CCAMP Working Group
Internet Draft
Expiration Date: May 2003

CCAMP GMPLS P&R Design Team

Dimitri Papadimitriou (Editor)
Eric Mannie (Editor)

Deborah Brungard (AT&T)
Sudheer Dharanikota (Consult)
Jonathan Lang (Calient)
Guangzhi Li (AT&T)
Bala Rajagopalan (Tellium)
Yakov Rekhter (Juniper)

November 2002

Analysis of Generalized MPLS-based Recovery Mechanisms (including Protection and Restoration)

draft-papadimitriou-ccamp-gmpls-recovery-analysis-03.txt

Status of this Memo

1. Abstract

This document provides an analysis grid that can be used to evaluate, compare and contrast the numerous Generalized MPLS (GMPLS)-based recovery mechanisms currently proposed at the CCAMP Working Group. A detailed analysis of each of the recovery phases is provided using the terminology defined in [CCAMP-TERM]. Also, this

document focuses on transport plane survivability and recovery issues and not on control plane resilience and related aspects.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [2].

3. Introduction

This document provides an analysis grid that can be used to evaluate, compare and contrast the numerous Generalized MPLS (GMPLS) based recovery mechanisms currently proposed in the CCAMP Working Group. Here, the focus will only be on transport plane survivability and recovery issues and not on control plane resilience related aspects. Although the recovery mechanisms described in this document impose different requirements on recovery protocols, the protocol(s) specifications will not be covered in this document. Though the concepts discussed here are technology independent, this document will implicitly focus on Sonet/SDH and pre-OTN technologies except when specific details need to be considered (for instance, in the case of failure detection). Details for applicability to other technologies such as Optical Transport Networks (OTN) [ITUT-G709] will be covered in a future release of this document.

In the present release, a detailed analysis is provided for each of the recovery phases as identified in [CCAMP-TERM]. These phases define the sequence of generic operations that need to be performed when a LSP/Span failure (or any other event generating such failures) occurs:

- Phase 1: Failure detection
- Phase 2: Failure localization and isolation
- Phase 3: Failure notification
- Phase 4: Recovery (Protection/Restoration)
- Phase 5: Reversion (normalization)

Failure detection, localization and notification phases together are referred to as fault management. Within a recovery domain, the entities involved during the recovery operations are defined in [CCAMP-TERM]; these entities include ingress, egress and intermediate nodes.

In this document the term ôrecovery mechanismö is used to cover both protection and restoration mechanisms. Specific terms such as protection and restoration are only used when differentiation is

required. Likewise the term ôfailureö is used to represent both signal failure and signal degradation. In addition, a clear distinction is made between partitioning (horizontal hierarchy) and layering (vertical hierarchy) when analyzing hierarchical recovery mechanisms including disjointness related issues. We also introduce the dimensions from which each of the recovery mechanisms described in this document can be further analyzed and provide an analysis grid with respect to these dimensions. Last, we conclude by

detailing the applicability of the current GMPLS protocol building blocks for recovery purposes.

Note: Any other recovery-related terminology used in this document conforms to the one defined in [CCAMP-TERM].

4. Fault Management

4.1 Failure Detection

Transport failure detection is the only phase that can not be achieved by the control plane alone since the latter needs a hook to the transport plane to collect the related information. It has to be emphasized that even if failure events themselves are detected by the transport plane, the latter, upon failure condition, MUST trigger the control plane for subsequent actions through the use of GMPLS signalling capabilities (see [GMPLS-SIG]) or Link Management Protocol (see [LMP], Section 6) capabilities.

Therefore, by definition, transport failure detection is transport technology dependent (and so exceptionally, we keep here the ôtransport planeö terminology). In transport fault management, distinction is made between a defect and a failure. Here, the discussion addresses failure detection (persistent fault cause). In the technology dependent descriptions, a more precise specification will be provided.

As an example, Sonet/SDH (see [G.707], [G.783] and [G.806]) provides supervision capabilities covering:

- Continuity: monitors the integrity of the continuity of a trail (i.e. section or path). This operation is performed by monitoring the presence/absence of the signal. Examples are Loss of Signal (LOS) detection for the physical layer, Unequipped (UNEQ) Signal detection for the path layer, Server Signal Fail Detection (e.g. AIS) at the client layer.

- Connectivity: monitors the integrity of the routing of the signal between end-points. Connectivity monitoring is needed if the layer provides flexible connectivity, either automatically (e.g. cross-connects controlled by the TMN) or manually (e.g. fiber distribution frame). An example is the Trail (i.e. section or path) Trace Identifier used at the different layers and the corresponding Trail Trace Identifier Mismatch detection.

- Alignment: checks that the client and server layer frame start can be correctly recovered from the detection of loss of alignment. The specific processes depend on the signal/frame structure and may include: (multi-)frame alignment, pointer processing and alignment of several independent frames to a common frame start in case of inverse multiplexing. Loss of alignment is a generic term. Examples are loss of frame, loss of multi-frame, or loss of pointer.

- Payload type: checks that compatible adaptation functions are used at the source and the sink. This is normally done by adding a signal type identifier at the source adaptation function and comparing it with the expected identifier at the sink. For instance, the payload signal label and the corresponding payload signal mismatch detection.

- Signal Quality: monitors the performance of a signal. For instance, if the performance falls below a certain threshold a defect û excessive errors (EXC) or degraded signal (DEG) - is detected.

The most important point to keep in mind is that the supervision processes and the corresponding failure detection (used to initiate the recovery phase(s)) result in either:

- Signal Degrade (SD): A signal indicating that the associated data has degraded in the sense that a degraded defect condition is active (for instance, a dDEG declared when the Bit Error Rate exceeds a preset threshold).

- Signal Fail (SF): A signal indicating that the associated data has failed in the sense that a signal interrupting near-end defect condition is active (as opposed to the degraded defect).

In Optical Transport Networks (OTN) equivalent supervision capabilities are provided at the optical/digital section layers (OTS, OMS and OTUk) and at optical/digital path layers (OCh and ODUk). Interested readers are referred to the ITU-T Recommendations

[G.798] and [G.709] for more details.

The above are examples where the failure detection, reporting and recovery responsible entities are co-located.

On the other hand, in pre-OTN networks, a failure may be masked by intermediate O/E/O based Optical Line System (OLS), preventing a Photonic Cross-Connect (PXC) from detecting upstream failures. In such cases, failure detection may be assisted by an out-of-band communication channel and failure condition reported to the PXC control plane. This can be provided by using [LMP-WDM] extensions that delivers IP message-based communication between the PXC and the OLS control plane. Also, since PXCs are framing format independent, failure conditions can only be triggered either by detecting the absence of the optical signal or by measuring its optical quality, mechanisms which are less reliable than electrical (digital) mechanisms. Both types of detection mechanisms are out of the scope of this document. If the intermediate OLS supports electrical (digital) mechanisms, using the LMP communication channel, these failure conditions are reported to the PXC and subsequent recovery actions performed as described in Section 5. As such from the control plane viewpoint, this mechanism makes the OLS-PXC composed system appearing as a single logical entity allowing considering for

such entity the same failure management mechanisms as for any other O/E/O capable device.

This example is to illustrate the scenario where the failure detection and reporting (recovery responsible) entities are not co-located.

More generally, the following are typical failure conditions in Sonet/SDH and pre-OTN networks:

- Loss of Light (LOL)/Loss of Signal (LOS): Signal Failure (SF) condition where the optical signal is not detected anymore on a given interfaceÆs receiver.
- Signal Degrade (SD): detection of the signal degradation over a specific period of time.
- For Sonet/SDH payloads, all of the above-mentioned supervision capabilities can be used, resulting in SD or SF condition.

In summary, the following cases are considered to illustrate the communication between the detecting and reporting (also recovery responsible) entities:

- Co-located detecting and reporting entities: both the detecting and reporting entities are on the same node (e.g., Sonet/SDH

equipment, Opaque cross-connects, and, with some limitations, for Transparent cross-connects, etc.).

- Non co-located detecting and reporting entities:
  - with In-band communication between entities:
    Entities are separated but transport plane (in-band) communication is provided between them (e.g., Server Signal Failures (AIS), etc.)
  - with Out-of-band communication between entities:
    Entities are separated but out-of-band communication is provided between them (e.g., using [LMP]).

4.2 Failure Localization and Isolation

Failure localization provides the required information in order to perform the subsequent recovery action(s) at the LSP/span end-points.

In some cases, accurate failure localization may be less urgent; the need is to identify the failure as occurring within the recovery domain. This is particularly the case when edge-to-edge LSP recovery (edge referring to a sub-network end-node for instance) is performed based on a simple failure notification (including the identification of the failed working LSPs) so that a more accurate localization can be performed after LSP recovery.

Failure localization should be triggered immediately after the fault detection phase. This operation can be performed at the transport management plane and/or, if unavailable (via the transport plane),

the control plane level where dedicated signaling messages can be used.

When performed at the control plane level, a protocol such as LMP (see [LMP], Section 6) can be used for failure localization and isolation purposes.

4.3 Failure Notification

Failure notification is used 1) to inform intermediate nodes that a LSP/span failure has occurred and has been detected 2) to inform the recovery deciding entities (which can correspond to any intermediate or end-point of the failed LSP/span) that the corresponding service is not available. In general, these deciding entities will be the ones taking the appropriate recovery decision. When co-located with the recovering entity, these entities will also perform the

corresponding recovery action(s).

Failure notification can be either provided by the transport or by the control plane. As an example, let us first briefly describe the failure notification mechanism defined at the Sonet/SDH transport plane level (also referred to as maintenance signal supervision):

- AIS (Alarm Indication Signal) occurs as a result of a failure condition such as Loss of Signal and is used to notify downstream nodes (of the appropriate layer processing) that a failure has occurred. AIS performs two functions 1) inform the intermediate nodes (with the appropriate layer monitoring capability) that a failure has been detected 2) notify the connection end-point that the service is no longer available.

For a distributed control plane supporting one (or more) failure notification mechanism(s), regardless of the mechanismÆs actual implementation, the same capabilities are needed with more (or less) information provided about the LSPs/Spans under failure condition, their detailed status, etc.

