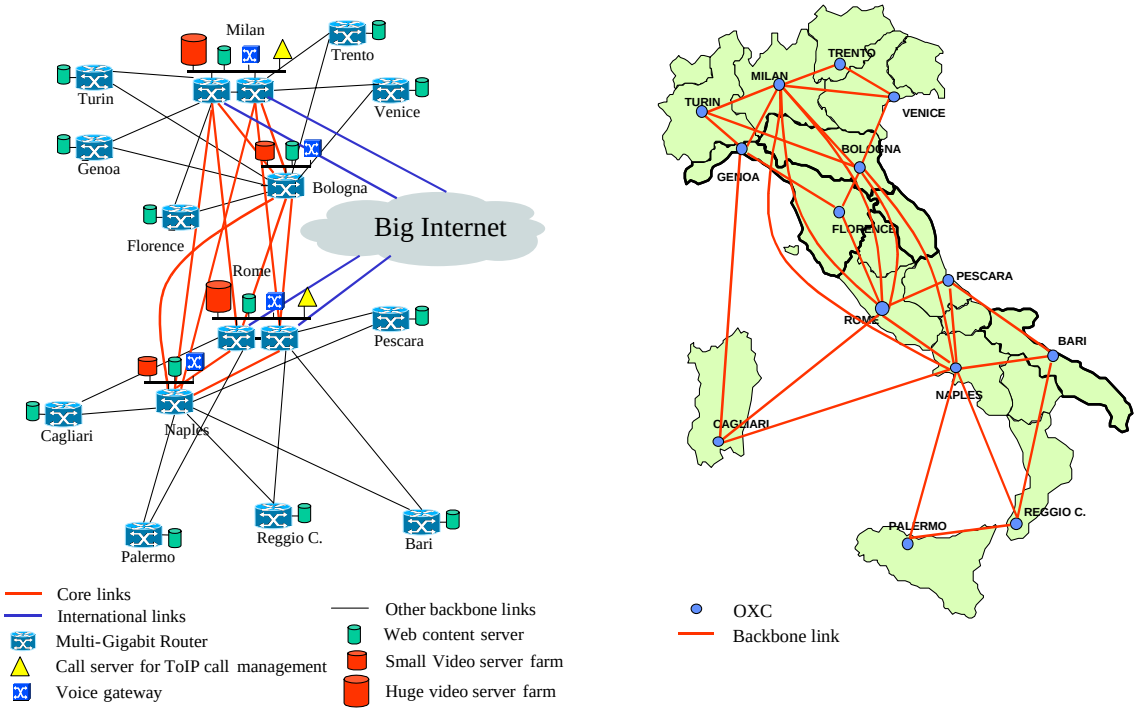


Figure 20: MPLS topology (left) and OTN topology (right)

Starting point of the design is the IP traffic matrix, which combines the demands of the various IP services (voice-over-IP, video-on-demand, web-based services, e-mail, etc). This matrix is asymmetric as some of the routers are connected to server farms and thus generating large amounts of traffic, that is downloaded by users scattered all over the country. Based on the IP traffic demand and the MPLS topology, the MPLS layer is dimensioned, using an MPLS-based planning tool. It routes the unidirectional IP traffic along the shortest path between its source and destination. Because the individual unidirectional flows are routed one at a time, it is

possible that both directions between the same <source, destination> pair are routed along different paths (with equal lengths). In order to provide recovery for LSR failures (or any other failure isolating an LSR), the network can be dimensioned for MPLS local Protection (see Section 2.2).

The MPLS dimensioning tool gives as output the routing of the traffic (on each link) and thus the capacity needed on the MPLS links. These are fed into the OTN planning tool together with the OTN topology. The maximum capacity on both directions of a logical link is considered as the capacity needed on that link. Or in other words the number of bidirectional lightpaths to be setup between two LSRs.

