Traffic Engineering Working Group
Internet Draft
<draft-ietf-tewg-restore-hierarchy-00.txt>
Category: Informational
Expiration Date: March 2002

Wai Sum Lai, AT&T
Dave McDysan, WorldCom
(Co-Editors)

Jim Boyle, Protocol
Driven Networks
Malin Carlzon
Rob Coltun, Redback
Tim Griffin, AT&T
Ed Kern, Cogent
Tom Reddington, Lucent

September 2001

Network Hierarchy and Multilayer Survivability

Status of this Memo

1. Abstract

This document is the deliverable out of the Network Hierarchy and Survivability Techniques Design Team established within the Traffic Engineering Working Group. This team was requested to try to determine what the current and near term requirements are for survivability and hierarchy in service provider environments. The team determined that there appears to be a need for common, interoperable survivability approaches in packet and non-packet networks. Suggested approaches include path-based as well as one that repairs connections in proximity to the network fault. For clarity, an expanded set of definitions is included. As for hierarchy, there did not appear to be as much need for work on "vertical hierarchy," defined as communication between network layers such as TDM/optical and MPLS. In particular, instead of direct exchange of signaling and routing between vertical layers,

some looser form of coordination and communication is a nearer term need. For "horizontal hierarchy" in data networks, there does appear to be a pressing need. This requirement is often presented in the context of layer 2 and layer 3 VPN services where SLAs would appear to necessitate signaling from the edges into the core of a network. Issues include potential current protocols limitations in networks which are hierarchical (e.g. multi-area OSPF) and scalability concerns of potentially O(N^2) connection growth in larger networks.

Please send comments to te-wg@ops.ietf.org

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 presents a proposal of the tangible requirements for network survivability and hierarchy in current service provider environments. With feedback from the working group solicited, the objective is to help focus the work that is being addressed in the TEWG (Traffic Engineering Working Group), CCAMP (Common Control and Measurement Plane Working Group), and other working groups. A main goal of this work is to provide some expedience for required functionality in multi-vendor service provider networks. The initial focus is primarily on intra-domain operations. However, to maintain consistency in the provision of end-to-end service in a multi-provider environment, rules governing the operations of survivability mechanisms at domain boundaries must also be specified. While such issues are raised and discussed, where appropriate, they will not be treated in depth in the initial release of this document.

The document first develops a set of definitions to be used later in this document and potentially in other documents as well. It then addresses the requirements and issues associated with service restoration, hierarchy, and finally a short discussion of survivability in hierarchical context.

4. Terminology and Concepts

[Editor's note: The terminology and concepts sections should preferably be read in parallel, but unfortunately, the presentation is necessarily sequential. The order of terminology followed by concepts is chosen to simplify the description with terms already defined.]

4.1 Hierarchy Terminology

Network hierarchy is an abstraction of part of a network's topology and the routing and signaling mechanism needed to support the topological abstraction. Abstraction may be used as a mechanism to build large networks or as a technique for enforcing administrative, topological or geographic boundaries. For example, network hierarchy might be used to separate the metropolitan and long-haul regions of a network or to separate the regional and backbone sections of a network [Bert Wijnen], or to interconnect service provider networks (with BGP which reduces a network to an Autonomous System). In this document, network hierarchy is considered from two perspectives:
(1) Horizontally oriented: between two areas or administrative subdivisions within the same network layer
(2) Vertically oriented: between two network layers

Horizontal hierarchy is the abstraction necessary to allow a network at one network layer, for instance a packet network, to grow. Examples of horizontal hierarchy include BGP and multi-area OSPF.

Vertical hierarchy is the abstraction, or reduction in information, which would be of benefit when communicating information across network layers, as in propagating information between optical and router networks.

4.2 Hierarchy Concepts

[Editor's note: It was suggested to change the term "hierarchy" to "partition." This section is added to clarify the meaning of these two terms. Partition is a means to an end. Hierarchy captures the significance of time scale and levels of abstraction - essential concepts that provide guidance to the process of partitioning.]

Hierarchy is a technique to build scalable complex systems. New levels of hierarchical controls are added to oversee and manage the increasing complexity of a system, and to allow for effective open-ended growth. To provide a parsimonious and meaningful description of the entire system, it is necessary to abstract at each level what is most significant from the details of the levels below. This is also needed to clarify relationships between parts and wholes of hierarchical systems.

