



**MSF Technical Report
MSF-TR-ARCH-008-FINAL**

Network Engineering to Support the Bandwidth Manager Architecture

Author: Cisco Systems
John Evans
joevans@cisco.com

www.msforum.org

May 2006

Abstract:

Network level behaviours can impact the determinism of call admission control decisions for a particular bandwidth management deployment. This white paper augments the bandwidth management technical report with a discussion of the impact of different network routing / forwarding models when used in conjunction with the bandwidth manager.

Disclaimer:

The following is a technical report of the MultiService Forum. The information in this publication is believed to be accurate as of its publication date. Such information is subject to change without notice and the MultiService Forum is not responsible for any errors or omissions. The MultiService Forum does not assume any responsibility to update or correct any information in this publication. Notwithstanding anything contained herein to the contrary, neither the MultiService Forum nor the publisher make any representation or warranty, expressed or implied, concerning the completeness, accuracy, or applicability of any information contained in this publication. No liability of any kind whether based on theories of tort, contract, warranty, strict liability or otherwise, shall be assumed or incurred by the MultiService Forum, its member companies, or the publisher as a result of reliance upon or use by any party of any information contained in this publication. All liability for any implied or express warranty of merchantability or fitness for a particular purpose, or any other warranty, is hereby disclaimed.

For additional information contact:

MultiService Forum
39355 California Street, Suite 307, Fremont, CA 94538
+1 (510) 608-5922
+1 (510) 608-5917 (fax)
info@msforum.org
<http://www.msforum.org>
Copyright © MultiService Forum 2006

1 Introduction

The bandwidth management technical report [1] outlines the issues surrounding bandwidth management for PSTN grade voice and multi-media services over packet networks within the context of the MSF QoS solution. A key requirement for migrating PSTN voice services to IP / MPLS networks is to provide the same levels of determinism for voice services as are available with the PSTN today.

To provide guaranteed support for services such as PSTN grade voice, one approach is to provision sufficient class bandwidth throughout the network to be able to assure that the total voice load can be serviced. However, consideration needs to be given to the limitations of guaranteed bandwidth provisioning, especially during network failure conditions:

- *Network working case conditions.* If sufficient bandwidth provisioning to cope with the peak call load can be assured only in network working case (i.e. normal operation with no failures) conditions then in all but the most trivial of topologies (i.e. those that are non-resilient) call admission control (CAC) may be required to cover network failure case conditions. In these cases CAC provides the capability to reject new or rerouted service requests so that those already granted admission continue to maintain their committed service; without CAC, congestion may occur which can degrade all calls.

If there were insufficient bandwidth to support the peak call load in normal working case conditions, then CAC would be required to cover both working and failure cases.

- *Single network element failure conditions.* Network planning and provisioning methods may be applied which take single element failures into account, ensuring that sufficient bandwidth is provisioned when allowing for all single network element failure conditions. In cases such as this, admission control may not be required; if connectivity verification shows that connectivity exists, then sufficient bandwidth must exist also.

Even where planning and provisioning takes single element failures into account, in some topologies there can be unplanned (multiple) failure cases where there is insufficient bandwidth to support the service load even though IP connectivity exists. In these cases, CAC may be required.

- *Multiple network element failure conditions.* If sufficient bandwidth can be provisioned to allow for multiple network element failures then admission control is not required. However, in many meshed topologies, ensuring that sufficient bandwidth exists in multiple network element failure case may not be a viable approach.

In all cases, it is a provider choice as to whether the cost and complexity of deploying admission control mechanisms is justified by the prevalence of events which may lead to service degradation due to congestion, and duration and impact of the potentially degraded service.

The MSF bandwidth manager functionality is an example of an off-path¹ or path-decoupled topology-aware admission control system, which could be used to provide admission control in the cases described above. To provide deterministic service and ensure that transient congestion cannot occur following network failure cases, it is critical that traffic cannot be rerouted before a new admission control decision is made based upon an up-to-date network view. This capability is implicit in the PSTN, but not natively in IP networks. Section 2 of this contribution examines the network routing / forwarding models described in [1] and considers their ability to provide such deterministic admission control capabilities that are available with the PSTN, within the context an MSF bandwidth manager deployment. Section 3 includes other specific network level comments with respect to [1].

¹ With off-path approaches, the messages used for QoS signaling are routed through nodes that are not assumed to be on the data (bearer) path.