The most important difference between these mechanisms is related to the fact that transport plane notifications (as defined today) would initiate a protection scheme directly (such as those defined in [CCAMP-TERM]) or a restoration scheme via the management plane. On the other hand, using a failure notification mechanism through the control plane would provide the possibility to trigger either a protection or a restoration action via the control plane. This has the advantage that a control plane recovery responsible entity does not necessarily have to be co-located with a transport maintenance/recovery domain. A control plane recovery domain can be defined at entities not supporting a transport plane recovery.

Moreover, as specified in [GMPLS-SIG], notification message exchanges through a GMPLS control plane may not follow the same path as the LSP/spans for which these messages carry the status. In turn, this ensures a fast, reliable (through the use of either a dedicated

control plane network or disjoint control channels) and efficient (through the aggregation of several LSP/span status within the same message) failure notification mechanism.

The other important properties to be met by the failure notification mechanism are mainly the following:

- Notification messages must provide enough information such that the most efficient subsequent recovery action will be taken (in

most of the recovery schemes this action is even deterministic) at the recovering entities. Remember here that these entities can be either intermediate or end-points through which normal traffic flows. Based on local policy, intermediate nodes may not use this information for subsequent recovery actions (see for instance the APS protocol phases as described in [CCAMP-TERM]). In addition, since fast notification is a mechanism running in collaboration with the existing signalling (see for instance, [GMPLS-RSVP-TE]) allowing intermediate nodes to stay informed about the status of the working LSP/spans under failure condition.

The trade-off here is to define what information the LSP/span end-points (more precisely, the deciding entity) needs in order for the recovering entity to take the best recovery action: if not enough information is provided, the decision can not be optimal (note that in this eventuality, the important issue is to quantify the level of sub-optimality), if too much information is provided the control plane may be overloaded with unnecessary information and the aggregation/correlation of this notification information will be more complex and time consuming to achieve. Notice that more detailed quantification of the amount of information to be exchanged and processed is strongly dependent on the failure notification protocol specification.

- If the failure localization and isolation is not performed by one of the LSP/Span end-points or some intermediate points, they should receive enough information from the notification message in order to locate the failure otherwise they would need to (re-) initiate a failure localization and isolation action.

- Avoiding so-called notification storms implies that failure detection output is correlated (i.e. alarm correlation) and aggregated at the node detecting the failure(s), failure notifications are directed to a restricted set of destinations (in general the end-points) and notification suppression (i.e. alarm suppression) is provided in order to limit flooding in case of multiple and/or correlated failures appearing at several locations in the network

- Alarm correlation and aggregation (at the failure detecting node) implies a consistent decision based on the conditions for which a trade-off between fast convergence (at detecting node) and fast notification (implying that correlation and aggregation occurs at receiving end-points) can be found.

4.5 Correlating Failure Conditions

A single failure event (such as a span failure) can result into reporting multiple failures (such as individual LSP failures) conditions. These can be grouped (i.e. correlated) to reduce the number of failure conditions communicated on the reporting channel, for both in-band and out-of-band failure reporting.

In such a scenario, it can be important to wait for a certain period of time, typically called failure correlation time, and gather all the failures to report them as a group of failures (or simply group failure). For instance, this approach can be provided using LMP-WDM for pre-OTN networks (see [LMP-WDM]) or when using Signal Failure/ Degrade Group in the Sonet/SDH context.

Note that a default average time interval during which failure correlation operation can be performed is difficult to provide since it is strongly dependent on the underlying network topology. Therefore, it can be advisable to provide a per node configurable failure correlation time. The detailed selection criteria for this time interval are outside of the scope of this document.

When failure correlation is not provided, multiple failure notification messages may be sent out in response to a single failure (for instance, a fiber cut), each one containing a set of information on the failed working resources (for instance, the individual lambda LSP flowing through this fiber). This allows for a more prompt response but can potentially overload the control plane due to a large amount of failure notifications.

5. Recovery Mechanisms and Schemes

5.1 Transport vs. Control Plane Responsibilities

For both protection and restoration, and when applicable, recovery resources are provisioned using GMPLS signalling capabilities. Thus, these are control plane-driven actions (topological and resource-constrained) that are always performed in this context.

The following table gives an overview of the responsibilities taken by the control plane in case of LSP/Span recovery:

1. LSP/span Protection Schemes

- Phase 1: Failure detection — Transport plane
- Phase 2: Failure isolation/localization — Transport/Control plane
- Phase 3: Failure notification — Transport/Control plane
- Phase 4: Protection switching — Transport/Control plane
- Phase 5: Reversion (normalization) — Transport/Control plane

Note: in the LSP/span protection context control plane actions can be performed either for operational purposes and/or synchronization

purposes (vertical synchronization between transport and control plane) and/or notification purposes (horizontal synchronization between nodes at control plane level).

2. LSP/span Restoration Schemes

- Phase 1: Failure detection Transport plane
- Phase 2: Failure isolation/localization Transport/Control plane
- Phase 3: Failure notification Control plane
- Phase 4: Recovery switching Control plane
- Phase 5: Reversion (normalization) Control plane

Therefore, this document is primarily focused on provisioning of recovery resources, failure notification, LSP/span recovery and reversion operations. Moreover some additional considerations can be dedicated to the mechanisms associated to the failure localization/isolation phase.

5.2 Technology in/dependent mechanisms

The present recovery mechanisms analysis applies in fact to any circuit oriented data plane technology with discrete bandwidth increments (like Sonet/SDH, G.709 OTN, etc.) being controlled by an IP-centric distributed control plane.

The following sub-sections are not intended to favor one technology versus another. They just lists pro and cons for each of them in order to determine the mechanisms that GMPLS-based recovery must deliver to overcome their cons and take benefits of their pros in their respective applicability context.

5.2.1 OTN Recovery

OTN recovery specifics are left for further considerations.

5.2.2 Pre-OTN Recovery

Pre-OTN Recovery specifics (also referred to as ôlambda switchingö) presents mainly the following advantages:

- benefits from a simpler architecture making it more suitable for meshed-based recovery schemes (on a per channel basis).

- when providing suppression of intermediate node transponders (vs. use of non-standard masking of upstream failures) e.g. use of squelching, implies that failures (such as LoL) will propagate to

edge nodes giving the possibility to initiate upper layer driven recovery actions.

The main disadvantage comes from the lack of interworking due to the large amount of failure management (in particular failure notification protocols) and recovery mechanisms currently available.

Note also, that for all-optical networks, combination of recovery with optical physical impairments is left for a future release of this document since corresponding detection technologies are under specification.

5.2.3 Sonet/SDH Recovery

Some of the advantages of Sonet/SDH and more generically any TDM transport plane are:

- Protection schemes are standardized (see [G.841]) and can operate across protected domains and interwork (see [G.842]).

- Provides failure detection, notification and path/section Automatic Protection Switching (APS) mechanisms.

- Provides greater control over the granularity of the TDM LSPs/Links that can be recovered with respect to coarser optical channel (or whole fiber content) recovery switching

Some of the current limitations of the Sonet/SDH layer recovery are:

- Limited topological scope: Inherently the use of ring topologies (Dedicated SNCP or Shared Protection Rings) has a reduced flexibility with respect to the somewhat more complex but potentially more resource efficient mesh-based recovery schemes.

- Inefficient use of spare capacity: Sonet/SDH protection is largely applied for ring topologies, where spare capacity often remains idle, making the efficiency of bandwidth usage an issue.

- Support of meshed recovery requires intensive network management development, and the functionality is limited by both the network elements and the element management systems capabilities.

5.3 Specific Aspects of Control Plane-based Recovery Mechanisms

5.3.1 In-band vs Out-of-band Signalling

The nodes communicate through the use of (IP terminating) control channels defining the control plane (transport) topology. In this context, two classes of transport mechanisms can be considered here i.e. in-fiber or out-of-fiber (through a dedicated physically diverse control network referred to as the Data Communication Network or DCN). The potential impact of the usage of an in-fiber (signalling) transport mechanism is briefly considered here.

In-fiber transport mechanism can be further subdivided into in-band and out-of-band. As such, the distinction between in-fiber in-band and in-fiber out-of-band signalling reduces to the consideration of a logically versus physically embedded control plane topology with respect to the transport plane topology. In the scope of this document, since we assume that (IP terminating) channels between

nodes must be continuously available in order to enable the exchange of recovery-related information and messages, one considers that in either case (i.e. in-band or out-of-band) at least one logical channel or one physical channel between nodes is available.

Therefore, the key issue when using in-fiber signalling is whether we can assume independence between the fault-tolerance capabilities of control plane and the failures affecting the transport plane (including the nodes). Note also that existing specifications like the OTN provide a limited form of independence for in-fiber signaling by dedicating a separate optical supervisory channel (OSC, see [ITU-T G.709] and [ITU-T G.874]) to transport the overhead and other control traffic. For OTNs, failure of the OSC does not result in failing the optical channels. Similarly, loss of the control channel must not result in failing the data (transport plane).

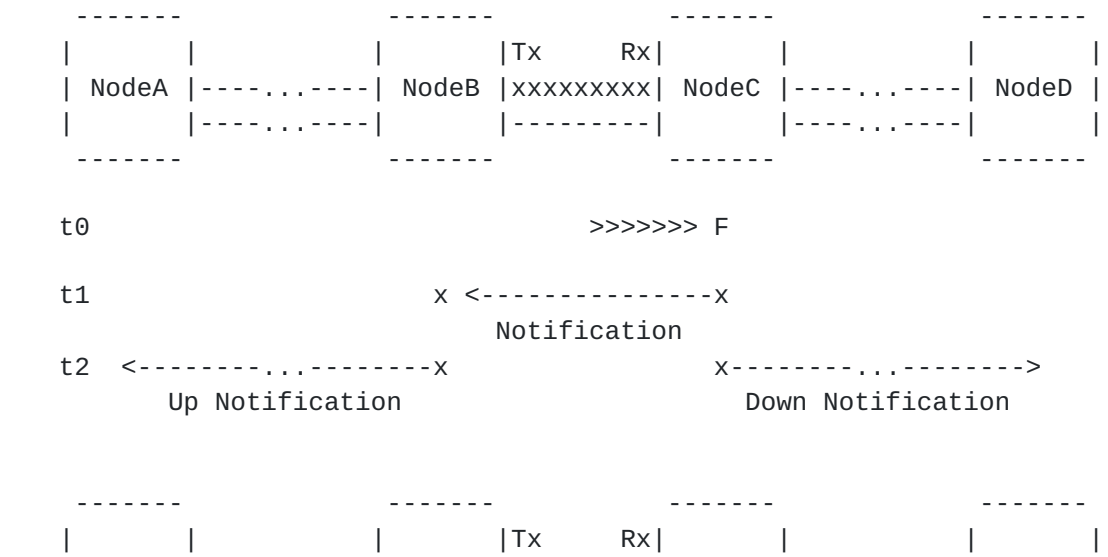
5.3.2 Uni- versus Bi-directional Failures

The failure detection, correlation and notification mechanisms (described in Section 4) can be triggered when either a unidirectional or a bi-directional LSP/Span failure occurs (or a combination of both). As illustrated in Figure 1 and 2, two alternatives can be considered here:

1. Uni-directional failure detection: the failure is detected on the receiver side i.e. it is only is detected by the downstream node to the failure (or by the upstream node depending on the failure propagation direction, respectively)

2. Bi-directional failure detection: the failure is detected on the receiver side of both downstream node AND upstream node to the failure.

