Network Working Group
Request for Comments: 3386
Category: Informational

W. Lai, Ed.
AT&T
D. McDysan, Ed.
WorldCom

November 2002

Network Hierarchy and Multilayer Survivability

Status of this Memo

Copyright Notice

Abstract

This document presents a proposal of the near-term and practical requirements for network survivability and hierarchy in current service provider environments.

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 BCP 14, RFC 2119 [2].

Table of Contents

1. Introduction

This document is the result of the Network Hierarchy and Survivability Techniques Design Team established within the Traffic Engineering Working Group. This team collected and documented current and near term requirements for survivability and hierarchy in service provider environments. For clarity, an expanded set of definitions is included. The team determined that there appears to be a need to define a small set of 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. They operate primarily at a single network layer. For hierarchy, there did not appear to be a driving near-term need for work on "vertical hierarchy," defined as communication between network layers such as Time Division Multiplexed (TDM)/optical and Multi-Protocol Label Switching (MPLS). In particular, instead of direct exchange of signaling and routing between vertical layers, some looser form of coordination and communication, such as the specification of hold-off timers, is a nearer term need. For "horizontal hierarchy" in data networks, there are several pressing needs. The requirement is to be able to set up many Label Switched Paths (LSPs) in a service provider network with hierarchical Interior Gateway Protocol (IGP). This is necessary to support layer 2 and layer 3 Virtual Private Network (VPN) services that require edge-to-edge signaling across a core network.

This document presents a proposal of the near-term and practical 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.

Here is a summary of the findings:

A. Survivability Requirements

- o need to define a small set of interoperable survivability approaches in packet and non-packet networks
- o suggested survivability mechanisms include
  - 1:1 path protection with pre-established backup capacity (non-shared)
  - 1:1 path protection with pre-planned backup capacity (shared)

  - local restoration with repairs in proximity to the network fault
  - path restoration through source-based rerouting
- timing bounds for service restoration to support voice call cutoff (140 msec to 2 sec), protocol timer requirements in premium data services, and mission critical applications
- use of restoration priority for service differentiation

B. Hierarchy Requirements

B.1. Horizontally Oriented Hierarchy (Intra-Domain)

- ability to set up many LSPs in a service provider network with hierarchical IGP, for the support of layer 2 and layer 3 VPN services
- requirements for multi-area traffic engineering need to be developed to provide guidance for any necessary protocol extensions

B.2. Vertically Oriented Hierarchy

The following functionality for survivability is common on most routing equipment today.

- near-term need is some loose form of coordination and communication based on the use of nested hold-off timers, instead of direct exchange of signaling and routing between vertical layers
- means for an upper layer to immediately begin recovery actions in the event that a lower layer is not configured to perform recovery

C. Survivability Requirements in Horizontal Hierarchy

- protection of end-to-end connection is based on a concatenated set of connections, each protected within their area
- mechanisms for connection routing may include (1) a network element that participates on both sides of a boundary (e.g., OSPF ABR) - note that this is a common point of failure; (2) a route server
- need for inter-area signaling of survivability information (1) to enable a "least common denominator" survivability mechanism at the boundary; (2) to convey the success or failure of the service restoration action; e.g., if a part of a "connection" is down on one side of a boundary, there is no need for the other side to recover from failures

2. Terminology and Concepts

2.1 Hierarchy

Hierarchy is a technique used to build scalable complex systems. It is based on an abstraction, at each level, of what is most significant from the details and internal structures of the levels further away. This approach makes use of a general property of all hierarchical systems composed of related subsystems that interactions between subsystems decrease as the level of communication between subsystems decreases.

Network hierarchy is an abstraction of part of a network's topology, routing and signaling mechanisms. 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, 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) Vertically oriented: between two network technology layers.
(2) Horizontally oriented: between two areas or administrative subdivisions within the same network technology layer.

2.1.1 Vertical Hierarchy

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

In the vertical hierarchy, the total network functions are partitioned into a series of functional or technological layers with clear logical, and maybe even physical, separation between adjacent layers. Survivability mechanisms either currently exist or are being developed at multiple layers in networks [3]. The optical layer is now becoming capable of providing dynamic ring and mesh restoration functionality, in addition to traditional 1+1 or 1:1 protection. The Synchronous Digital Hierarchy (SDH)/Synchronous Optical NETwork (SONET) layer provides survivability capability with automatic protection switching (APS), as well as self-healing ring and mesh restoration architectures. Similar functionality has been defined in the Asynchronous Transfer Mode (ATM) Layer, with work ongoing to also provide such functionality using MPLS [4]. At the IP layer,

rerouting is used to restore service continuity following link and node outages. Rerouting at the IP layer, however, occurs after a period of routing convergence, which may require a few seconds to several minutes to complete [5].