The above approach is based on a general property of all hierarchical systems that interactions between systems decrease in strength with distance. This is because the internal interactions of the components of the lower-level subsystems in a hierarchy tend to occur at a smaller time scale with a higher frequency. On a suitably larger time scale, these interactions will be so rapid that the corresponding subsystems will appear to be in equilibrium. The observable dynamics of interaction of the major subsystems will then be nearly independent of the detail of the internal structure of the

subsystems. The extent to which lower-level details can be ignored depends largely on how sharply the different time scales can be delineated. Using such an approximation, a hierarchical structure is developed by ordering the interactions among different functions or components according to their orders of magnitude.

In the vertical hierarchy, the total network functions are partitioned into a series of functional layers with clear logical, and may be even physical, separation between adjacent layers. In this structuring, a general goal is for lower layers to have faster restoration times than higher layers.

In the horizontal hierarchy, a large network is partitioned into multiple smaller, non-overlapping sub-networks. The partitioning criteria can be based on topology, network function, administrative policy, or service domain demarcation. Two networks at the *same* hierarchical level, e.g., two Autonomous Systems in BGP, may share a peer relation with each other through some loose form of coupling. On the other hand, for routing in large networks using multi-area OSPF, abstraction through the aggregation of routing information is achieved through a hierarchical partitioning of the network.

4.3 Survivability Terminology

[Editor's note. This section needs more work to present a natural order.]

Extra traffic is the traffic carried over the protection entity while the working entity is active. Extra traffic is not protected, i.e., when the protection entity is required to protect the traffic that is being carried over the working entity (e.g., due to a failure affecting the working entity), the extra traffic is preempted.

Normalization is the return to the normal state of a network upon completing the repair of the network failure. This could include the rerouting of affected traffic to the original working entities or new routes. The term revertive mode is used when traffic is returned to the working entity (switch back).

Protection, also called protection switching, is a survivability technique based on predetermined failure recovery: as the working entity is established, resources are reserved for the protection entity. These resources may be used by low-priority traffic (referred to as extra traffic) if traffic preemption is allowed. Depending on the amount of reserved resources, not all of the affected traffic may be protected. (For further discussion of concepts related to protection, see the Sub-section below on Survivability Concepts.)

Protection entity (also called back-up entity or recovery entity) is the entity that is used to carry protected traffic in protection

operation mode, i.e., when the working entity is in error or has failed.

Recovery is the sequence of actions taken by a network after the detection of a failure to maintain the required performance level for existing services (e.g., according to service level agreements) and to allow normalization of the network. The actions include notification of the failure followed by two parallel processes: (1) a repair process with fault isolation and repair of the failed components, and (2) a reconfiguration process with path selection and rerouting for the affected traffic. (For a description of the sequence of events of how network failures are monitored, detected, and mitigated, see [MPLS Recovery Framework].)

Rerouting is placement of affected traffic from the working entity to the protection entity, when the path for the protection entity has been selected after the detection of a fault on the working entity. ~~This~~ In protection techniques, rerouting is synonymous with switch-over ~~in protection techniques~~. For dynamic restoration, the "protection" resources for rerouting are the currently unassigned (unreserved) resources in that layer. (In [3], rerouting is synonymous with restoration.)

Restoration is a survivability technique that dynamically discovers the alternate path from spare resources in network, or establishes new paths on demand, for affected traffic once the failure is detected and the affected traffic is identified for rerouting. The new path may be based on preplanned configurations or current network status. Thus, restoration involves a path selection process followed by traffic rerouting. (In [3], restoration is referred to as recovery by rerouting.)

Restoration, or more specifically, service restoration, refers to the actions taken by a network to maintain service continuity after the detection of a failure. In this second usage, restoration has a meaning very similar to recovery, except that restoration covers only the reconfiguration process and not the repair process. Also, in this usage, it should be clear from the context that it is irrelevant whether the survivability technique used to achieve service continuity is based on protection or restoration techniques.

Restoration priority is a method of giving preference to protect higher-priority traffic ahead of lower-priority traffic. Its use is to help determine the order of rerouting traffic after a failure has occurred. The purpose is to differentiate service restoration time as well as to control access to available spare capacity for different classes of traffic.

Restoration time is the time interval from the occurrence of a network impairment to the instant when the affected traffic is either completely rerouted or until spare resources are exhausted and/or no more preemptable traffic to make room.

Revertive mode is a procedure in which revertive action, i.e., switch back from the protection entity to the working entity, is taken once the failed working entity has been repaired. In non-revertive mode, such action is not taken. To minimize service interruption, switch-back in revertive mode should be performed at a time when there is the least impact on the traffic concerned, or by using the make-before-break concept.