Notice that after the failure detection time, if only control plane based failure management is provided, the peering node is unaware of the failure detection status of its neighbor.



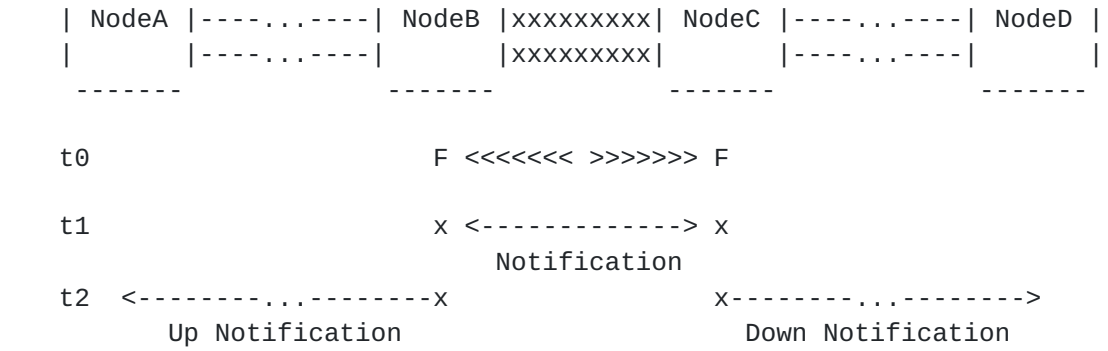


Fig. 1 & 2. Uni- and Bi-directional Failure Detection/Notification

After failure detection, the following failure management operations can be subsequently considered:

- Each detecting entity sends a notification message to the corresponding transmitting entity. For instance, in Fig. 1 (Fig. 2), node C sends a notification message to node B (while node B sends a notification message to node A). To ensure reliable failure notification, a dedicated acknowledgment message can be returned back to the sender node.

- Next, within a certain (and pre-determined) time window, nodes impacted by the failure occurrences perform their correlation. In case of unidirectional failure, node B only receives the notification message from node C and thus the time for this operation is negligible. However, in case of bi-directional failure, node B (and node C) must correlate the received notification message from node C (node B, respectively) with the corresponding locally detected information.

- After some (pre-determined) period of time, referred to as the hold-off time, after which local recovery actions were not successful, the following occurs. In case of unidirectional failure and depending on the directionality of the connection, node B should send an upstream notification message to the ingress node A or node C should send a downstream notification to the egress node D. However, in such a case only node A (node D, respectively) referred to as the master and node D, to as the slave per [CCAMP-TERM], would initiate a edge to edge recovery action. Note that the connection terminating node (i.e. node D or node A) may be optionally notified.

  In case of bi-directional failure, node B may send an upstream notification message to the ingress node A or node C a downstream notification to the egress node D. However, due to the dependence on the connection directionality, only ingress node A or egress node D would initiate an edge to edge recovery action. Note that the connection terminating node (i.e. node D or node A) should be also notified of this event using upstream and downstream fast

  notification (see [GMPLS-SIG]). For instance, if a connection directed from D to A is under failure condition, only the notification sent by from node C to D would initiate a recovery action. Here as well, per [CCAMP-TERM], the deciding (and recovering) node D is referred to as the "master" while the node A is referred to as the "slave" (i.e. recovering only entity).

  Note: The determination of the master and the slave may be based either on configured information or dedicated protocol capability.

In the above scenarios, the path followed by the notification messages does not have to be the same as the one followed by the failed LSP (see [GMPLS-SIG], for more details on the notification message exchange). The important point, concerning this mechanism, is that either the detecting/reporting entity (i.e. the nodes B and C) are also the deciding/recovery entity or the detecting/reporting entities are simply intermediate nodes in the subsequent recovery

process. One refers to local recovery in the former case and to edge-to-edge recovery in the latter one.

5.3.3 Partial versus Full Span Recovery

When given span carries more than one LSPs or LSP segments, an additional aspect must be considered during span failure carrying several LSPs. These LSPs can be either individually recovered or recovered as a group (aka bulk LSP recovery) or independent sub-groups. The selection of this mechanism would be triggered independently of the failure notification granularity when correlation time windows are used and simultaneous recovery of several LSPs can be performed using single request. Moreover, criteria by which such sub-groups can be formed are outside of the scope of this document.

An additional complexity arises in case of (sub-)group LSP recovery. Between a given node pair, the LSPs a given (sub-)group contains may have been created from different source (i.e. initiator) nodes toward different destinations nodes. Consequently the failure notification messages sub-sequent to a bi-directional span failure affecting several LSPs (or the whole group of LSPs it carries) are not necessarily directed toward the same initiator nodes. In particular these messages may be directed to both the upstream and downstream nodes to the failure. Therefore, such span failure may trigger recovery actions to be performed from both sides (i.e. both from the upstream and the downstream node to the failure). In order to facilitate the definition of the corresponding recovery mechanisms (and their sequence), one assumes here as well, that per [CCAMP-TERM] the deciding (and recovering) entity, referred to as the "master" is the only initiator of the recovery of the whole LSP (sub-)group.

5.3.4 Difference between LSP, LSP Segment and Span Recovery

The recovery definitions given in [CCAMP-TERM] are quite generic and apply for link (or local span) and LSP recovery. The major difference between LSP, LSP Segment and span recovery is related to the number of intermediate nodes that the signalling messages have to travel. Since nodes are not necessarily adjacent in case of LSP (or LSP Segment) recovery, signalling message exchanges from the reporting to the deciding/recovery entity will have to cross several intermediate nodes. In particular, this applies for the notification

messages due to the number of hops separating the failure occurrence location from their destination. This results in an additional propagation and forwarding delay. Note that the former delay may in certain circumstances be non-negligible e.g. in case of copper out-of-band network one has to consider approximately 1 ms per 200km.

Moreover, the recovery mechanisms applicable to end-to-end LSP and to the segments (i.e. edge-to-edge) that may compose an end-to-end LSP can be exactly the same. However, one expects in the latter case, that the destination of the failure notification message will be the ingress of each of these segments. Therefore, taking into account the mechanism described in Section 5.3.2, failure notification can be first exchanged between the LSP segments terminating points and after expiration of the hold-off time directed toward end-to-end LSP terminating points.

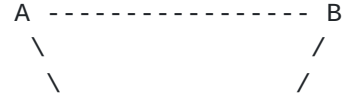
5.4 Difference between Recovery Type and Scheme

Section 4.6 of [CCAMP-TERM] defines the basic recovery types. The purpose of this section is to describe the schemes that can be built using these recovery types. In brief, a recovery scheme is defined as the combination between different ingress-egress node pairs of a set of identical recovery types. Several examples are provided in order to illustrate the difference between a recovery type such as 1:1 and a recovery scheme such as (1:1)^n.

1. (1:1)^n with recovery resource sharing

The exponent, n, indicates the number of times a 1:1 recovery type is applied between at most n different ingress-egress node pairs. Here, at most n pairs of disjoint working and recovery LSPs/spans share at most n times a common resource. Since the working LSPs/spans are mutually disjoint, simultaneous requests for use of the shared (common) resource will only occur in case of simultaneous failures, which is less likely to happen.

For instance, in the (1:1)^2 common case if the 2 recovery LSPs in the group overlap the same common resource, then it can handle only single failures; any multiple working LSP failures will cause at least one working LSP to be denied automatic recovery. Consider for instance, the following example, with working LSPs A-B and E-F and recovery LSPs A-C-D-B and E-C-D-F sharing a common C-D resource.

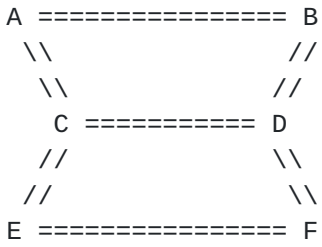


2. (M:N)^n with recovery resource sharing

The exponent, n, indicates the number of times a M:N recovery type is applied between at most n different ingress-egress node pairs. So the interpretation follows from the previous case, expect that here disjointness applies to the N working LSPs/spans and to the M recovery LSPs/spans while sharing at most n times M common resources.

In both schemes, one may see the following at the LSP level: we have a ôgroupö of sum{n=1}^N N{n} working LSPs and a pool of shared backup resources, not all of which are available to any given working path. In such conditions, defining a metric that describes the amount of overlap among the recovery LSPs would give some indication of the groupÆs ability to handle multiple simultaneous failures.

For instance, in the simple (1:1)^n case situation if n recovery LSPs in a (1:1)^n group overlap, then it can handle only single failures; any multiple working LSP failures will cause at least one working LSP to be denied automatic recovery. But if one consider for instance, a (2:2)^2 group in which there are two pairs of overlapping recovery LSPs, then two LSPs (belonging to the same pair) can be simultaneously recovered. The latter case can be illustrated as follows: 2 working LSPs A-B and E-F and 2 recovery LSPs A-C-D-B and E-C-D-F sharing the two common C-D resources.



Moreover, in all these schemes, (working) path disjointness can be reinforced by exchanging working LSP related information during the recovery LSP signalling.