2 Network Routing / Forwarding Models

2.1 IGP (no MPLS TE)

Although this model is not explicitly defined in [1] it is considered because it is the most basic IP network behaviour and some of the models that are defined in [1] exhibit the same characteristics.

With a conventional link state Interior Gateway Protocol (IGP) deployment, such as OSPF [2] or ISIS [3] – where MPLS Traffic Engineering (TE) [4, 5, 6] is not used – each router receives link state database information which is flooded throughout the network, and using a local copy of this database, each router will autonomously calculate its own routing table. If the database is consistent across the network, then the routing will also be consistent.

In such a deployment, if the bandwidth manager also has a copy of the same link state database (e.g. from passively participating in the IGP), in normal network working case conditions, the bandwidth manager should be able to process admission control decisions based upon the same view of the network routing topology. Hence, in working conditions, the bandwidth manager should be able to provide a deterministic CAC function.

Following network failures, the routers connecting to the failing network element will issue updated link state advertisements (LSAs, for OSPF) / link state PDUs (LSPs, for ISIS), which are then flooded throughout the network. Each router that receives the flooded update will perform a new Shortest Path First (SPF) calculation to recalculate their routing tables; each router will do this autonomously of the other routers and autonomously of the bandwidth manager. Consequently, traffic may be rerouted at the network level before a new admission control decision can be made by the bandwidth manager. Hence, transient congestion following network failures is possible when the bandwidth manager is used in conjunction with conventional IP routing; this transient congestion may impact both calls rerouted as a result of the failure and calls which had previously been successfully admitted onto the same path.

To illustrate this point consider the network topology shown in Figure 1 (for simplicity assume all links = 1Mbps) and the following sequence of events:

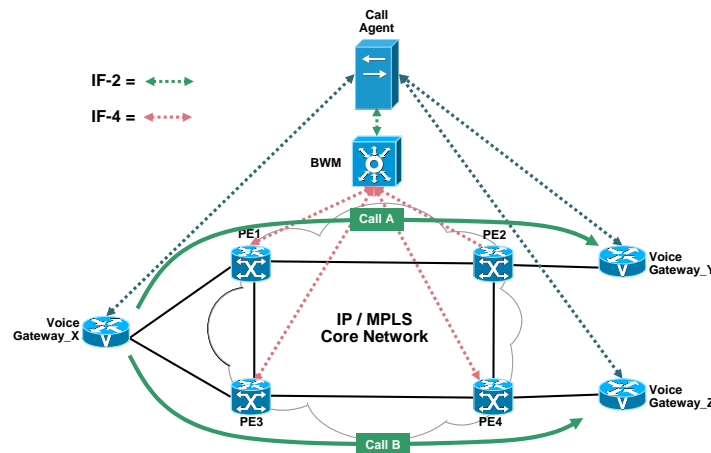


Figure 1. IGP (no MPLS TE) example: step #4

1. The call agent requests admission for a 1Mbps call (Call A) to be setup from X to Y (for simplicity only call legs from X to Y are considered in this example; not from Y to X)
2. The bandwidth manager verifies that sufficient bandwidth is available on the link(s) impacted by the call (PE1→PE2); it accepts the call and responds positively to the call agent
3. The call agent requests admission for another 1Mbps call (Call B) to be setup from X to Z

4. The bandwidth manager verifies that sufficient bandwidth is available on the links impacted by the call (PE3→PE4); it accepts the call and responds positively to the call agent
5. (t1) Link PE1⇔PE2 fails
6. (t2) Router PE1 recalculates its routing table autonomously, and as a result Call A is rerouted and now follows the path PE1 → PE3 → PE4 → PE2 and link PE3→PE4 is now congested, which impacts both calls, as shown in Figure 2.