Line-systems of 32 wavelengths were assumed, with each wavelength carrying an STM-16 signal. The routing in the OTN layer starts from an initial shortest path routing and tries to remove inefficiently used line-systems by rerouting the traffic along other line-systems which have enough unused capacity left. The tool can calculate the spare resources needed for different recovery schemes: no protection, Link or Path Restoration and 1+1 Protection [38] In our design, the OTN layer was chosen to provide resilience against expected failures (this includes single link and node failures). However, as described in Section 3.1.1, a recovery scheme in the OTN layer alone doesn't suffice to provide resilience against MPLS router failures (or any other failure isolating a router). An appropriate recovery scheme in the MPLS layer (e.g. MPLS Local Protection) is needed. This will result in an increase of the overall cost, because extra capacity in the OTN is needed to support the spare resources of the MPLS layer. Figure 21 shows a comparison in terms of cost between the various possible recovery schemes in the OTN, with and without the use of MPLS Local Protection in the MPLS layer.

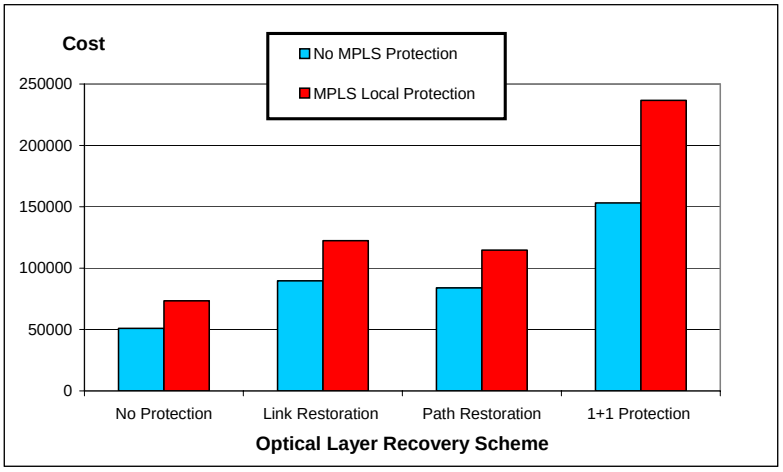


Figure 21: presents the overall network cost for different resilience strategies. Per recovery technique in the optical layer two values are given: one for the case with and one for the case without Local MPLS Protection against router failures.

The cost is modeled as the sum of the number of wavelengths needed on the various links multiplied by the link length. Important here is also the assumption that all OXCs are able to perform wavelength conversion.

A first conclusion that can be drawn from these results is that the use of 1+1 Protection in the OTN layer leads to the most expensive solution, 1.7 to 1.8 times more expensive than restoration (in the case where no MPLS protection mechanisms are used). Path Restoration is in this case the cheapest solution. A second result is that the introduction of MPLS Local Protection has a serious impact on the overall cost. On average, the network cost increases with a factor 1.4 due to its use. In this case the extra cost of 1+1 Protection compared to restoration is even higher: 1.9 to 2 times more expensive. Again, Path Restoration is the cheapest solution. Of course, the network is now also protected against MPLS router (isolating) failures, which was not true in the former case. However, part of the cost increase can be explained by the fact that spare resources are now needed in both layer networks. This results in what is called redundant or double protection: spare resources in the OTN layer also protect spare resources from the MPLS layer, which is superfluous. This can be avoided by supporting the MPLS spare resources as unprotected traffic in the OTN. Even better results can be obtained by adopting a multi-layer survivability strategy based on the common pool concept [31], [32]. The basic idea behind this concept is to support higher layer spare resources as unprotected pre-emptible traffic in the lower layer network.

5. Conclusions

A roadmap has been outlined in this paper, showing how current core networks will evolve from a rather complex IP/ATM/SDH/WDM towards a simplified IP-directly-over-OTN paradigm. In particular the survivability features of such data-centric optical networks have been investigated. Special attention has been paid to the application of MPLS recovery techniques.