2.1.2 Horizontal Hierarchy

Horizontal hierarchy is the abstraction that allows a network at one technology layer, for instance a packet network, to scale. Examples of horizontal hierarchy include BGP confederations, separate Autonomous Systems, and multi-area OSPF.

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.

2.2 Survivability Terminology

In alphabetical order, the following terms are defined in this section:

backup entity, same as protection entity (section 2.2.2)
extra traffic (section 2.2.2)
non-revertive mode (section 2.2.2)
normalization (section 2.2.2)
preemptable traffic, same as extra traffic (section 2.2.2)
preemption priority (section 2.2.4)
protection (section 2.2.3)
protection entity (section 2.2.2)
protection switching (section 2.2.3)
protection switch time (section 2.2.4)
recovery (section 2.2.2)
recovery by rerouting, same as restoration (section 2.2.3)
recovery entity, same as protection entity (section 2.2.2)
restoration (section 2.2.3)
restoration priority (section 2.2.4)
restoration time (section 2.2.4)
revertive mode (section 2.2.2)
shared risk group (SRG) (section 2.2.2)
survivability (section 2.2.1)
working entity (section 2.2.2)

2.2.1 Survivability

Survivability is the capability of a network to maintain service continuity in the presence of faults within the network [6]. Survivability mechanisms 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.

2.2.2 Generic Operations

This document does not discuss the sequence of events of how network failures are monitored, detected, and mitigated. For more detail of this aspect, see [4]. Also, the repair process following a failure is out of the scope here.

A working entity is the entity that is used to carry traffic in normal operation mode. Depending upon the context, an entity can be a channel or a transmission link in the physical layer, an Label Switched Path (LSP) in MPLS, or a logical bundle of one or more LSPs.

A protection entity, also called backup entity or recovery entity, is the entity that is used to carry protected traffic in recovery operation mode, i.e., when the working entity is in error or has failed.

Extra traffic, also referred to as preemptable 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, the extra traffic is preempted.

A 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. The concept of SRG can be extended to represent a "risk domain" and its associated capabilities and summarization for traffic engineering purposes. See [7] for further discussion.

Normalization is the sequence of events and actions taken by a network that returns the network to the preferred state upon completing repair of a failure. This could include the switching or rerouting of affected traffic to the original repaired working entities or new routes. Revertive mode refers to the case where traffic is automatically returned to a repaired working entity (also called switch back).

Recovery is the sequence of events and 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 using survivability mechanisms to maintain service continuity. In protection, reconfiguration involves switching the affected traffic from a working entity to a protection entity. In restoration, reconfiguration involves path selection and rerouting for the affected traffic.

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.

Non-revertive mode is the case where there is no preferred path or it may be desirable to minimize further disruption of the service brought on by a revertive switching operation. A switch-back to the original working path is not desired or not possible since the original path may no longer exist after the occurrence of a fault on that path.

2.2.3 Survivability Techniques

Protection, also called protection switching, is a survivability technique based on predetermined failure recovery: as the working entity is established, a protection entity is also established. 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 other end (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.

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. When the working entity fails, the protected traffic is switched to the protection entity. The two ends of the protected domain must signal detection of the fault and initiate the switchover.

In the 1:n protection architecture, a dedicated protection entity is shared by n working entities. In this case, not all of the affected traffic may be protected.

The m:n architecture is a generalization of the 1:n architecture. Typically m <= n, where m dedicated protection entities are shared by n working entities.

Restoration, also referred to as recovery by rerouting [4], is a survivability technique that establishes new paths or path segments on demand, for restoring affected traffic after the occurrence of a fault. The resources in these alternate paths are the currently unassigned (unreserved) resources in the same layer. Preemption of extra traffic may also be used if spare resources are not available to carry the higher-priority protected traffic. As initiated by detection of a fault on the working path, the selection of a recovery path may be based on preplanned configurations, network routing policies, or current network status such as network topology and fault information. Signaling is used for establishing the new paths to bypass the fault. Thus, restoration involves a path selection process followed by rerouting of the affected traffic from the working entity to the recovery entity.

2.2.4 Survivability Performance

Protection switch time is the time interval from the occurrence of a network fault until the completion of the protection-switching operations. It includes the detection time necessary to initiate the protection switch, any hold-off time to allow for the interworking of protection schemes, and the switch completion time.