Shared risk group (SRG) is a set of network elements that are collectively impacted by a specific fault or fault type. For example, a shared risk link group (SRLG) is the union of all the links on those fibers that are routed in the same physical conduit in a fiber-span network. This concept includes, besides shared conduit, other types of compromise such as shared fiber cable, shared right of way, shared optical ring, shared office without power sharing, etc. The span of an SRG, such as the length of the sharing for compromised outside plant, needs to be considered on a per fault basis. (See [4] for further discussion.)

Survivability is the capability of a network to maintain service continuity in the presence of faults within the network [5]. Survivability techniques such as protection and restoration are implemented either on a per-link basis, on a per-path basis, or throughout an entire network to alleviate service disruption at affordable costs. The degree of survivability is determined by the network's capability to survive single failures, multiple failures, and equipment failures.

Working entity is the entity that is used to carry traffic in normal operation mode. Depending on the context, an entity can be, e.g., a channel or a transmission link in the physical layer, an LSP in MPLS, or a logical bundle of one or more LSPs.

4.4 Survivability Concepts

[Editor's note. This section needs to add text on restoration and references for restoration methods such as [6, 7, 8, …].]

In a survivable network design, spare capacity and diversity must be built into the network from the beginning to support some degree of self-healing whenever failures occur. A common strategy is to associate each working entity with a protection entity having either dedicated resources or shared resources that are pre-reserved or reserved-on-demand. According to the methods of setting up a protection entity, different approaches to providing survivability can be classified. Generally, protection techniques are based on having a dedicated protection entity set up prior to failure. Such is not the case in restoration techniques, which mainly rely on the use of spare capacity in the network. Hence, in terms of trade-offs, protection techniques usually offer fast recovery from failure with enhanced availability, while restoration techniques usually achieve better resource utilization.

Protection techniques can be implemented by several architectures: 1+1, 1:1, 1:n, and m:n. In the context of SDH/SONET, they are referred to as Automatic Protection Switching (APS).

In the 1+1 protection architecture, a protection entity is dedicated to each working entity. The dual-feed mechanism is used whereby the working entity is permanently bridged onto the protection entity at the source of the protected domain. In normal operation mode, identical traffic is transmitted simultaneously on both the working and protection entities. At the sink of the protected domain, both feeds are monitored for alarms and maintenance signals. A selection between the working and protection entity is made based on some predetermined criteria, such as the transmission performance requirements or defect indication. This architecture is rather expensive since resource duplication is required. It is generally used for specific services that need a very high availability.

In the 1:1 protection architecture, a protection entity is also dedicated to each working entity. The protected traffic is normally transmitted by the working entity. If the working entity has failed, the protected traffic is rerouted to the protection entity. This architecture is inherently slower in recovering from failure than a 1+1 architecture since communication between both ends of the protection domain is required to perform the switch-over operation. An advantage is that the protection entity can optionally be used to carry preemptable "extra traffic" in normal operation. Also, in packet networks, a protection path can be pre-established for later use with pre-planned but not pre-reserved capacity. (If no packets are sent into a link, no bandwidth is consumed.) This is not the case in channelized transport networks.

In the 1:n protection architecture, a dedicated protection entity is shared by n working entities. Traffic is normally sent on the working entities. When multiple working entities have failed simultaneously, only one of them can be restored by the common protection entity. This contention is resolved by assigning a different preemptive priority to each working entity. As in the 1:1 case, the protection entity can optionally be used to carry preemptable "extra traffic" in normal operation

The m:n architecture is a generalization of the 1:n architecture. Typically m <= n, m dedicated protection entities are shared by n working entities. While this architecture can improve system availability with small cost increases, it has rarely been implemented or standardized.

5. Survivability

5.1 Scope

Interoperable approaches to network survivability were determined to be an immediate requirement in packet networks as well as in SDH/SONET framed TDM networks. Not as pressing at this time were techniques which would cover all-optical networks (e.g., where framing is unknown), as the control of these networks in a multi-vendor environment appeared to have some other hurdles to first deal with. Also, not of immediate interest were approaches to coordinate or explicitly communicate survivability mechanisms across network layers (such as from a TDM or optical network to/from an IP network). However, a capability should be provided for a network operator to control the operation of survivability mechanisms among different layers. Such issues and those related to OAM are currently outside the scope of this document. (For proposed MPLS OAM requirements, see [9, 10]).

The types of network failures that cause a restoration to be performed include link/span and node failures (which might include span failures at lower layers). Other more complex failure mechanisms such as systematic control-plane failure or breach of security are not within the scope of the survivability mechanisms discussed in this document.