Since a data-centric optical network contains at least an IP-MPLS layer and an optical layer, one of the main questions to be answered was: “In which layer to provide survivability features?”. It was shown that each layer has its pros and cons. Therefore, a likely solution seems to be providing survivability at multiple layers, in order to combine the advantages of these layers. However, in order to avoid inefficiencies or conflicts between these layers, the recovery actions of these layers may require coordination. Therefore, in addition to the uncoordinated approach, a sequential (e.g., by means of a hold-off timer or recovery token) and an integrated approach have been proposed.

Finally, some case studies illustrated the relevance of those multi-layer survivability issues. One of the conclusions was that MPLS protection allows fast recovery of traffic at the electrical MPLS level and even more that this is typically cheaper than MPλS protection, but that protection switching at the MPLS level may have a negative impact on TCP goodput during a rather long period (in the order of a (few) second(s)) after the failure and the protection switch. Another case study illustrated that protecting against MPLS router failures, while trying to recover as much traffic as possible in the OTN, without appropriate precautions, may have a significant negative impact on the overall network cost.

Acknowledgements

Part of this work has been supported by the European Commission through the IST-projects LION, TEQUILA and DAVID and by the Flemish Government through the IWT-project ITA/980272/INTEC and an IWT-scholarship. The third and fifth author are a Research Assistant of the Fund for Scientific Research – Flanders (F.W.O.-V., Belgium).

References

[1] NetSizer from Telcordia: http://www.netsizer.com

[2] U.S. Dept. of Commerce, The Emerging Digital Economy, http://www.ecommerce.gov/emerging.htm , (1998).

[3] Evolution of Information and Communication and its impact on research activities, Issue 2, August 30, 1999, EISI-WAY

[4] The European Information Technology Observatory http://www.eito.com

[5] Paul Green, Progress in Optical Networking, IEEE Communications Magazine, Vol. 39, No. 1, (2001), pp. 54-61

[6] K. Struyve, et al., Application, Design and Evolution of WDM in GTS's Pan-European Transport Network, IEEE Communications Magazine, Vol. 38, No. 3, (2000), pp. 114-121.

[7] Nasir Ghani, Lambda-Labeling: A Framework for IP-over-WDM using MPLS, Optical Networks Magazine, Vol. 1, No. 2, (April 2000), pp. 45-58.

[8] Jon Anderson, Protocols and Architectures for IP Optical Networking, Lucent White-paper, May 2000 http://www.lucent.com/minds/techjournal/jan-mar1999/pdf/paper06.pdf

[9] Kohei Shiomoto et. al., Scalable Multi-QoS IP+ATM Switch Router Architecture, IEEE Communications Magazine, Vol. 38, No. 12, (2000), pp. 86-92

[10] G.-S. Kuo, Multi-Protocol Label Switching, special issue of IEEE Communications Magazine, Vol. 37, No. 12, (1999).

[11] R. Callon, et al., A Framework for Multi-Protocol Label Switching, IETF Internet Draft <draft-ietf-mpls-framework-05.txt>, (1997), work in progress.

[12] Grenville Armitage, MPLS: The Magic Behind the Myths, IEEE Communications Magazine, Vol. 38, No. 1, (January 2000), pp. 124-131.

[13] George Swallow, MPLS Advantages for Traffic Engineering, IEEE Communications Magazine, Vol. 37, No. 12, (December 1999), pp. 54-57.

[14] "LDP Specification", L. Andersson, RFC3036, January 2001

[15] RSVP-TE: Extensions to RSVP for LSP Tunnels, Daniel O. Awduche et. al., work in progress, internet-draft February 2001 http://www.ietf.org/internet-drafts/draft-ietf-mpls-rsvp-lsp-tunnel-08.txt

[16] Nico Wauters et. al., Survivability in a New Pan-European Carriers'-Carrier Network Based on WDM and SDH Technology: Current Implementation and Future Requirements, IEEE Communications Magazine, Vol. 37, No. 8, (1999), pp. 63-69

[17] Armand Neukermans and Rajiv Ramaswami, MEMS Technology for Optical Networking Applications, IEEE Communications Magazine, Vol. 39, No. 1, (2001), pp. 62-69

[18] OIF2000.125.3, " User Network Interface (UNI) 1.0 Signaling Specification", Dec.8, 2000

[19] ITU-T G.astn v.0.3, "Architecture for the Automatic Switched Transport Network"

[20] D. Awduche, et al., Multi-Protocol Lambda Switching : Combining MPLS Traffic Engineering Control With Optical Crossconnects, IETF Internet Draft <draft-awduche-mpls-te-optical-01.txt>, (1999), work in progress.