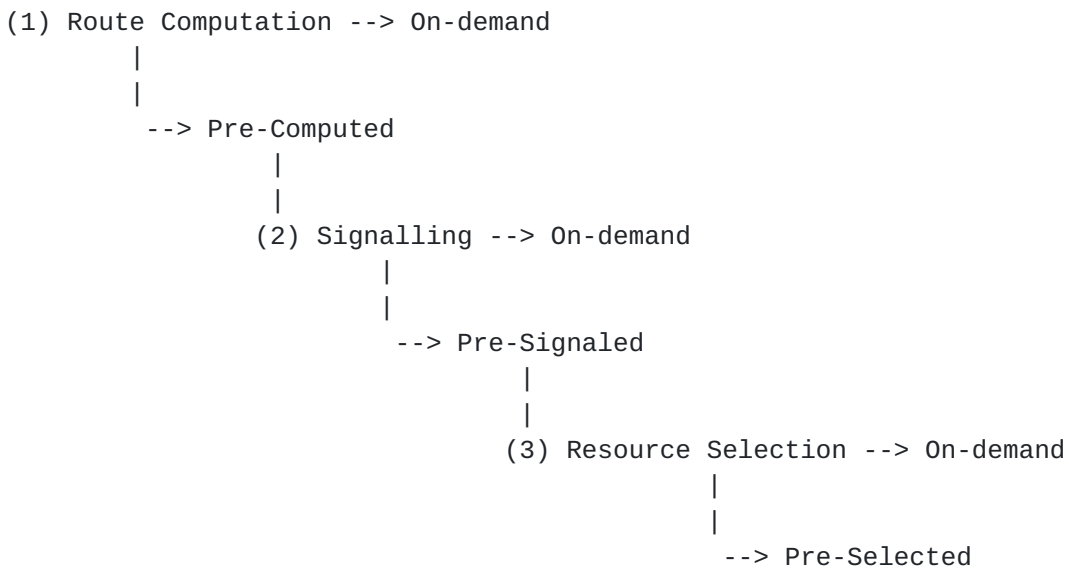
Specific issues related to the combination of shared (discrete) bandwidth and disjointness for recovery schemes are described in Section 8.4.2.

5.5 LSP Restoration Schemes

5.5.1 Classification

LSPs/spans recovery time and ratio depend on the proper recovery LSP (soft) provisioning and the level of recovery resources overbooking (i.e. over-provisioning). A proper balance of these two mechanisms will result in a desired LSP/span recovery time and ratio when single or multiple failure(s) occur(s).

Recovery LSP Provisioning phases:



Overbooking Levels:

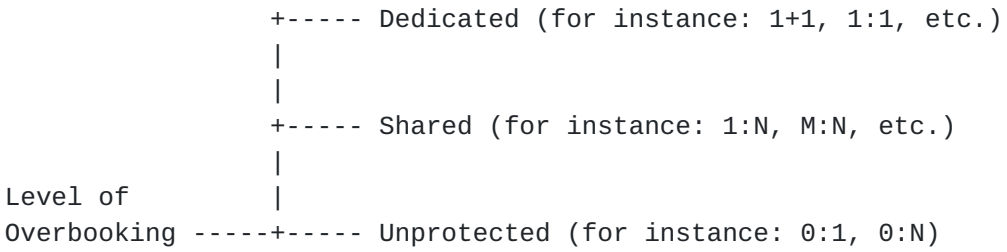


Fig 3. LSP Provisioning and Overbooking Classification

In this figure, we present a classification of different options under LSP provisioning and overbooking. Although we acknowledge these operations are run mostly during planning (using network planning) and provisioning time (using signaling and routing) activities, we keep them in analyzing the recovery schemes.

Proper LSP/span provisioning will help in alleviating many of the failures. As an example, one may compute primary and secondary paths, either end-to-end or segment-per-segment, to recover an LSP from multiple failure events affecting link(s), node(s), SRLG(s)

and/or SRG(s). Such primary and secondary LSP/span provisioning can be categorized, as shown in the above figure, based on:

(1) the recovery path (i.e. route) can be either pre-computed or computed on demand.
(2) when the recovery path is pre-computed: pre-signaled (implying recovery resource reservation) or signaled on demand.
(3) and when the recovery resources are reserved, they can be either pre-selected or selection on-demand.

Note that these different options give rise to different LSP/span recovery times. The following subsections will consider all these cases in analyzing the schemes.

There are many mechanisms available allowing the overbooking of the recovery resources. This overbooking can be done per LSP (such as the example mentioned above), per link (such as span protection) or per domain (such as ring topologies). In all these cases the level of overbooking, as shown in the above figure, can be classified as dedicated (such as 1+1 and 1:1), shared (such as 1:N and M:N) or unprotected (i.e. restorable if enough recovery resources are available).

Under a shared restoration scheme one may support preemptable (preempt low priority connections in case of resource contention) extra-traffic. In this document we keep in mind all the above-mentioned overbooking mechanisms in analyzing the recovery schemes.

5.5.2 Dynamic LSP Restoration

We first define the following times in order to provide a quantitative estimation about the time performance of the different dynamic and pre-signaled LSP restoration:

- Path Computation Time - Tpc
- Path Selection Time - Tps
- End-to-end LSP Resource Reservation û Trr (a delta for resource selection is also considered, the corresponding total time is then referred to as Trrs)
- End-to-end LSP Resource Activation Time û Tra (a delta for resource selection is also considered, the corresponding total time is then referred to as Tras)

Path Selection Time (Tps) is considered when a pool of recovery LSPÆs paths between a given source/destination is pre-computed and after failure occurrence one of these paths is selected for the recovery of the LSP under failure condition.

Note: failure management operations such as failure detection, correlation and notification are considered as equivalently time consuming for all the mechanisms described here below:

1. With Route Pre-computation (or LSP re-provisioning)

An end-to-end restoration LSP is established after the failure(s) occur(s) based on a pre-computed path (i.e. route). As such, one can define this as an ôLSP re-provisioningö mechanism. Here, one or more (disjoint) routes for the restoration path are computed (and optionally pre-selected) before a failure occurs.

No reservation or selection of resources is performed along the restoration path before failure. As a result, there is no guarantee that a restoration connection is available when a failure occurs.

The expected total restoration time T is thus equal to Tps + Trrs or when a dedicated computation is performed for each working LSP to Trrs.

2. Without Route Pre-computation (or LSP re-routing)

An end-to-end restoration LSP is established after the failure(s) occur(s). Here, one or more (disjoint) explicit routes for the restoration path are dynamically computed and one is selected after failure. As such, one can define this as an ôLSP re-routingö mechanism.

No reservation or selection of resources is performed along the restoration path before failure. As a result, there is no guarantee that a restoration connection is available when a failure occurs.

The expected total restoration time T is thus equal to Tpc (+ Tps) + Trrs. Therefore, time performance between these two approaches differs by the time required for route computation Tpc (and its potential selection time, Tps).

5.5.3 Pre-signaled Restoration LSP

1. With resource reservation without pre-selection

An end-to-end restoration path is pre-selected from a set of one or more pre-computed (disjoint) explicit route before failure. The restoration LSP is signaled along this pre-selected path to reserve resources (i.e. signaled) at each node but resources are not selected.

In this case, the resources reserved for each restoration LSP may be dedicated or shared between different working LSP that are not expected to fail simultaneously. Local node policies can be applied to define the degree to which these resources are shared across independent failures.

Upon failure detection, signaling is initiated along the restoration path to select the resources, and to perform the appropriate operation at each node entity involved in the restoration connection (e.g. cross-connections).

The expected total restoration time T is thus equal to Tras (post-failure activation) while operations performed before failure occurrence takes Tpc + Tps + Trr.

2. With resource reservation and pre-selection

An end-to-end restoration path is pre-selected from a set of one or more pre-computed (disjoint) explicit route before failure. The restoration LSP is signaled along this pre-selected path to reserve AND select resources at each node but not cross-connected. Such that the selection of the recovery resources is fixed at the control plane level. However, no cross-connections are performed along the restoration path.

In this case, the resources reserved for each restoration LSP may only be shared between different working LSPs that are not expected to fail simultaneously. Since one considers restoration schemes here, the sharing degree should not be limited to working (and recovery) LSPs starting and ending at the same ingress and egress nodes. Therefore, one expects to receive some feedback information on the recovery resource sharing degree at each node participating to the recovery scheme.

Upon failure detection, signaling is initiated along the restoration path to activate the reserved and selected resources and to perform the appropriate operation at each node involved in the restoration connection (e.g. cross-connections).

The expected total restoration time T is thus equal to Tra (post-failure activation) while operations performed before failure occurrence takes Tpc + Tps + Trrs. Therefore, time performance between these two approaches differs only by the time required for resource selection during the activation of the recovery LSP (i.e. Tras û Tra).

5.5.4 LSP Segment Restoration

The above approaches can be applied on a sub-network basis rather than end-to-end basis (in order to reduce the global recovery time).

It should be also noted that using the horizontal hierarchical approach described in Section 7.1, that a given end-to-end LSP can be recovered by multiple recovery mechanisms (e.g. 1:1 protection in a metro edge network but M:N protection in the core). These mechanisms are ideally independent and may even use different failure localization and notification mechanisms.

6. Normalization

Normalization is defined as the mechanism allowing switching normal traffic from the recovery LSP/span to the working LSP/span previously under failure condition.

Use of normalization is under the discretion of the recovery domain policy. Normalization (reversion) may impact the normal traffic (a second hit) depending on the normalization mechanism used.

If normalization is supported 1) the LSP/span must be returned to the working LSP/span when the failure condition clears 2) capability to de-activate (turn-off) the use of reversion should be provided. De-activation of reversion should not impact the normal traffic (regardless if currently using the working or recovery LSP/span).

Note: during the failure, the reuse of any non-failed resources (e.g. LSP spans) belonging to the working LSP/span is under the discretion of recovery domain policy.

6.1 Wait-To-Restore

A specific mechanism (Wait-To-Restore) is used to prevent frequent recovery switching operation due to an intermittent defect (e.g. BER fluctuating around the SD threshold).

First, an LSP/span under failure condition must become fault-free, e.g. a BER less than a certain recovery threshold. After the recovered LSP/span (i.e. the previously working LSP/span) meets this criterion, a fixed period of time shall elapse before normal traffic uses the corresponding resources again. This duration called Wait-To-Restore (WTR) period or timer is generally of the order of a few minutes (for instance, 5 minutes) and should be capable of being set. The WTR timer may be either a fixed period, or provide for incremental longer periods before retrying. An SF or SD condition on the previously working LSP/span will override the WTR timer value (i.e. the WTR cancels and the WTR timer will restart).