Restoration time is the time interval from the occurrence of a network fault to the instant when the affected traffic is either completely restored, or until spare resources are exhausted, and/or no more extra traffic exists that can be preempted to make room.

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 restoring 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.

Preemption priority is a method of determining which traffic can be disconnected in the event that not all traffic with a higher restoration priority is restored after the occurrence of a failure.

2.3 Survivability Mechanisms: Comparison

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.

A 1+1 protection architecture is rather expensive since resource duplication is required for the working and protection entities. It is generally used for specific services that need a very high availability.

A 1:1 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 low-priority extra traffic in normal operation, if traffic preemption is allowed. Packet networks can pre-establish a protection path for later use with pre-planned but not pre-reserved capacity. That is, if no packets are sent onto a protection path,

then no bandwidth is consumed. This is not the case in transmission networks like optical or TDM where path establishment and resource reservation cannot be decoupled.

In the 1:n protection architecture, 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 could be 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 traffic in normal operation.

While the m:n architecture can improve system availability with small cost increases, it has rarely been implemented or standardized.

When compared with protection mechanisms, restoration mechanisms are generally more frugal as no resources are committed until after the fault occurs and the location of the fault is known. However, restoration mechanisms are inherently slower, since more must be done following the detection of a fault. Also, the time it takes for the dynamic selection and establishment of alternate paths may vary, depending on the amount of traffic and connections to be restored, and is influenced by the network topology, technology employed, and the type and severity of the fault. As a result, restoration time tends to be more variable than the protection switch time needed with pre-selected protection entities. Hence, in using restoration mechanisms, it is essential to use restoration priority to ensure that service objectives are met cost-effectively.

Once the network routing algorithms have converged after a fault, it may be preferable in some cases, to reoptimize the network by performing a reroute based on the current state of the network and network policies.

3. Survivability

3.1 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 that 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

perform fault notification and to control the operation of survivability mechanisms among different layers. This may require the development of corresponding OAM functionality. However, such issues and those related to OAM are currently outside the scope of this document. (For proposed MPLS OAM requirements, see [8, 9]).

The initial scope is to address only "backhoe failures" in the inter-office connections of a service provider network. A link connection in the router layer is typically comprised of multiple spans in the lower layers. Therefore, the types of network failures that cause a recovery to be performed include link/span failures. However, linecard and node failures may not need to be treated any differently than their respective link/span failures, as a router failure may be represented as a set of simultaneous link failures.

Depending on the actual network configuration, drop-side interface (e.g., between a customer and an access router, or between a router and an optical cross-connect) may be considered either inter-domain or inter-layer. Another inter-domain scenario is the use of intra-office links for interconnecting a metro network and a core network, with both networks being administered by the same service provider. Failures at such interfaces may be similarly protected by the mechanisms of this section.

Other more complex failure mechanisms such as systematic control-plane failure, configuration error, or breach of security are not within the scope of the survivability mechanisms discussed in this document. Network impairment such as congestion that results in lower throughput are also not covered.

3.2 Required initial set of survivability mechanisms

3.2.1 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 requires 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 should be capable of carrying preemptable extra traffic in non-packet networks. When protection capacity is called into service during recovery, 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.

3.2.2 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". Operators need the ability to provision different amounts of protection capacity according to expected failure modes and service level agreements. Thus, an operator may wish to provision sufficient restoration capacity to handle a single failure affecting all connections in an SRG, or may wish to provision less or more restoration capacity. Mechanisms should be provided to allow restoration capacity on each link to be shared by SRG-disjoint failures. 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, e.g., see the proposals in [10, 11]. If capacity is planned but not allocated, some form of signaling could be required before traffic may be sent on protection connections, especially in TDM networks.

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.

3.2.3 Local Restoration

Due to the time impact of signal propagation, dynamic recovery of an entire path may not meet the service requirements of some networks. The solution to this is to restore connectivity of the link or span in immediate proximity to the fault, e.g., see the proposals in [12, 13]. At a minimum, this approach should be able to protect against connectivity-type SRGs, though protecting against node-based SRGs might be worthwhile. Also, this approach is applicable to support restoration on the inter-domain and inter-layer interconnection scenarios using intra-office links as described in the Scope Section.

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 the definition of an object in signaling.

Since local restoration may be suboptimal, a means for head end systems to later perform path-level re-grooming must be supported for this approach.