Other types of network failures (may be out of scope?): drop-side interface failure (e.g. between customer and access router, or a router to an OXC). These types of failures may be either inter-layer or inter-domain, depending on the actual network configuration.

5.2 Required initial set of survivability mechanisms

5.2.1 1:1 Path Protection with Pre-Established Capacity

In this protection mode, the head end of a working connection establishes a protection connection to the destination. There should be the ability to maintain relative restoration priorities between working and protection connections, as well as between different classes of protection connections.

In normal operation, traffic is only sent on the working connection, though the ability to signal that traffic will be sent on both connections (1+1 Path for signaling purposes) would be valuable in non-packet networks. Some distinction between working and protection connections is likely, either through explicit objects, or preferably through implicit methods such as general classes or priorities. Head ends need the ability to create connections that are as failure disjoint as possible from each other. This would require SRG information that can be generally assigned to either nodes or links and propagated through the control or management plane. In this mechanism, capacity in the protection connection is pre-established, however it can be used to carry preemptable extra traffic. ~~Protect capacity is first come first served.~~ When protect capacity is called into service during restoration, there should be

the ability to promote the protection connection to working status (for non-revertive mode operation) with some form of make-before-break capability.

5.2.2 1:1 Path Protection with Pre-Planned Capacity

Similar to the above 1:1 protection with pre-established capacity, the protection connection in this case is also pre-signaled. The difference is in the way protection capacity is assigned. With pre-planned capacity, the mechanism supports the ability for the protection capacity to be shared, or "double-booked". It would be expected that should operator predicted failures occur, which potentially could rely on enumeration in SRGs, that only a limited set of protection connections would be put into service, and that the protection capacity available in the network would be able to fulfill this traffic (given proper sizing and planning of the network). In a sense, this is 1:1 from a path perspective, however the protection capacity in the network (on a link by link basis) is shared in a 1:n fashion. Some form of information propagation could be required before traffic may be sent on protection connections, especially in TDM networks. In data networks, a desirable operating approach for this mechanism might be where the protection capacity is not accurately booked against SRGs (e.g. non-predictive).

The use of this approach improves network resource utilization, but may require more careful planning. So, initial deployment might be based on 1:1 path protection with pre-established capacity and the local restoration mechanism to be described next.

5.2.3 Local Restoration

Due to the time impact of signal propagation, path-based approaches may not be able to meet the service requirements desired in some networks. The solution to this is to restore connectivity in immediate proximity to the fault. At a minimum, this approach should be able to protect against connectivity-type SRGs, though protecting against node-based SRGs might be worthwhile. After local restoration is in place, it is likely that head end systems would later perform some path-level re-grooming.

Head end systems must have some control as to whether their connections are candidates for or excluded from local restoration. For example, best-effort and preemptable traffic may be excluded from local restoration; they only get restored if there is bandwidth available. This type of control may require an object in the signaling.

5.2.4 Path Restoration

In this approach, connections that are impacted by a fault are rerouted by the originating network element upon notification of connection failure. This source-based approach is efficient for network resources, but typically takes longer restoration times. It

does not involve any new mechanisms. It merely is a mention of another common approach to protecting against faults in a network.

5.3 Applications Supported

With service continuity under failure as a goal, a network is "survivable" if, in the face of a network failure, connectivity is interrupted for a brief period and then restored before the network failure ends. The length of this interrupted period is dependent on the application supported. Here are some typical applications that need to be considered:

- Best-effort data: restoration of network connectivity by rerouting at the IP layer would be sufficient
- Premium data service: need to meet TCP or application protocol timer requirements
- Voice: call cutoff is in the range of 140 msec to 2 sec
- Other real-time service (e.g., streaming, fax)
- Mission-critical applications

5.4 Timing Bounds for Service Restoration

The approach to picking the types of survivability mechanisms recommended was to consider a spectrum of mechanisms that can be used to protect traffic with varying characteristics of survivability and speed of restoration, and then attempt to select a few general points which provide some coverage across that spectrum. The focus of this work is to provide requirements to which a small set of detailed proposals may be developed, allowing the operator some (limited) flexibility in approaches to meeting their design goals in engineering multi-vendor networks. Requirements of different applications as listed in the previous sub-section were discussed generally, however none on the team would likely attest to the scientific merit of the ability of the timing bounds below to meet any specific application’s needs. A few assumptions include:

1. Approaches that protection switch without propagation of information are likely to be faster than those that do require some form of fault notification to some or all elements in a network.
2. Approaches that require some form of signaling after a fault will also likely suffer some timing impact.