[21] LSP Hierarchy with MPLS TE, Kireeti Kompella et. al., work in progress, internet-draft March 2001 http://search.ietf.org/internet-drafts/draft-ietf-mpls-lsp-hierarchy-02.txt

[22] Didier Colle et. al., Porting MPLS-Recovery Techniques to the MPλS Paradigm, Special Issue on 'Protection and Survivability' of the 'Optical Networks Optical Networks' Magazine, Vol. 2, No. 3, May 2001, to be published

[23] Didier Colle, et. al., MPLS Recovery Mechanisms for IP-over-WDM networks, special issue on 'IP over WDM and Optical Packet Switching' of 'Photonic Network Communications' Magazine, Vol.3, No.1, January, 2001, to be published.

[24] P. Van heuven, et al., Recovery in IP based networks using MPLS, Proc. of IEEE Workshop on IP-oriented Operations & Management (Cracow, Poland, September 2000)

[25] "A Method for Setting an Alternative Label Switched Paths to Handle Fast Reroute", work in progress, internet-draft November 2000: http://infonet.aist-nara.ac.jp/member/nori-d/mlr/id/draft-haskin-mpls-fast-reroute-05.txt

[26] "Framework for MPLS based recovery", Makam et al, work in progress, internet-draft July 2000: http://www.watersprings.org/links/mlr/id/draft-makam-mpls-recovery-frmwrk-01.txt

[27] "Protection/Restoration of MPLS networks", Makam et al, work in progress, internet-draft October 1999: http://www.watersprings.org/links/mlr/id/draft-makam-mpls-protection-00.txt

[28] "A Path Protection/Restoration Mechanism for MPLS Networks", Changcheng Huang et al, work in progress, internet-draft November 2000: http://search.ietf.org/internet-drafts/draft-chang-mpls-path-protection-02.txt

[29] Yinghua Ye, Sudhir Dixit and Mohamed Ali, On Joint Protection/Restoration in IP-centric DWDM-based Optical Transport Networks, IEEE Communications Magazine, Vol. 37, No. 8, (August 1999), pp. 174-183.

[30] RSVP Label Allocation for Backup Tunnels, Robert Goguen, work in progress, internet-draft November 2000 http://www.ietf.org/internet-drafts/draft-swallow-rsvp-bypass-label-01.txt

[31] Gryseels M. et al., Common Pool Survivability for Meshed SDH-Based ATM Networks, Proc. International Symposium on Broadband European Networks SYBEN'98 (Zurich, Switzerland, May 1998, pp. 267-278).

[32] P. Demeester, et al., Resilience in Multi-layer Networks, IEEE Communications Magazine, Vol. 37, No. 8, (August 1999), pp. 70-76.

[33] Giles Heron, Level 3: MPLS over DWDM, IPoWDM conference, Paris, November 2000.

[34] Blake S. et. al., "An Architecture for Differentiated Services", RFC 2475, December 1998

[35] "OSPF version 2", J. Moy, RFC1247, July 1991

[36] Marco Listani and Roberto Sabella, Architectural and Technological Issues for Future Optical Internet Networks, IEEE Communications Magazine, Vol. 38, No. 9, (2000), pp. 82-92

[37] D. Colle et. al., Envisaging Next-Generation Data-Centric Optical Networks, submitted to DRCN2001 conference

[38] Peter Arijs, et. al., Design of Ring and Mesh based WDM Transport Networks, Optical Networks Magazine, Vol. 1, No. 3, (July 2000), pp. 25-40.