6.2 Revertive Mode Operation

In revertive mode of operation, when the recovery LSP/span is no longer required, i.e. the failed working LSP/span is no longer in SD or SF condition, a local Wait-to-Restore (WTR) state will be activated before switching the normal traffic back to the recovered working LSP/span.

During the reversion operation, since this state becomes the highest in priority, signalling must maintain the normal traffic on the recovery LSP/span from the previously failed working LSP/span. Moreover, during this WTR state, any null traffic or extra traffic (if applicable) request is rejected.

However, deactivation (cancellation) of the wait-to-restore timer may occur in case of higher priority request attempts. That is the recovery LSP/span usage by the normal traffic may be preempted if a higher priority request for this recovery LSP/span is attempted.

6.3 Orphans

When a reversion operation is requested normal traffic must be switched from the recovery to the recovered working LSP/span. A particular situation occurs when the previously working LSP/span can not be recovered such that normal traffic can not be switched back. In such a case, the LSP/span under failure condition (also referred to as ôorphanö) must be cleared i.e. removed from the pool of resources allocated for normal traffic. Otherwise, potential de-synchronization between the control and transport plane resource usage can appear. Depending on the signalling protocol capabilities and behavior different mechanisms are to be expected here.

Therefore any reserved or allocated resources for the LSP/span under failure condition must be unreserved/de-allocated. Several ways can be used for that purpose: wait for the elapsing of the clear-out time interval, or initiate a deletion from the ingress or the egress node, or trigger the initiation of deletion from an entity (such as an EMS or NMS) capable to react on the reception of an appropriate notification message.

7. Hierarchies

Recovery mechanisms are being made available at multiple (if not each) transport layers within so-called ôIP-over-opticalö networks. However, each layer has certain recovery features and one needs to determine the exact impact of the interaction between the recovery mechanisms provided by these layers.

Hierarchies are used to build scalable complex systems. Abstraction is used as a mechanism to build large networks or as a technique for enforcing technology, topological or administrative boundaries. The same hierarchical concept can be applied to control the network survivability. In general, it is expected that the recovery action is taken by the recoverable LSP/span closest to the failure in order to avoid the multiplication of recovery actions. Moreover, recovery hierarchies can be also bound to control plane logical partitions (e.g. administrative or topological boundaries). Each of them may apply different recovery mechanisms.

In brief, commonly accepted ideas are generally that the lower layers can provide coarse but faster recovery while the higher layers can provide finer but slower recovery. Moreover, it is also more than desirable to avoid too many layers with functional overlaps. In this context, this section intends to analyze these hierarchical aspects including the physical (passive) layer(s).

7.1 Horizontal Hierarchy (Partitioning)

A horizontal hierarchy is defined when partitioning a single layer network (and its control plane) into several recovery domains. Within a domain, the recovery scope may extend over a link (or span), LSP segment or even an end-to-end LSP. Moreover, an

administrative domain may consist of a single recovery domain or can be partitioned into several smaller recovery domains. The operator can partition the network into recovery domains based on physical network topology, control plane capabilities or various traffic engineering constraints.

An example often addressed in the literature is the metro-core-metro application (sometimes extended to a metro-metro/core-core) within a single transport layer (see Section 7.2). For such a case, an end-to-end LSP is defined between the ingress and egress metro nodes, while LSP segments may be defined within the metro or core sub-networks. Each of these topological structures determines a so-called ôrecovery domainö since each of the LSPs they carry can have its own recovery type (or even scheme). The support of multiple recovery schemes within a sub-network is referred to as a multi-recovery capable domain or simply multi-recovery domain.

7.2 Vertical Hierarchy (Layers)

It is a very challenging task to combine in a coordinated manner the different recovery capabilities available across the path (i.e. switching capable) and section layers to ensure that certain network survivability objectives are met for the different services supported by the network.

As a first analysis step, one can draw the following guidelines for a vertical coordination of the recovery mechanisms:
- The lower the layer the faster the notification and switching
- The higher the layer the finer the granularity of the recoverable entity and therefore the granularity of the recovery resource (and subsequently its sharing ratio)

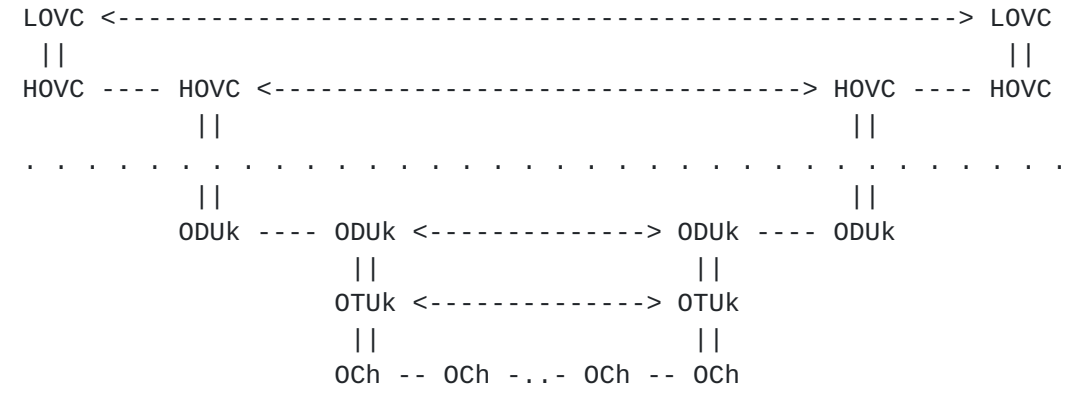
Therefore, in the scope of this analysis, a vertical hierarchy consists of multiple layered transport planes providing different:
- Discrete bandwidth granularities for non-packet LSPs such as OCh, ODUk, STS SPE/HOVC and VT SPE/LOVC LSPs and continuous bandwidth granularities for packet LSPs
- Potentially, recovery capabilities with different temporal granularities: ranging from milliseconds to tens of seconds

Note: based on the bandwidth granularity we can determine four classes of vertical hierarchiesÆ (1) packet over packet (2) packet over circuit (3) circuit over packet and (4) circuit over circuit. Here below we extend a little bit more on (4), (2) being covered in [TE-RH] on the other hand (1) is extensively covered at the MPLS Working Group, and (3) at the PWE3 Working Group.

In Sonet/SDH environments, one typically considers the VT/LOVC and STS SPE/HOVC as independent layers, VT/LOVC LSP using the underlying STS SPE/HOVC LSPs as links, for instance. In OTN, the ODUk path layers will lie on the OCh path layer i.e. the ODUk LSPs using the underlying OCh LSPs as links. Notice here that server layer LSPs may

simply be provisioned and not dynamically triggered or established (control driven approach).

The following figure (including only the path layers) illustrates the hierarchical layers that can be covered by the recovery architecture of a transmission network comprising a SDH/Sonet and an OTN part:



In this context, the important points are the following:
- these layers are path layers; i.e. the ones controlled by the GMPLS (in particular, signalling) protocol suite.
- an LSP at the lower layer for instance an optical channel (= network connection) appears as a section (= link) for the OTUk layer i.e. the links that are typically controlled by link management protocols such as LMP.

The first key issue with multi-layer recovery is that achieving control plane individual or bulk LSP recovery will be as efficient as the underlying link (local span) recovery. In such a case, the span can be either protected or unprotected, but the LSP it carries MUST be (at least locally) recoverable. Therefore, the span recovery process can either be independent when protected (or restorable), or triggered by the upper LSP recovery process. The former requires coordination in order to achieve subsequent LSP recovery. Therefore, in order to achieve robustness and fast convergence, multi-layer recovery requires a fine-tuned coordination mechanism.

Moreover, in the absence of adequate recovery mechanism coordination (pre-determined for instance by the hold-off timer), a failure notification may propagate from one layer to the next within a recovery hierarchy. This can cause "collisions" and trigger simultaneous recovery actions that may lead to race conditions and in turn, reduce the optimization of the resource utilization and/or generate global instabilities in the network (see [MANCHESTER]). Therefore, a consistent and efficient escalation strategy is needed to coordinate recovery across several layers.

Therefore, one can expect that the definition of the recovery

mechanisms and protocol(s) is technology independent such that they can be consistently implemented at different layers; this would in turn simplify their global coordination. Moreover, as mentioned in

[TE-RH], some looser form of coordination and communication between (vertical) layers such a consistent hold-off timer configuration (and setup through signalling during the working LSP establishment) can be considered in this context, allowing synchronization between recovery actions performed across these layers.

Note: Recovery Granularity

In most environments, the design of the network and the vertical distribution of the LSP bandwidth are such that the recovery granularity is finer for higher layers. The OTN and SDH/Sonet layers can only recover the whole section or the individual connections it transports whereas IP/MPLS layer(s) can recover individual packet LSPs or groups of packet LSPs.

Obviously, the recovery granularity at the sub-wavelength (i.e. Sonet/SDH) level can be provided only when the network includes devices switching at the same granularity level (and thus not with optical channel switching capable devices). Therefore, the network layer can deliver control-plane driven recovery mechanisms on a per-LSP basis if and only if the LSPs class has the corresponding switching capability at the transport plane level.

7.3 Escalation Strategies

There are two types of escalation strategies (see [DEMEESTER]): bottom-up and top-down.

The bottom-up approach assumes that lower layer recovery schemes are more expedient and faster than the upper layer one. Therefore we can inhibit or hold-off higher layer recovery. However this assumption is not entirely true. Imagine a Sonet/SDH based protection mechanism (with a less than 50 ms protection switching time) lying on top of an OTN restoration mechanism (with a less than 200 ms restoration time). Therefore, this assumption should be (at least) clarified as: lower layer recovery schemes are faster than upper level one but only if the same type of recovery mechanism is used at each layer (assuming that the lower layer one is faster).

Consequently, taking into account the recovery actions at the different layers in a bottom-up approach, if lower layer recovery mechanisms are provided and sequentially activated in conjunction with higher layer ones, the lower layers MUST have an opportunity to

recover normal traffic before the higher layers do. However, if lower layer recovery is slower than higher layer recovery, the lower layer MUST either communicate the failure related information to the higher layer(s) (and allow it to perform recovery), or use a hold-off timer in order to temporarily set the higher layer recovery action in a ôstandby modeö. Note that the a priori information exchange between layers concerning their efficiency is not within the current scope of this document. Nevertheless, the coordination functionality between layers must be configurable and tunable.