Proposed timing bounds for service restoration for different mechanisms are as follows (all bounds are exclusive of signal propagation):

1:1 path protection with pre-established capacity:	100-500 ms
1:1 path protection with pre-planned capacity:	100-750 ms
Local restoration:	50 ms
Path restoration:	1-5 seconds

To ensure that the service requirements for different applications can be met within the above timing bounds, restoration priority is used to determine the order in which connections are restored (to minimize service restoration time as well as to gain access to available spare capacity). For example, mission critical applications may require high restoration priority. At the fiber layer, instead of specific applications, it may be possible that priority be given to certain classifications of customers with their traffic types enclosed within the customer aggregate. Preemption priority should only be used in the event that all connections cannot be restored, in which case connections with lower preemption priority should be released. Depending on a service provider's strategy in provisioning network resources for backup, preemption may or not be needed in the network.

5.5 Coordination Among Layers

A common design goal for multi-layered networks is to provide the desired level of service in the most cost-effective manner. The use of multilayer survivability might allow the optimization of spare resources through the improvement of resource utilization by sharing spare capacity across different layers, though further investigations are needed. Coordination during service restoration among different network layers (e.g. IP, SDH/SONET, optical layer) might necessitate development of vertical hierarchy. The benefits of providing survivability mechanisms at multiple layers, and the optimization of the overall approach, must be weighed with the associated cost and service impacts.

A default coordination mechanism for inter-layer interaction could be the use of nested timers and current SDH/SONET fault monitoring, as has been done traditionally for backward compatibility. Thus, when lower-layer restoration happens in a longer time period than higher-layer restoration, a hold-off timer is utilized to avoid contention between the different single-layer recovery schemes. In other words, multilayer interaction is addressed by having successively higher multiplexing levels operate at restoration time scale greater than the next lowest layer. Currently, if SDH/SONET protection switching is used, MPLS recovery timers must wait until SDH/SONET has had time to switch. On the contrary, if the lower layer does not have rerouting capability or is not expected to protect, say an unprotected SDH/SONET linear circuit, then there must be a mechanism for the lower layer to trigger the higher layer to take recovery actions immediately. This necessitates the allowance for adjustment of hold-off timer values.

It was felt that the current approach to coordination of survivability approaches currently did not have significant operational shortfalls. These approaches include protecting traffic solely at one layer (e.g. at the IP layer over linear WDM, or at the SDH/SONET layer). Where survivability mechanisms might be deployed at several layers, such as when a routed network rides a SDH/SONET protected network, it was felt that current coordination approaches

were sufficient in many cases. One exception is the hold-off of MPLS recovery until the completion of SDH/SONET protection switching as described above. This limits the recovery time of fast MPLS restoration. ~~Also, note that failures within a layer can be guarded against by techniques either in that layer or at a higher layer, but not in reverse. Thus, the optical layer cannot guard against failures in the IP layer such as router system failures, line card failures.~~Also, by design, the operations and mechanisms within a given layer are usually somewhat opaque to other layers, thus making its less easy for its failures to be protected by others.

5.6 Evolution Toward IP Over Optical

As more pressing requirements for survivability and horizontal hierarchy for edge-to-edge signaling are met with technical proposals, it is believed that the benefits of merging (in some manner) the control planes of multiple layers will be outlined. When these benefits are self-evident, it would then seem to be the right time to review if vertical hierarchy mechanisms are needed, and what the requirements might be. (For example, a future requirement might be to provide a better match between the restoration requirements of IP networks with the restoration capability of optical transport. One such proposal is described in [11].)

6. Hierarchy Requirements

Efforts in the area of network hierarchy should focus on mechanisms that would allow more scalable edge-to-edge signaling, or signaling across networks with existing network hierarchy (such as multi-area OSPF). This would appear to be a more immediate need than mechanisms that might be needed to interconnect networks at different layers. (An issue here may be that at optical data rates, OSPF updates may not be quick enough to recover from the tremendous loss in data. This may only work if each possible optical channel has its own logical IP address associated with it. This could cause an explosion in routing tables.)