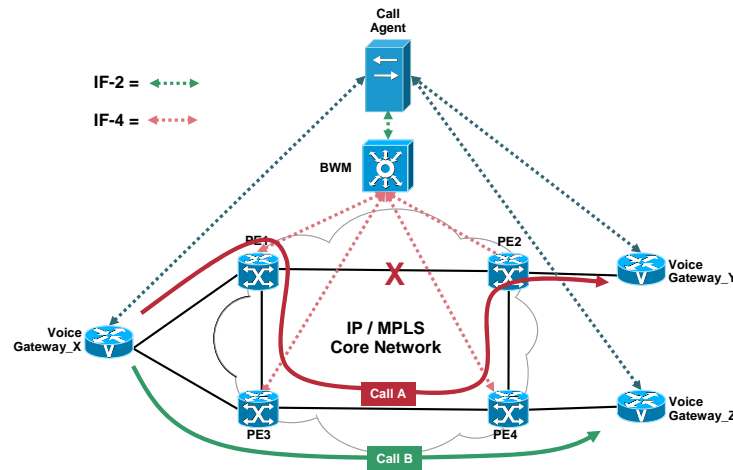


Figure 2. IGP (no MPLS TE) example: step #6

7. (t3) The bandwidth manager recalculates its routing table and determines that Call A is rerouted and now follows the path PE1 → PE3 → PE4 → PE2 and that link PE3→PE4 is congested.
8. (t4) The bandwidth manager may clear down Call A or Call B by signalling to the call agent or Call A and/or Call B may be cleared down by bearer level packet loss detection in the gateways.

The time difference between t1 and t2 is the IGP convergence time; during this time there is loss of connectivity and service disruption to those calls directly impacted by the failed network element. In well-designed networks, the IGP convergence time can be sub-second [7]. This time can be further reduced through the use of MPLS TE Fast Reroute (FRR) [8] (see section 2.4).

In the time difference between t2 and t4, traffic is being rerouted onto a path which as a result becomes congested; this causes loss of connectivity and service disruption both to the rerouted calls (Call A) and to those calls which had previously been successfully admitted onto that new path (Call B). With conventional IGP deployments, such transient congestion following network failures is possible, as each router will make an IGP rerouting decision autonomously of the bandwidth manager, i.e. traffic can be rerouted before the bandwidth manager can make a new admission control decision based upon an up-to-date network view. This is symptomatic of the fact that in this case there is no mechanism to keep the network rerouting behaviour in lock-step with the bandwidth manager's view of the network. The possibility of transient congestion network following failures can be removed by using MPLS TE with specified tunnel bandwidths, as per section 2.3.

In this scenario, loss of call connectivity and service disruption lasts from t1 to t4.

2.2 MPLS TE with zero bandwidth tunnels

A common initial step in TE deployments is to deploy edge-to-edge zero bandwidth tunnels, which is the model described in section 1.4.1.1 of [1].

With MPLS TE, tunnel paths can either be dynamically calculated online in a distributed fashion by the TE tunnel sources (known as head-ends) themselves (i.e. using a dynamic path option) or can be calculated

by a offline centralised function (i.e. path computation element, or tunnel server) which then specifies the explicit tunnel path a head-end should use for a particular tunnel (using an explicit path option). With either approach, a constraint-based shortest path first (CSFP) calculation is used to determine the path that a particular tunnel will take. This CSFP calculation is similar to a conventional IGP SPF calculation, but it also takes into account bandwidth and administrative constraints, to determine the shortest path to a destination, which also satisfies those constraints. Whether online or offline path calculation is used, the output is an explicit route object (ERO) which defines the hop-by-hop tunnel path and which is handed to RSVP in order to signal the tunnel label switched path (LSP).

If CSFP is used with zero bandwidth tunnels then, assuming no administrative constraints are applied, tunnels will implicitly follow the IGP shortest path, i.e. for zero bandwidth tunnels the CSFP path is effectively the same as the SPF path and the bandwidth constraint has no effect. If this is used in combination with dynamic tunnel path options (where the tunnel head-end routers determine the tunnel paths themselves) then this suffers the same issue as using no TE at all (section 2.1), which in this case is that following network failures each head-end router will make an autonomous rerouting decision before a new CAC decision can be made based upon an updated view of the network topology. To illustrate this point consider the network topology shown in Figure 3 (for simplicity assume all links = 1Mbps) and the following sequence of events:

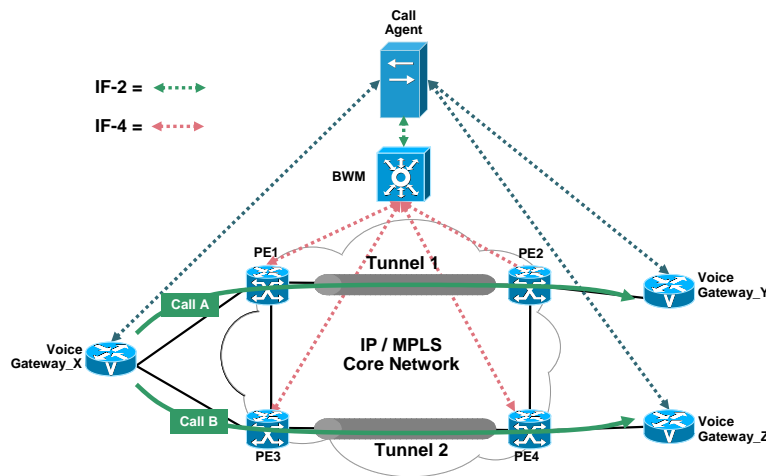


Figure 3. MPLS TE with zero bandwidth tunnels example: step #6

1. PE1 uses CSFP to calculate the path for a zero bandwidth tunnel (Tunnel 1) to PE2; as it is a zero bandwidth tunnel it follows the IGP shortest path, i.e. it traverses the directly connecting link PE1→PE2; PE1 uses RSVP signalling to setup Tunnel 1
2. PE3 uses CSFP to calculate the path for a zero bandwidth tunnel (Tunnel 2) to PE4, as it is a zero bandwidth tunnel it follows the IGP shortest path, i.e. it traverses the directly connecting link PE3→PE4; PE3 uses RSVP signalling to setup Tunnel 2
3. The call agent requests admission for a 1Mbps call (Call A) to be setup from X to Y (only the call legs from X are considered in this example; it is noted that TE tunnels are unidirectional entities and therefore additional tunnels would be required to support the call legs to X)
4. The bandwidth manager verifies that sufficient bandwidth is available on the links impacted by the call (PE1→PE2); it accepts the call (which uses Tunnel 1) and responds positively to the call agent²
5. The call agent requests admission for another 1Mbps call (Call B) to be setup from X to Z
6. The bandwidth manager verifies that sufficient bandwidth is available on the links impacted by the call (PE3→PE4); it accepts the call (which uses Tunnel 2) and responds positively to the call

² Note that although there is only a single call using the tunnel in this and subsequent examples, there may in practise be many calls using a single TE tunnel, with the only limit being imposed by the available tunnel bandwidth.

agent

7. (t1) Link PE1↔PE2 fails
8. (t2) Router PE1 uses CSPF to recalculate the path for Tunnel 1 which it re-signals; as the tunnel has zero bandwidth it follows the IGP shortest path with the result that both Tunnel 1 and Call A (which is carried on Tunnel 1) now follow the path PE1 → PE3 → PE4 → PE2 and link PE3→PE4 is now congested, which impacts both calls, as shown in Figure 4.

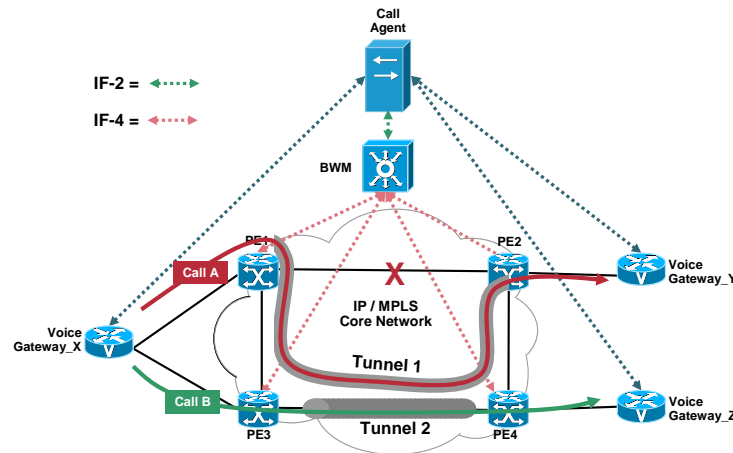


Figure 4. MPLS TE with zero bandwidth tunnels example: step #8

9. (t3) The bandwidth manager recalculates its routing table and determines that Call A is rerouted and now follows the path PE1 → PE3 → PE4 → PE2 and that link PE3→PE4 is congested.
10. (t4) The bandwidth manager may clear down Call A or Call B by signalling to the call agent or Call A and/or Call B may be cleared down by bearer level packet loss detection in the gateways.

In this case, the time difference between t1 and t2 is the time to re-signal the TE tunnel; during this time, there is loss of connectivity and service disruption to those calls directly impacted by the failed network element. This is typically in the order of a 1-2 seconds; this time can be reduced to in the order of 50ms through the use of MPLS TE FRR (see section 2.4).

In the time difference between t2 and t4, traffic is being rerouted onto a path which as a result becomes congested; this causes loss of connectivity and service disruption both to rerouted calls (Call A in this case) and to those calls which had previously been admitted onto that path (Call B). Hence, as with a plain IGP deployment (section 2.1) transient congestion following network failures is possible if zero bandwidth tunnels are used in conjunction with dynamic tunnel path options. This possibility of transient congestion can be removed by using MPLS TE with specified tunnel bandwidths, as per section 2.3.