An example of coordination between the optical and packet layer control plane enables for instance letting the optical layer performing the failure management operations (in particular, failure detection and notification) while giving to the packet layer control plane the authority to perform the recovery actions. In case of packet layer unsuccessful recovery action, fallback at the optical layer can be subsequently performed.

The Top-down approach attempts service recovery at the higher layers before invoking lower layer recovery. Higher layer recovery is service selective, and permits "per-CoS" or "per-connection" re-routing. With this approach, the most important aspect is that the upper layer must provide its own reliable and independent failure detection mechanism from the lower layer.

The same reference suggests also recovery mechanisms incorporating a coordinated effort shared by two adjacent layers with periodic status updates. Moreover, at certain layers, some of these recovery operations can be pre-assigned, e.g. a particular link will be handled by the packet layer while another will be handled by the optical layer.

7.4 Disjointness

Having link and node diverse working and recovery LSPs/spans does not guarantee working and recovery LSPs/Spans disjointness. Due to the common physical layer topology (passive), additional hierarchical concepts such as the Shared Risk Link Group (SRLG) and mechanisms such as SRLG diverse path computation must be developed to provide a complete working and recovery LSP/span disjointness (see [IPO-IMP] and [CCAMP-SRLG]). Otherwise, a failure affecting the working LSP/span would also potentially affect the recovery LSP/span resources, one refers to such event as a common failure.

7.4.1 SRLG Disjointness

A Shared Risk Link Group (SRLG) is defined as the set of optical spans (or links or optical lines) sharing a common physical resource (for instance, fiber links, fiber trunks or cables) i.e. sharing a common risk. For instance, a set of links L belongs to the same SRLG s, if they are provisioned over the same fiber link f.

The SRLG properties can be summarized as follows:

1) A link belongs to more than one SRLG if and only if it crosses one of the resources covered by each of them.

2) Two links belonging to the same SRLG can belong individually to (one or more) other SRLGs.

3) The SRLG set S of an LSP is defined as the union of the individual SRLG s of the individual links composing this LSP.

SRLG disjointness for LSP:

The LSP SRLG disjointness concept is based on the following postulate: an LSP (i.e. sequence of links) covers an SRLG if and only if it crosses one of the links belonging to that SRLG.

Therefore, the SRLG disjointness for LSPs can be defined as follows: two LSPs are disjoint with respect to an SRLG s if and only if none of them covers simultaneously this SRLG.

While the LSP SRLG disjointness with respect of a set S of SRLGs is defined as follows: two LSPs are disjoint with respect to a set of SRLGs S if and only if the sets of SRLGs they cover are completely and mutually disjoint.

The impact on recovery is obvious: SRLG disjointness is a necessary (but not a sufficient) condition to ensure optical network survivability. With respect to the physical network resources, a working-recovery LSP/span pair must be SRLG disjoint in case of dedicated recovery type while a working-recovery LSP/span group must be SRLG disjoint in case of shared recovery.

7.4.2 SRG Disjointness

By extending the previous definition from a link to a more generic structure, referred to as a ôrisk domainö, one comes to the SRG (Shared Risk Group) notion (see [CCAMP-SRG]). A risk domain is a group of arbitrarily connected nodes and spans that together can provide certain like-capabilities (such as a chain of dedicated/

shared protected links and nodes, or a ring forming nodes and links, or a protected hierarchical TE Link).

In turn, an SRG represents the risk domain capabilities and other parameters, which assist in computing diverse paths through the domain (it can also be used in assessing the risk associated with the risk domain.)

Note that the SRLG set of a risk domain constitutes a subset of the SRGs. SRLGs address only risks associated with the links (physical) and passive elements within the risk domain, whereas SRGs may contain nodes and other topological information in addition to the links. The key difference between an SRLG and an SRG is that an SRLG translates to only one link share risk with respect to server layer topology (even hierarchical TE Links) while an SRG translates a sequence of SRLGs over the same layer from one source to one or more than one destination located within the same area.

As for SRLG disjointness, the impact on recovery is that SRG disjointness is a necessary (but not a sufficient) condition to ensure optical network survivability. With respect to the physical and logical network resources (and topology), a working-recovery LSP/span pair must be SRG disjoint in case of dedicated recovery

type while a working-recovery LSP/span group must be SRG disjoint in case of shared recovery.

8. Recovery Scheme/Strategy Selection

In order to provide a structured selection and analysis of the recovery scheme/strategy, the following dimensions can be defined:

1. Fast convergence (performance): provide a mechanism that aggregates multiple failures (this implies fast failure detection and correlation mechanisms) and fast recovery decision independently of the number of failures occurring in the optical network (implying also a fast failure notification).

2. Efficiency (scalability): minimize the switching time required for LSP/span recovery independently of number of LSPs/spans being recovered (this implies an efficient failure correlation, a fast failure notification and timely efficient recovery mechanism(s)).

3. Robustness (availability): minimize the LSP/span downtime independently of the underlying topology of the transport plane (this implies a highly responsive recovery mechanism).

4. Resource optimization (optimality): minimize the resource capacity, including LSP/span and nodes (switching capacity), required for recovery purposes; this dimension can also be referred to as optimize the sharing degree of the recovery resources.

5. Cost optimization: provide a cost-effective recovery strategy.

However, these dimensions are either out of the scope of this document such as cost optimization and recovery path computational aspects or going in opposite directions. For instance, it is obvious that providing a 1+1 recovery type for each LSP minimizes the LSP downtime (in case of failure) while being non-scalable and recovery resource consuming without enabling any extra-traffic.

The following sections try to provide a first response in order to select a recovery strategy with respect to the dimensions described above and the recovery schemes proposed in [CCAMP-TERM].

8.1 Fast Convergence (Detection/Correlation and Hold-off Time)

Fast convergence is related to the failure management operations. It refers to the elapsing time between the failure detection/ correlation and hold-off time, point at which the recovery switching actions are initiated. This point has been already discussed in Section 4.

8.2 Efficiency (Switching Time)

In general, the more pre-assignment/pre-planning of the recovery LSP/span, the more rapid the recovery scheme is. Since protection implies pre-assignment (and cross-connection in case of LSP recovery) of the protection resources, in general, protection schemes recover faster than restoration schemes.

Span restoration (since using control plane) is also likely to be slower than most span protection types; however this greatly depends on the span restoration signalling efficiency. LSP Restoration with pre-signaled and pre-selected recovery resources is likely to be faster than fully dynamic LSP restoration, especially because of the elimination of any potential crank-back during the recovery LSP establishment.

If one excludes the crank-back issue, the difference between dynamic and pre-planned restoration depends on the restoration path computation and path selection time. Since computational considerations are outside of the scope of this document, it is up to the vendor to determine the average path computation time in different scenarios and to the operator to decide whether or not dynamic restoration is advantageous over pre-planned schemes depending on the network environment. This difference depends also on the flexibility provided by pre-planned restoration with respect to dynamic one: the former implies a limited number of failure scenarios (that can be due for instance to local storage limitation). This, while the latter enables an on-demand path computation based on the information received through failure notification and as such more robust with respect to the failure scenario scope.

Moreover, LSP segment restoration, in particular, dynamic restoration (i.e. no path pre-computation so none of the recovery resource is pre-signaled) will generally be faster than end-to-end LSP schemes. However, local LSP restoration assumes that each LSP segment end-point has enough computational capacity to perform this operation while end-to-end requires only that LSP end-points provides this path computation capability.

Recovery time objectives for Sonet/SDH protection switching (not including time to detect failure) are specified in [G.841] at 50 ms, taking into account constraints on distance, number of connections involved, and in the case of ring enhanced protection, number of nodes in the ring. Recovery time objectives for restoration mechanisms have been proposed through a separate effort [TE-RH].

8.3 Robustness

In general, the less pre-assignment (protection)/pre-planning (restoration) of the recovery LSP/span, the more robust the recovery type/scheme is to a variety of (single) failures, provided that adequate resources are available. Moreover, the pre-selection of the

recovery resources gives less flexibility for multiple failure scenarios than no recovery resource pre-selection. For instance, if failures occur that affect two LSPs sharing a common link along their restoration paths, then only one of these LSPs can be recovered. This occurs unless the restoration path of at least one of these LSPs is re-computed or the local resource assignment is modified on the fly.

In addition, recovery schemes with pre-planned recovery resources, in particular spans for protection and LSP for restoration purposes, will not be able to recover from failures that simultaneously affect both the working and recovery LSP/span. Thus, the recovery resources should ideally be chosen to be as disjoint as possible (with respect to link, node and SRLG) from the working ones, so that any single failure event will not affect both working and recovery LSP/span. In brief, working and recovery resource must be fully diverse in order to guarantee that a given failure will not affect simultaneously the working and the recovery LSP/span. Also, the risk of simultaneous failure of the working and restoration LSP can be reduced by re-computing a restoration path whenever a failure occurs along the corresponding recovery LSP or by re-computing a restoration path and re-provisioning the corresponding recovery LSP whenever a failure occurs along a working LSP/span. This method enables to maintain the number of available recovery path constant.

The robustness of a recovery scheme is also determined by the amount of reserved (i.e. signaled) recovery resources within a given shared resource pool: as the amount of recovery resources sharing degree increases, the recovery scheme becomes less robust to multiple failure occurrences. Recovery schemes, in particular restoration, with pre-signaled resource reservation (with or without pre-selection) should be capable to reserve the adequate amount of resource to ensure recovery from any specific set of failure events, such as any single SRLG failure, any two SRLG failures etc.

8.4 Resource Optimization

It is commonly admitted that sharing recovery resources provides network resource optimization. Therefore, from a resource utilization perspective, protection schemes are often classified with respect to their degree of sharing recovery resources with respect to the working entities. Moreover, non-permanent bridging protection types allow (under normal conditions) for extra-traffic over the recovery resources.

From this perspective 1) 1+1 LSP/Span protection is the more resource consuming protection type since it doesnÆt allow for any extra-traffic 2) 1:1 LSP/span protection type requires dedicated recovery LSP/span allowing carrying extra preemptible traffic 3) 1:N and M:N LSP/span recovery types require 1 (or M, respectively) recovery LSP/span (shared between the N working LSP/span) while allowing carrying extra preemptible traffic. Obviously, 1+1 protection precludes and 1:1 recovery type does not allow for

recovery LSP/span sharing whereas 1:N and M:N recovery types do allow sharing of 1 (M, respectively) recovery LSP/spans between N working LSP/spans.