6.1 Historical Context

One reason for horizontal hierarchy is functionality (e.g., metro versus backbone). Geographic “islands” reduce the need for interoperability and make administration and operations less complex. Using a simpler, more interoperable, survivability scheme at metro/backbone boundaries is natural for many provider network architectures. In transmission networks, creating geographic islands of different vendor equipment has been done for a long time because multi-vendor interoperability has been difficult to achieve. Traditionally, providers have to coordinate the equipment on either end of a "connection," and making this interoperable reduces complexity. A provider should be able to concatenate survivability mechanisms in order to provide a "protected link" to the next higher level. Think of SDH/SONET rings connecting to TDM DXCs with 1+1

line-layer protection between the ADM and the DXC port. The TDM connection, e.g., a DS3 is protected, but usually all equipment on each SDH/SONET ring is from a single vendor. The DXC cross connections are controlled by the provider and the ports are physically protected resulting in a highly available design. Thus, concatenation of survivability approaches can be used to cascade across horizontal hierarchy. While not perfect, it is workable in the near- to mid-term until multi-vendor interoperability is achieved.

While the problems associated with multi-vendor interoperability may necessitate horizontal hierarchy as a practical matter (at least this has been the case in TDM networks), there may be no technical reason for it. Members of the team with more experience on IP networks felt there should be no need for this in core networks, or even most access networks.

Some of the largest service provider networks currently run a single area/level IGP. Some service providers, as well as many large enterprise networks, run multi-area OSPF to gain increases in scalability. Often, this was from an original design, so it is difficult to say if the network truly required the hierarchy to reach its current size.

Some proposals on improved mechanisms to address network hierarchy have been suggested [12, 13, 14, 15, 16]. This document aims to provide the concrete requirements so that these and other proposals can first aim to meet some limited objectives.

6.2 Applications for Horizontal Hierarchy

A primary driver for intra-domain horizontal hierarchy is signaling scalability in the context of edge-to-edge VPNs, potentially across traffic-engineered data networks. There are a number of different approaches to VPNs and they are currently being addressed by different emerging protocols: RFC 2547bis BGP/MPLS VPNs, provider-provisioned VPNs based upon MPLS tunnels (e.g., virtual routers), Pseudo Wire Edge-to-edge Emulation (PWE3), etc. These may or not need explicit signaling from edge to edge, but it is a common perception that in order to meet SLAs, some form of edge-to-edge signaling is required.

For signaling scalability, there are probably two types of network scenarios to consider:

- Large SP networks with flat routing domains where edge-to-edge (MPLS) signaling as implemented today would probably not scale.
- Networks which would like to signal edge-to-edge, and might even scale in a limited application. However, they are hierarchically routed (e.g. OSPF areas) and current implementations, and potentially standards prevent signaling across areas. This requires the development of signaling standards that support

dynamic establishment and potentially restoration of LSPs across a 2-level IGP hierarchy.

Scalability is concerned with the O(N^2) properties of edge-to-edge signaling. For a large network, maintaining a "connection" between every edge is simply not scalable. Even if establishing and maintaining connections is feasible, there might be an impact on core survivability mechanisms which would cause restoration times to grow with N^2, which would be undesirable. While some value of N may be inevitable, approaches to reduce N (e.g. to pull in from the edge to aggregation points) might be of value.

For routing scalability, especially in data applications, a major concern is the amount of processing/state that is required in the variety of network elements. If some nodes might not be able to communicate and process the state of every other node, it might be preferable to limit the information by conveying abstracted topologies across areas as in multi-area OSPF. There is one way of thought that says that the amount of information contained by a horizontal barrier should be significant, and that impacts this might have on optimality in route selection and ability to provide global survivability are accepted tradeoffs.

6.3 Horizontal Hierarchy Requirements

Mechanisms are required to allow for edge-to-edge signaling of connections through a network. The types of network scenarios include large networks with a large number of edge devices and flat interior routing, as well as medium to large networks which currently have hierarchical interior routing such as multi-area OSPF or multi-level IS-IS. The primary context of this is edge-to-edge signaling which is thought to be required to assure the SLAs for the layer 2 and layer 3 VPNs that are being carried across the network. Another possible context would be edge-to-edge signaling in TDM SDH/SONET networks with IP control, where metro and core networks again might either be in a flat or hierarchical interior routing domain.

To support scalable edge-to-edge signaling in the above network scenarios, multi-area TE must be accommodated within the framework of existing horizontal hierarchies. Proposals for multi-area TE and related mechanisms such as route server, route query, crankback, and TE feedback are discussed in [12, 13, 14, 15, 16]. Corresponding protocol extensions need to be developed.

7. Survivability and Hierarchy

When horizontal hierarchy exist in a network layer, a question arises as to how survivability can be provided along a connection which crosses hierarchical boundaries.