In this scenario, as per the plain IGP (i.e. no TE) case (section 2.1) loss of call connectivity and service disruption lasts from t1 to t4.

It may seem that this model provides little benefit over the conventional IGP approach, as it provides neither the benefits of network level CAC nor bandwidth optimisation (the ability to make use of paths other than the IGP shortest path). This model is often used, however, as an initial step towards a full TE deployment where tunnel bandwidths are explicitly specified (as per section 2.3), in order to determine the core traffic matrix, and hence the resulting tunnel bandwidth settings which will later be used. This model may also be used to gain benefit from MPLS TE FRR as deploying “primary” TE tunnels is a prerequisite for FRR, however, if this is the main requirement a simpler approach may be to use one-hop zero bandwidth tunnels between all adjacent PE and P routers, rather than a mesh tunnels between PEs.

Where an offline tunnel server is used to determine the tunnel explicit paths, then the bandwidth specified for each tunnel at the head-end has less significance than in the context of head-end CSPF, as it is only used in the tunnel signalling, and not in the path computation. Hence, two sub cases can be considered:

- If the tunnel server calculates the tunnel explicit paths taking into account defined edge-to-edge bandwidth demands, then this is effectively equivalent to the case where MPLS TE is used with specified tunnel bandwidths as per section 2.3
- If the requested bandwidth demands are not taken into account when calculating the tunnel explicit paths, then effectively the tunnels only provide an explicit (static) routing capability across the core network. In this case, the bandwidth manager would need to understand the current state of the entire core network topology (rather than just the state of tunnels) when it processes a CAC decision, to be able to make deterministic CAC decisions across the core network. In addition, the bandwidth manager would need to make a new CAC decision based upon an updated view of the network topology before any tunnels are rerouted by the tunnel server.

2.3 MPLS TE with specified tunnel bandwidth

If CSFP is used with non-zero bandwidth tunnels either calculated by the tunnel head-ends or by an offline tunnel server, then implicitly TE bandwidth reservations are made at the network level. In this case, it is not possible for a tunnel to be rerouted without a new admission control decision being made. To illustrate this point consider the network topology shown in Figure 5 (for simplicity assume all links = 1Mbps) and the following sequence of events:

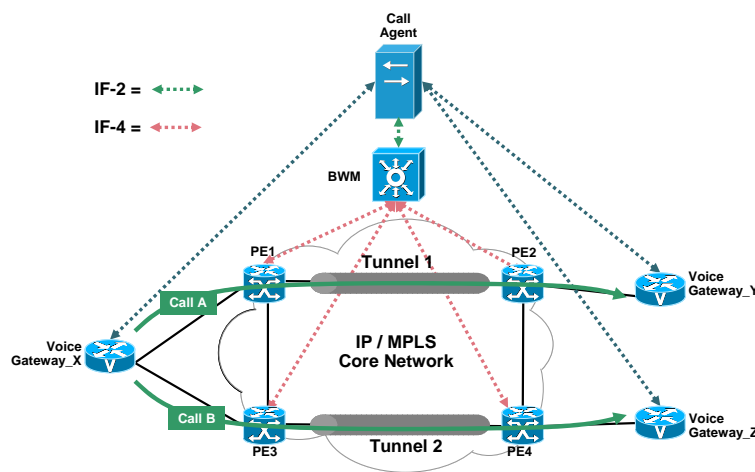


Figure 5. MPLS TE with specified tunnel bandwidth example: step #6

1. PE1 uses CSPF to calculate the path for a 1Mbps tunnel (Tunnel 1) to PE2, which traverses the directly connecting link PE1→PE2; PE1 uses RSVP signalling to setup Tunnel 1
2. PE3 uses CSPF to calculate the path for a 1Mbps tunnel (Tunnel 2) to PE4, which traverses the directly connecting link PE3→PE4; PE3 uses RSVP signalling to setup Tunnel 2
3. The call agent requests admission for a 1Mbps call (Call A) to be setup from X to Y (for simplicity only the call legs from X are considered in this example; it is noted that TE tunnels are unidirectional entities and therefore additional tunnels would be required to support the call legs to X)
4. The bandwidth manager verifies that sufficient bandwidth is available on the tunnel impacted by the call (Tunnel 1); it accepts the call and responds positively to the call agent
5. The call agent requests admission for a 1Mbps call (Call B) to be setup from X to Z
6. The bandwidth manager verifies that sufficient bandwidth is available on the tunnel impacted by the call (Tunnel 2); it accepts the call and responds positively to the call agent
7. (t1) Link PE1↔PE2 fails
8. Router PE1 uses CSPF to recalculate the path for Tunnel 1; as there are no paths available

which meet the bandwidth constraint, Tunnel 1 is not able to be re-established and Call A is impacted by the failure but congestion on Link PE3→PE4 is avoided (hence the service to Call B is preserved), and the bandwidth manager is informed.