3.2.4 Path Restoration

In this approach, connections that are impacted by a fault are rerouted by the originating network element upon notification of connection failure. Such a source-based approach is efficient for network resources, but typically takes longer to accomplish restoration. It does not involve any new mechanisms. It merely is a mention of another common approach to protecting against faults in a network.

3.3 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 recovered before the network failure ends. The length of this interrupted period is dependent upon the application supported. Here are some typical applications and considerations that drive the requirements for an acceptable protection switch time or restoration time:

- Best-effort data: recovery of network connectivity by rerouting at the IP layer would be sufficient
- Premium data service: need to meet TCP timeout or application protocol timer requirements
- Voice: call cutoff is in the range of 140 msec to 2 sec (the time that a person waits after interruption of the speech path before hanging up or the time that a telephone switch will disconnect a call)
- Other real-time service (e.g., streaming, fax) where an interruption would cause the session to terminate
- Mission-critical applications that cannot tolerate even brief interruptions, for example, real-time financial transactions

3.4 Timing Bounds for Survivability Mechanisms

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 protection/restoration, and then attempt to select a few general points that 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 in which 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 different survivability 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 must be implemented to determine the order in which connections are restored (to minimize service restoration time as well as to gain access to available spare capacity on the best paths). 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 not all connections can 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 may not be needed in the network.

3.5 Coordination Among Layers

A common design goal for networks with multiple technological layers is to provide the desired level of service in the most cost-effective manner. Multilayer survivability may 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 recovery 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 recovery happens in a longer time period than higher-layer recovery, a hold-off timer is utilized to avoid contention between the different single-layer survivability schemes. In other words, multilayer interaction is addressed by having successively higher multiplexing levels operate at a protection/restoration time scale greater than the next lowest layer. This can impact the overall time to recover service. For example, if SDH/SONET protection switching is used, MPLS recovery timers must wait until SDH/SONET has had time to switch. Setting such timers involves a tradeoff between rapid recovery and creation of a race condition where multiple layers are responding to the same fault, potentially allocating resources in an inefficient manner.

In other configurations where the lower layer does not have a restoration 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 difference in network configuration means that implementations must allow for adjustment of hold-off timer values and/or a means for a lower layer to immediately indicate to a higher layer that a fault has occurred so that the higher layer can take restoration or protection actions.

Furthermore, faults at higher layers should not trigger restoration or protection actions at lower layers [3, 4].

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, by design, the operations and mechanisms within a given layer tend to be invisible to other layers.

3.6 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 whether 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 recovery requirements of IP networks with the recovery capability of optical transport. One such proposal is described in [14].

4. 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 appears to be a more urgent need than mechanisms that might be needed to interconnect networks at different layers.

4.1 Historical Context

One reason for horizontal hierarchy is functionality (e.g., metro versus backbone). Geographic "islands" or partitions 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 a 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 in the near to mid-term (at least this has been the case in TDM networks), there should not be a technical reason for it in the standards developed by the IETF for core networks, or even most access networks. Establishing interoperability of survivability mechanisms between multi-vendor equipment in core IP networks is urgently required to enable adoption of IP as a viable core transport technology and to facilitate the traffic engineering of future multi-service IP networks [3].

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 Open Shortest Path First (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 [15, 16, 17, 18, 19]. This document aims to provide the concrete requirements so that these and other proposals can first aim to meet some limited objectives.

4.2 Applications for Horizontal Hierarchy

A primary driver for intra-domain horizontal hierarchy is signaling capabilities in the context of edge-to-edge VPNs, potentially across traffic-engineered data networks. There are a number of different approaches to layer 2 and layer 3 VPNs and they are currently being addressed by different emerging protocols in the provider-provisioned VPNs (e.g., virtual routers) and Pseudo Wire Edge-to-Edge Emulation (PWE3) efforts based on either MPLS and/or IP tunnels. These may or may 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 may be required.

With a large number of edges (N), scalability is concerned with avoiding the O(N^2) properties of edge-to-edge signaling. However, the main issue here is not with the scalability of large amounts of signaling, such as in O(N^2) meshes with a "connection" between every edge-pair. This is because, even if establishing and maintaining connections is feasible in a large network, there might be an impact on core survivability mechanisms which would cause

protection/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.

Thus, most service providers feel that O(N^2) meshes are not necessary for VPNs, and that the number of tunnels to support VPNs would be within the scalability bounds of current protocols and implementations. That may be the case, as there is currently a lack of ability to signal MPLS tunnels from edge to edge across IGP hierarchy, such as OSPF areas. This may require the development of signaling standards that support dynamic establishment and potentially the restoration of LSPs across a 2-level IGP hierarchy.