In designing protocols to meet the requirements of hierarchy, an approach to consider is that boundaries are either clean, or are of

minimal value. However, the concept of network elements that participate on both sides of a boundary might be a consideration (e.g. OSPF ABRs). That would allow for devices on either side to take an intra-area approach within their region of knowledge, and for the ABR to do this in both areas, and splice the two protected connections together at a common point (granted it is a common point of failure now). If the limitations of this approach start to appear in operational settings, then perhaps it would be time to start thinking about route-servers and signaling propagated directives. However, one initial approach might be to signal through a common border router, and to consider the service as protected as it consist of a concatenated set of connections which are each protected within their area. Another approach might be to have a least common denominator mechanism at the boundary, e.g., 1+1 port protection. There should also be some standardized means for a survivability scheme on one side of such a boundary to communicate with the scheme on the other side regarding the success or failure of the service restoration action. For example, if a part of a "connection" is down on one side of such a boundary, there is no need for the other side to recover from failures.

In summary, at this time, approaches that allow concatenation of survivability schemes across hierarchical boundaries should provide sufficient.

8. Security Considerations

~~Security is not considered in this initial version.~~There may be the possibility of a hostile agent to create different issues with the signaling schema to ensure survivability.

- Access to a multilayer device may allow for false signaling of a working entity(s) failure.
- Protection entity information could be altered.
- Prioritization values for recovery/restoration could be altered.

9. References

1 Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

2 Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

3 V. Sharma, B. Crane, S. Makam, K. Owens, C. Huang, F. Hellstrand, J. Weil, L. Andersson, B. Jamoussi, B. Cain, S. Civanlar, and A. Chiu, "Framework for MPLS-based Recovery," Internet-Draft, Work in Progress, July 2001.

4 S. Dharanikota, R. Jain, D. Papadimitriou, R. Hartani, G. Bernstein, V. Sharma, C. Brownmiller, Y. Xue, and J. Strand,

"Inter-domain routing with Shared Risk Groups," Internet-Draft, Work in Progress, July 2001.

5 D.O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and Principles of Internet Traffic Engineering," Internet-Draft, Work in Progress, August 2001.

6 D.H. Gan, P. Pan, A. Ayyangar, and K. Kompella, "A Method for MPLS LSP Fast-Reroute Using RSVP Detours," Internet-Draft, Work in Progress, April 2001.

7 A. Atlas, C. Villamizar, and C. Litvanyi, "MPLS RSVP-TE Interoperability for Local Protection/Fast Reroute," Internet-Draft, Work in Progress, July 2001.

8 G. Li, C. Kalmanek, J. Yates, G. Bernstein, F. Liaw, and V. Sharma, RSVP-TE Extensions For Shared-Mesh Restoration in Transport Networks," Internet-Draft, Work in Progress, July 2001.

9 N. Harrison, P. Willis, S. Davari, E. Cuevas, B. Mack-Crane, E. Franze, H. Ohta, T. So, S. Goldfless, and F. Chen, "Requirements for OAM in MPLS Networks," Internet-Draft, Work in Progress, May 2001.

10 D. Allan and M. Azad, "A Framework for MPLS User Plane OAM," Internet-Draft, Work in Progress, July 2001.

11 A. Chiu and J. Strand, "Joint IP/Optical Layer Restoration after a Router Failure," Proc. OFC'2001, Anaheim, CA, March 2001.

12 K. Kompella and Y. Rekhter, "Multi-area MPLS Traffic Engineering," Internet-Draft, Work in Progress, March 2001.

13 G. Ash, et al, "Requirements for Multi-Area TE," Internet-Draft, Work in Progress, September 2001.

14 A. Iwata, N. Fujita, G.R. Ash, and A. Farrel, "Crankback Routing Extensions for MPLS Signaling," Internet-Draft, Work in Progress, July 2001.

15 C-Y Lee, A Celer, N Gammage, S Ghanti, G. Ash, "Distributed Route Exchangers," Internet-Draft, Work in Progress, March 2001.

16 C-Y Lee and S Ghanti, "Path Request and Path Reply Message," Internet-Draft, Work in Progress, July 2001.

10. Acknowledgments

A lot of the direction taken in this document, and by the team in its initial effort, was steered by the insightful questions provided

by Bala Rajagoplan, Greg Bernstein, Yangguang Xu, and Avri Doria. The set of questions is attached as Appendix A in this document.

After the release of the first draft, a number of comments were received. Thanks to the inputs from Jerry Ash, Sudheer Dharanikota, Dan Koller, Lyndon Ong, Steve Plote, and Yong Xue.