However, despite the fact that the 1:1 recovery type does not allow recovery LSP/span sharing, the recovery schemes (see Section 5.4) that can be built from them (e.g.(1:1)^n) do allow for sharing of recovery resources these entities includes. In addition, the flexibility in the usage of shared recovery resources (in particular, shared links) may be limited because of network topology restrictions, e.g. fixed ring topology for traditional enhanced protection schemes.

On the other hand, in restoration with pre-signaled resource reservation, the amount of reserved restoration capacity is determined by the local bandwidth reservation policies. In restoration schemes with re-provisioning, a pool of restoration resource can be defined from which all (spare) restoration resources are selected after failure occurrence for recovery path computation purpose. The degree to which restoration schemes allow sharing amongst multiple independent failures is then directly dictated by the size of the restoration pool. Moreover, in all restoration schemes, spare resources can be used to carry preemptible traffic (thus over preemptible LSP/span) when the corresponding resources have not been committed for LSP/span recovery purposes.

From this, it clearly follows that less recovery resources (i.e. LSP/spans and switching capacity) have to be allocated to a shared recovery resource pool if a greater sharing degree is allowed. Thus, the degree to which the network is survivable is determined by the policy that defines the amount of reserved (shared) recovery resources and the maximum sharing degree allowed.

8.4.1. Recovery Resource Sharing

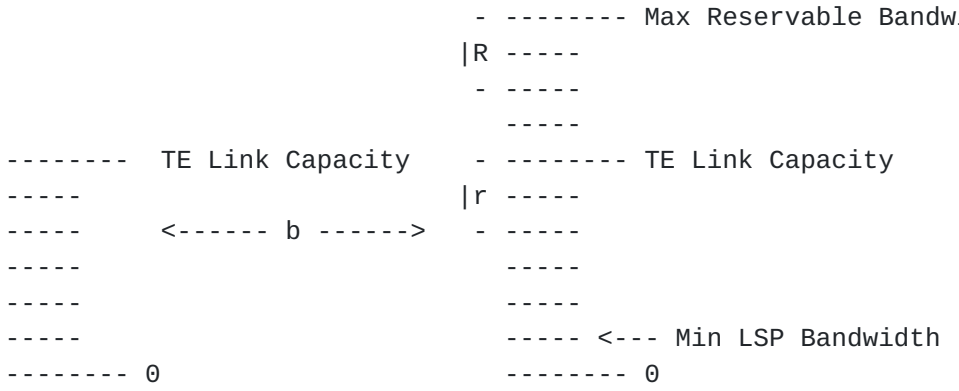
When recovery resources are shared over several LSP/Spans, [GMPLS-RTG], the use of the Maximum LSP Bandwidth, the Maximum Reservable Bandwidth and the Unreserved Bandwidth TE Link sub-TLVs provides only part of the information needed to obtain the optimization of the network resources allocated for shared recovery purposes.

Here, one has to additionally consider a recovery resource sharing ratio (or degree) in order to optimize the shared resource usage, since the distribution of the bandwidth utilization per component Link ID over a given TE Link is by definition unknown. For this purpose, we define the difference between Maximum Reservable Bandwidth (for recovery) and the Maximum Capacity per TE Link i as the Maximum Sharable Bandwidth or max_R[i]. Within this quantity, the amount of bandwidth currently allocated for shared recovery per TE Link i is defined as R[i]. Both quantities are expressed in terms of component link bandwidth unit (and thus equivalently the Min LSP

Bandwidth is of one bandwidth unit).

From these definitions, it results that the usage of this information available per TE Link can be considered in order to optimize the usage of the resources allocated (per TE Link) for shared recovery. If one refers to r[i] as the actual bandwidth per TE Link i (in terms of per component bandwidth unit) committed for shared recovery, then the following quantity must be maximized over the potential TE Link candidates: sum {i=1}^N [(R{i} + r{i})/(t{i} û b{i})] or equivalently: sum {i=1}^N [(R{i} + r{i})/r{i}] with R{i} >= 1 and r{i} >= 1 (in terms of per component bandwidth unit). In this formula, N is the total number of links traversed by a given LSP, t[i] the Maximum LSP Bandwidth per TE Link i and b[i] the sum per TE Link i of the bandwidth committed for working LSPs and dedicated recovery. The quantity [(R{i} + r{i})/r{i}] is defined as the Shared (Recovery) Bandwidth Ratio per TE Link i. In addition, TE Links for which R[i] = max_R[i] or for which r[i] = 0 are pruned during recovery path computation. Note also that the TE Links for which R[i] = max_R[i] = r[i] can not be shared more than twice (their sharing ratio equals 2).

More generally, one can draw the following mapping between the available bandwidth at the transport and control plane level:



Note that the above approach does not require the flooding of any per LSP information or a detailed distribution of the bandwidth allocation per component link (or individual ports). Moreover, it has been demonstrated that this Partial Information Routing approach can also be extended to resource shareability with respect to the number of times each SRLG is protected by a recovery resource, in particular an LSP (see also Section 8.4.2). This method also referred to as stochastic approach is described in [BOUILLET]. By flooding this summarized information using a link-state protocol,

recovery path computation and selection for SRLG diverse recovery paths can be optimized with respect to resource sharing giving a performance difference of less than 5% compared to a Full Information Flooding approach. The latter can be found in [GLI] for instance. Strictly speaking both methods rely on deterministic knowledge of the network topology and resource (usage) status.

For GMPLS-based recovery purposes, the Partial Information Routing approach can be further enhanced by extending GMPLS signalling capabilities. This, by allowing the working LSP related information

and in particular, its explicit route to be exchanged over the recovery LSP in order to enable more efficient admission control at shared (link) resource upstream nodes.

8.4.2 Recovery Resource Sharing and SRLG Disjointness

As stated in the previous section, resource shareability should be maximized with respect to the number of times each SRLG is protected by a recovery resource.

Methods can be considered for avoiding contention for the shared recovery resources during a single SRLG/node failure (see Section 5). These allow the sharing of common reserved recovery resource between two (or more) recovery LSPs (only) if their respective working LSPs are mutually disjoint with respect to link, node or SRLG. A single failure then does not disrupt several (at least two) working LSPs simultaneously.

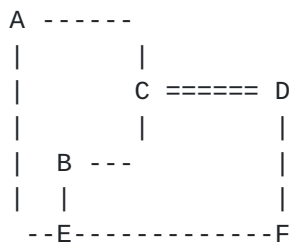
For this purpose, additional extensions to [GMPLS-RTG] in support of the path computation for shared mesh restoration may be considered. First, the information about the recovery resource sharing on a TE link such as the current number of recovered LSPs sharing the recovery resources reserved on the TE link (see also Section 8.4.1) and the current number of SRLGs recovered by this amount of shared recovery resource on the TE link, may be considered. The latter is equivalent to the total number of SRLGs that the (recovery) LSPs sharing the recovery resources shall recover. Then, if SRLG-disjointness has to be considered under strong recovery guarantee in the event of a single SRLG failure, the explicit list of SRLGs recovered by the currently recovery resources shared on the TE link together with their respective sharable recovery bandwidth (see also Section 8.4.1). The latter information is equivalent to the maximum sharable recovery bandwidth per SRLG or per group of SRLG (thus one considers a decreasing amount of sharable bandwidth and SRLG list over time).

Note: it has to be emphasized that a per (group of) SRLG maximum sharable recovery bandwidth is restricted by the length that the corresponding (sub-)TLV may take and thus the number of SRLGs that it can include.

Therefore, compared to the case of simple recovery resource sharing regardless of SRLG disjointness (as described in Section 8.4.1), the additional TE link information considered here should allow for better path selection (at distinct ingress node) during SRLG-disjoint LSP provisioning in shared meshed recovery scheme. The next section will demonstrate that such extensions are complementary to the exchange of the explicit route of working LSP over the recovery LSP path in order to achieve shared recovery resource contention avoidance.

8.4.3 Recovery Resource Sharing, SRLG Disjointness and Admission Control

Admission control is a strict requirement to be fulfilled by nodes giving access to shared links. This can be illustrated using the following recovery scheme:



Node A creates a working LSP to D, through C only, B creates simultaneously a working LSP to D through C and a recovery LSP (through E and F) to the same destination. Then, A decides to create a recovery LSP to D, but since C to D span carries both working LSPs node E should either assign a dedicated resource for this recovery LSP or if it has already reached its maximum shared recovery bandwidth level reject this request. Otherwise, in the latter case a C-D span failure would imply that one of the working LSP would not be recoverable.

Consequently, node E must have the required information (implying for instance that the explicit route followed by the primary LSPs to be carried with the corresponding recovery LSP request) in order to

perform an admission control for the recovery LSP requests.

Moreover, node E may securely (if its maximum shared recovery bandwidth ratio has not been reached yet for this link) accept the recovery LSP request and logically assign the same resource to these LSPs. This if and only if it can guarantee that A-C-D and B-C-D are SRLG disjoint over the C-D span (one considers here in the scope of this example, node failure probability as negligible). To achieve this, the explicit route of the primary LSP (and transported over the recovery path) is examined at each shared link ingress node. The latter uses the interface identifier as index to retrieve in the TE Link State DataBase (TE LSDB) the SRLG id list associated to the links of the working LSPs. If these LSPs have one or more SRLG id in common (in this example, one or more SRLG id in common over C-D), then node E should not assign the same resource to the recovery LSPs. Otherwise one of these working LSPs would not be recoverable in case of C-D span failure.