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. There is one school 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.

4.3 Horizontal Hierarchy Requirements

Mechanisms are required to allow for edge-to-edge signaling of connections through a network. One network scenario includes medium to large networks that currently have hierarchical interior routing such as multi-area OSPF or multi-level Intermediate System to Intermediate System (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 be in a hierarchical interior routing domain.

To support edge-to-edge signaling in the above network scenarios within the framework of existing horizontal hierarchies, current traffic engineering (TE) methods [20, 6] may need to be extended. Requirements for multi-area TE need to be developed to provide guidance for any necessary protocol extensions.

5. Survivability and Hierarchy

When horizontal hierarchy exists in a network technology layer, a question arises as to how survivability can be provided along a connection that 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 consists 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 recovery 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 as described above that allow concatenation of survivability schemes across hierarchical boundaries seem sufficient.

6. Security Considerations

The set of SRGs that are defined for a network under a common administrative control and the corresponding assignment of these SRGs to nodes and links within the administrative control is sensitive information and needs to be protected. An SRG is an acknowledgement that nodes and links that belong to an SRG are susceptible to a common threat. An adversary with access to information contained in an SRG could use that information to design an attack, determine the scope of damage caused by the attack and, therefore, be used to maximize the effect of an attack.

The label used to refer to a particular SRG must allow for an encoding such that sensitive information such as physical location, function, purpose, customer, fault type, etc. is not readily discernable by unauthorized users.

SRG information that is propagated through the control and management plane should allow for an encryption mechanism. An example of an approach would be to use IPSEC [21] on all packets carrying SRG information.

7. 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] K. Owens, V. Sharma, and M. Oommen, "Network Survivability Considerations for Traffic Engineered IP Networks", Work in Progress.

[4] 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", Work in Progress.

[5] M. Thorup, "Fortifying OSPF/ISIS Against Link Failure", http://www.research.att.com/~mthorup/PAPERS/lf_ospf.ps

[6] Awduche, D., Chiu, A., Elwalid, A., Widjaja, I. and X. Xiao, "Overview and Principles of Internet Traffic Engineering", RFC 3272, May 2002.

[7] 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", Work in Progress.

[8] 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," Work in Progress.

[9] D. Allan and M. Azad, "A Framework for MPLS User Plane OAM," Work in Progress.

[10] S. Kini, M. Kodialam, T.V. Lakshman, S. Sengupta, and C. Villamizar, "Shared Backup Label Switched Path Restoration," Work in Progress.

[11] G. Li, C. Kalmanek, J. Yates, G. Bernstein, F. Liaw, and V. Sharma, "RSVP-TE Extensions For Shared-Mesh Restoration in Transport Networks", Work in Progress.

[12] P. Pan (Editor), D.H. Gan, G. Swallow, J. Vasseur, D. Cooper, A. Atlas, and M. Jork, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", Work in Progress.

[13] A. Atlas, C. Villamizar, and C. Litvanyi, "MPLS RSVP-TE Interoperability for Local Protection/Fast Reroute", Work in Progress.

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

[15] K. Kompella and Y. Rekhter, "Multi-area MPLS Traffic Engineering", Work in Progress.

[16] G. Ash, et. al., "Requirements for Multi-Area TE", Work in Progress.

[17] A. Iwata, N. Fujita, G.R. Ash, and A. Farrel, "Crankback Routing Extensions for MPLS Signaling", Work in Progress.

[18] C-Y Lee, A. Celer, N. Gammage, S. Ghanti, G. Ash, "Distributed Route Exchangers", Work in Progress.

[19] C-Y Lee and S. Ghanti, "Path Request and Path Reply Message", Work in Progress.

[20] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M. and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.

[21] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

8. 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, Chuck Kalmanek, Dan Koller, Lyndon Ong, Steve Plote, and Yong Xue.

9. Contributing Authors

Jim Boyle (PDNets), Rob Coltun (Movaz), Tim Griffin (AT&T), Ed Kern, Tom Reddington (Lucent) and Malin Carlzon.

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 coordination 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 separation?

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 on top of an already hierarchical network architecture
- routing and signalling

For signalling scalability, is it
- manageability
- 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 SDH/SONET 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?

Editors' Addresses

Wai Sum Lai
AT&T
200 Laurel Avenue
Middletown, NJ 07748, USA

Phone: +1 732-420-3712
EMail: wlai@att.com

Dave McDysan
WorldCom
22001 Loudoun County Pkwy
Ashburn, VA 20147, USA

EMail: dave.mcdysan@wcom.com

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.