11. Author's Addresses

Wai Sum Lai
AT&T
200 Laurel Avenue
Middletown, NJ 07748, USA
Tel: +1 732-420-3712
wlai@att.com

Dave McDysan
WorldCom
22001 Loudoun County Pkwy
Ashburn, VA 20147, USA
dave.mcdysan@wcom.com

Jim Boyle
Protocol Driven Networks
Tel: +1 919-852-5160
jboyle@pdnets.com

Malin Carlzon
malin@sunet.se

Rob Coltun
rcoltun@redback.com

Tim Griffin
AT&T
180 Park Avenue
Florham Park, NJ 07932, USA
Tel: +1 973-360-7238
griffin@research.att.com

Ed Kern
Cogent Communications
ejk@tech.org

Tom Reddington
Lucent Technologies
67 Whippany Rd
Whippany, NJ 07981, USA
Tel: +1 973-386-7291
treddington@bell-labs.com

Appendix A: Questions used to help develop requirements

A. Definitions

1. In determining the specific requirements, the design team should precisely define the concepts "survivability", "restoration", "protection", "protection switching", "recovery", "re-routing" etc. and their relations. This would enable the requirements doc to describe precisely which of these will be addressed.
In the following, the term "restoration" is used to indicate the broad set of policies and mechanisms used to ensure survivability.

B. Network types and protection modes

1. What is the scope of the requirements with regard to the types of networks covered? Specifically, are the following in scope:

Restoration of connections in mesh optical networks (opaque or transparent)
Restoration of connections in hybrid mesh-ring networks
Restoration of LSPs in MPLS networks (composed of LSRs overlaid on a transport network, e.g., optical)
Any other types of networks?
Is commonality of approach, or optimization of approach more important?

2. What are the requirements with regard to the protection modes to be supported in each network type covered? (Examples of protection modes include 1+1, M:N, shared mesh, UPSR, BLSR, newly defined modes such as P-cycles, etc.)

3. What are the requirements on local span (i.e., link by link) protection and end-to-end protection, and the interaction between them? E.g.: what should be the granularity of connections for each type (single connection, bundle of connections, etc).

C. Hierarchy

1. Vertical (between two network layers):
What are the requirements for the interaction between restoration procedures across two network layers, when these features are offered in both layers? (Example, MPLS network realized over pt-to-pt optical connections.) Under such a case,

(a) Are there any criteria to choose which layer should provide protection?

(b) If both layers provide survivability features, what are the requirements to coordinate these mechanisms?

(c) How is lack of current functionality of cross-layer cooridnation currently hampering operations?

(d) Would the benefits be worth additional complexity associated with routing isolation (e.g. VPN, areas), security, address isolation and policy / authentication processes?

2. Horizontal (between two areas or administrative subdivisions within the same network layer):

(a) What are the criteria that trigger the creation of protocol or administrative boundaries pertaining to restoration? (e.g., scalability? multi-vendor interoperability? what are the practical issues?) multi-provider? Should multi-vendor necessitate hierarchical seperation?

When such boundaries are defined:

(b) What are the requirements on how protection/restoration is performed end-to-end across such boundaries?

(c) If different restoration mechanisms are implemented on two sides of a boundary, what are the requirements on their interaction?

What is the primary driver of horizontal hierarchy? (select one)
- functionality (e.g. metro -v- backbone)
- routing scalability
- signaling scalability
- current network architecture, trying to layer on TE ontop of already hiearchical network architecture
- routing and signalling

For signalling scalability, is it
- managability
- processing/state of network
- edge-to-edge N^2 type issue

For routing scalability, is it
- processing/state of network
- are you flat and want to go hierarchical
- or already hierarchical?
- data or TDM application?

D. Policy

1. What are the requirements for policy support during protection/restoration,
e.g., restoration priority, preemption, etc.

E. Signaling Mechanisms

1. What are the requirements on the signaling transport mechanism (e.g., in-band over sonet/sdh overhead bytes, out-of-band over an IP network, etc.) used to communicate restoration protocol
messages between network elements. What are the bandwidth and other requirements on the signaling channels?

2. What are the requirements on fault detection/localization mechanisms (which is the prelude to performing restoration procedures) in the case of opaque and transparent optical networks? What are the requirements in the case of MPLS restoration?

3. What are the requirements on signaling protocols to be used in restoration procedures (e.g., high priority processing, security, etc).

4. Are there any requirements on the operation of restoration protocols?

F. Quantitative

1. What are the quantitative requirements (e.g., latency) for completing restoration under different protection modes (for both local and end-to-end protection)?

G. Management

1. What information should be measured/maintained by the control plane at each network element pertaining to restoration events?

2. What are the requirements for the correlation between control plane and data plane failures from the restoration point of view?

Full Copyright Statement

MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.