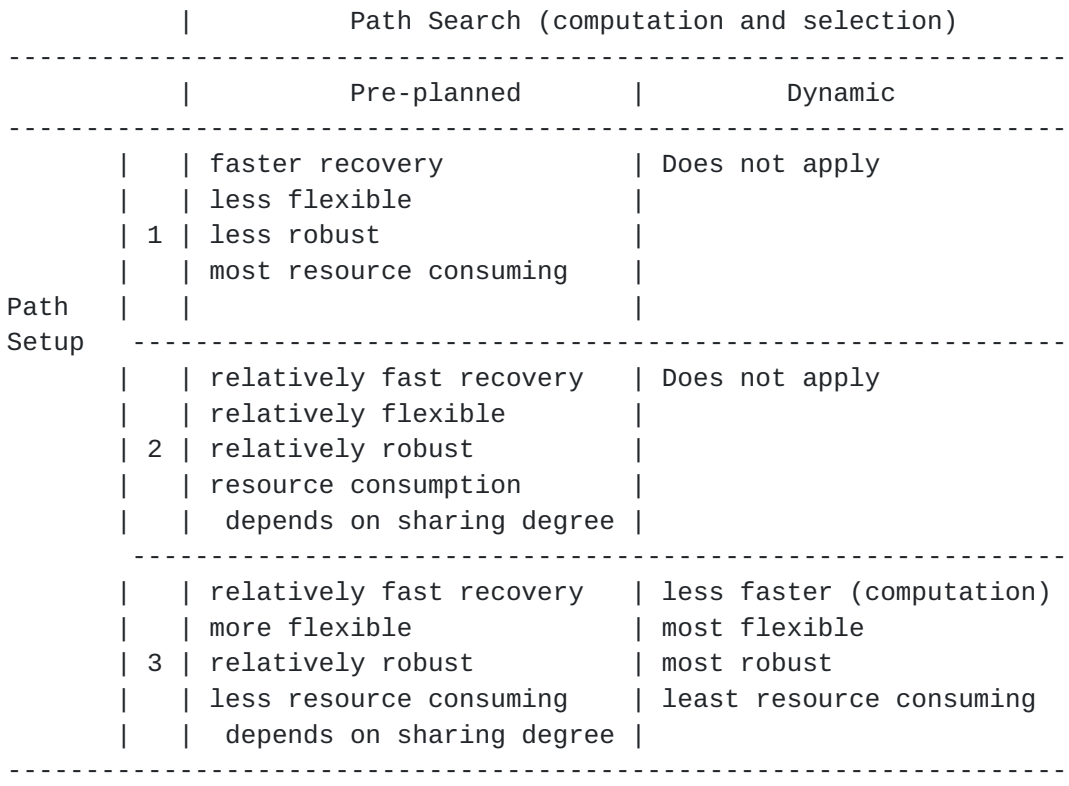
There are some issues related to this method, the major one being the number of SRLG Ids that a single link can cover (more than 100, in complex environments). Moreover, when using link bundles, this approach may generate the rejection of some recovery LSP requests. This because the SRLG sub-TLV corresponding to a link bundle

includes the union of the SRLG id list of all the component links belonging to this bundle (see [GMPLS-RTG] and [MPLS-BUNDLE]).

In order to overcome this specific issue, an additional mechanism may consist of querying the nodes where such an information would be available (in this case, node E would query C). The major drawback of this method, in addition to the dedicated mechanism it requires, is that it may become very complex when several common nodes are traversed by the working LSPs. Therefore, when using link bundles, a potential way of solving this issue tightly related to the sequence of the recovery operations (at least in a first step, since per component flooding of SRLG id would impact the link state routing protocol scalability), is to rely on the usage of dedicated queries to an on-line accessible network management system.

8.5 Summary

One can summarize by the following table the selection of a recovery scheme/strategy, using the recovery types proposed in [CCAMP-TERM] and the above discussion.



1. Path Setup with Resource Reservation (i.e. signalling) and Selection
2. Path Setup with Resource Reservation (i.e. signalling) w/o Selection
3. Path Setup w/o Resource Reservation (i.e. signalling) w/o Selection

As defined in [CCAMP-TERM], the term pre-planned refers to restoration resource pre-computation, signaling (reservation) and a priori selection (optional), but not cross-connection.

9. Conclusion

TBD.

10. Security Considerations

This document does not introduce or imply any specific security consideration.

11. References

[RFC-2026] S.Bradner, ôThe Internet Standards Process -- Revision 3ö, BCP 9, RFC 2026, October 1996.

[RFC-2119] S.Bradner, ôKey words for use in RFCs to Indicate Requirement Levelsö, BCP 14, RFC 2119, March 1997.

[BOUILLET] E.Bouillet et al., ôStochastic Approaches to Compute Shared Meshed Restored Lightpaths in Optical Network Architecturesö, INFOCOM 2002, New York City, June 2002.

[CCAMP-LI] G.Li et al. ôRSVP-TE Extensions For Shared-Mesh Restoration in Transport Networksö, Internet Draft, Work in progress, draft-li-shared-mesh-restoration-01.txt, November 2001.

[CCAMP-LIU] H.Liu et al. ôOSPF-TE Extensions in Support of Shared Mesh Restorationö, Internet Draft, Work in progress, draft-liu-gmpls-ospf-restoration-00.txt, October 2002.

[CCAMP-SRLG] D.Papadimitriou et al., ôShared Risk Link Groups Encoding and Processing,ö Internet Draft, Work in progress, draft-papadimitriou-ccamp-srlg-processing-01.txt, November 2002.

[CCAMP-SRG] S.Dharanikota et al., ôInter domain routing with Shared Risk Groups,ö Internet Draft, Work in progress, November 2001.

[CCAMP-TERM] E.Mannie and D.Papadimitriou (Editors), ôRecovery (Protection and Restoration) Terminology for GMPLS,ö Internet Draft, Work in progress, draft-ietf-ccamp-gmpls-recovery-terminology-00.txt, June 2002.

[DEMEESTER] P.Demeester et al., ôResilience in Multilayer Networksö, IEEE Communications Magazine, Vol. 37, No. 8, August 1998, pp. 70-76.

[G.707] ITU-T, ôNetwork Node Interface for the Synchronous Digital Hierarchy (SDH)ö, Recommendation G.707, October 2000.

[G.709] ITU-T, ôNetwork Node Interface for the Optical Transport Network (OTN)ö, Recommendation G.709, February 2001 (and Amendment n 1, October 2001).

[G.783] ITU-T, ôCharacteristics of Synchronous Digital Hierarchy (SDH) Equipment Functional Blocksö, Recommendation G.783, October 2000.

[G.798] ITU-T, ôCharacteristics of Optical Transport Network (OTN) Equipment Functional Blocksö, Recommendation G.798, January 2002.

[G.806] ITU-T, ôCharacteristics of Transport Equipment û Description Methodology and Generic Functionalityö, Recommendation G.806, October 2000.

[G.826] ITU-T, ôPerformance Monitoringö, Recommendation G.826, February 1999.

[G.841] ITU-T, ôTypes and Characteristics of SDH Network Protection Architecturesö, Recommendation G.841, October 1998.

[G.842] ITU-T, ôInterworking of SDH network protection architecturesö, Recommendation G.842, October 1998.

[G.GPS] ITU-T Draft Recommendation G.GPS, Version 2, ôGeneric Protection Switchingö, Work in progress, May 2002.

[GLI] Guangzhi Li et al., ôEfficient Distributed Path Selection for Shared Restoration Connectionsö, IEEE Infocom, New York, June 2002.

[GMPLS-ARCH] E.Mannie (Editor), ôGeneralized MPLS Architectureö, Internet Draft, Work in progress, draft-ietf-ccamp-gmpls-architecture-03.txt, August 2002.

[GMPLS-RTG] K.Kompella et al., ôRouting Extensions in Support of Generalized MPLS,ö Internet Draft, Work in Progress, draft-ietf-ccamp-gmpls-routing-05.txt, August 2002.

[GMPLS-SIG] L.Berger (Editor), ôGeneralized MPLS û Signaling Functional Descriptionö, Internet Draft, Work in progress, draft-ietf-mpls-generalized-signaling-09.txt, October 2002.

[LMP] J.Lang (Editor), ôLink Management Protocol (LMP) v1.0ö Internet Draft, Work in progress, draft-ietf-ccamp-lmp-06, September 2002.

[LMP-WDM] A.Fredette and J.Lang (Editors), ôLink Management Protocol (LMP) for DWDM Optical Line Systems,ö Internet Draft, Work in progress, draft-ietf-ccamp-lmp-wdm-01.txt, September 2002.

[MANCHESTER] J.Manchester, P.Bonenfant and C.Newton, ôThe Evolution of Transport Network Survivability,ö IEEE Communications Magazine, August 1999.

[MPLS-REC] V.Sharma and F.Hellstrand (Editors) et al., ôA Framework for MPLS Recoveryö, Internet Draft, Work in Progress, draft-ietf-mpls-recovery-frmwrk-06.txt, July 2002.

[MPLS-OSU] S.Seetharaman et al., ôIP over Optical Networks: A Summary of Issuesö, Internet Draft, Work in Progress, draft-osu-ipo-mpls-issues-02.txt, April 2001.

[T1.105] ANSI, "Synchronous Optical Network (SONET): Basic Description Including Multiplex Structure, Rates, and Formats", ANSI T1.105, January 2001.

[TE-NS] K.Owens et al., ôNetwork Survivability Considerations for Traffic Engineered IP Networksö, Internet Draft, Work in Progress, draft-owens-te-network-survivability-01.txt, July 2001.

[TE-RH] W.Lai, D.McDysan, J.Boyle, et al., ôNetwork Hierarchy and Multi-layer Survivabilityö, Internet Draft, Work in Progress, draft-ietf-tewg-restore-hierarchy-01.txt, June 2002.

12. Acknowledgments

The authors would like to thank Fabrice Poppe (Alcatel) and Bart Rousseau (Alcatel) for their revision effort, Richard Rabbat (Fujitsu), David Griffith (NIST) and Lyndon Ong (Ciena) for their useful comments.

13. Author's Addresses

Deborah Brungard (AT&T)
Rm. D1-3C22
200 S. Laurel Ave.
Middletown, NJ 07748, USA
Email: dbrungard@att.com

Sudheer Dharanikota (Nayna)
481 Sycamore Drive
Milpitas, CA 95035, USA
Email: sudheer@nayna.com

Jonathan P. Lang (Calient)
25 Castilian
Goleta, CA 93117, USA
Email: jplang@calient.net

Guangzhi Li (AT&T)
180 Park Avenue,
Florham Park, NJ 07932, USA
Email: gli@research.att.com
Phone: +1 973 360-7376

Eric Mannie (Consulting)
Email: eric_mannie@hotmail.com

Dimitri Papadimitriou (Alcatel)
Francis Wellesplein, 1
B-2018 Antwerpen, Belgium
Phone: +32 3 240-8491
Email: dimitri.papadimitriou@alcatel.be

Bala Rajagopalan (Tellium)
2 Crescent Place
P.O. Box 901
Oceanport, NJ 07757-0901, USA
Phone: +1 732 923-4237
Email: braja@tellium.com

Yakov Rekhter (Juniper)
Email: yakov@juniper.net

Full Copyright Statement