9. The bandwidth manager may clear down Call A by signalling to the call agent or Call A may be cleared down by bearer level packet loss detection in the gateways.
10. (t2) If there were another path available with sufficient bandwidth for the rerouted tunnel, it would be re-established.

Where MPLS TE is used with non-zero bandwidth tunnels, there is no possibility of traffic being rerouted onto a path, which as a result becomes congested, because implicitly a new network level admission control decision is made before the traffic is rerouted. In the example above, unlike the cases described in sections 2.1 and 2.2, the service to Call B is not disrupted because the use of TE for network level admission control ensures that the network rerouting behaviour is kept in lock-step with the bandwidth manager's view of the network.

If there is sufficient bandwidth on another path for the tunnel to be rerouted, the time difference between t1 and t2 is the time to re-signal the TE tunnel; during this time, there is loss of connectivity and service disruption to those calls directly impacted by the failed network element. This is typically in the order of 1-2 seconds; this time can be reduced to in the order of 50ms through the use of MPLS TE FRR (see section 2.4).

2.4 MPLS TE with specified tunnel bandwidth and FRR

This model builds on the one described in section 2.3 by adding MPLS TE Fast Reroute (FRR) backup tunnels to minimise the loss of connectivity and service disruption caused to those calls directly impacted by the failed network element while the TE tunnel is being re-signalled, i.e. the time difference between t1 and t2 from section 2.3. To illustrate this point consider the network topology shown in Figure 6 (for simplicity assume all links = 1Mbps) and the following sequence of events:

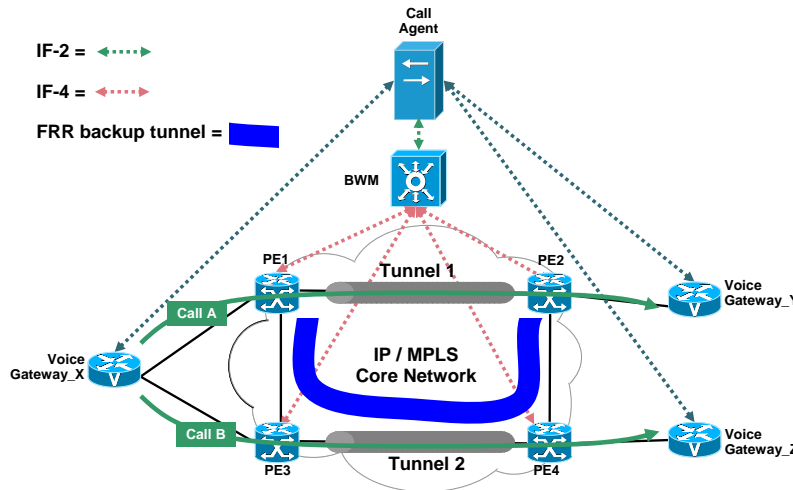


Figure 6. MPLS TE with specified tunnel bandwidth and FRR: example step #2

1. As per steps 1-6 from section 2.3
2. An FRR backup TE tunnel is setup from PE1→PE3→PE4→PE2 using an explicit (i.e. static) path option, to protect primary tunnels from PE1 on the failure of link PE1→PE2
3. (t1) Link PE1↔PE2 fails
4. (t2) router PE1 reroutes Tunnel 1 over the FRR backup tunnel (as shown in Figure 7) which typically happens in sub 50ms. Whether there is the possibility of congestion on the backup paths depends upon whether bandwidth protection is used for the backup tunnel:
 - a. If bandwidth protection is used (i.e. bandwidth is specified and reserved for the FRR

backup tunnel), there can be no possibility of congestion on the backup path; this is analogous to 1+1 protection and can be inefficient with respect to use of backup bandwidth

- b. If bandwidth protection is not used (i.e. bandwidth is not specified for the FRR protection tunnel), then there may be the possibility of congestion on the backup path. There are a number of potential techniques which can be used to reduce or mitigate this possibility, but they are not discussed in detail in this paper.

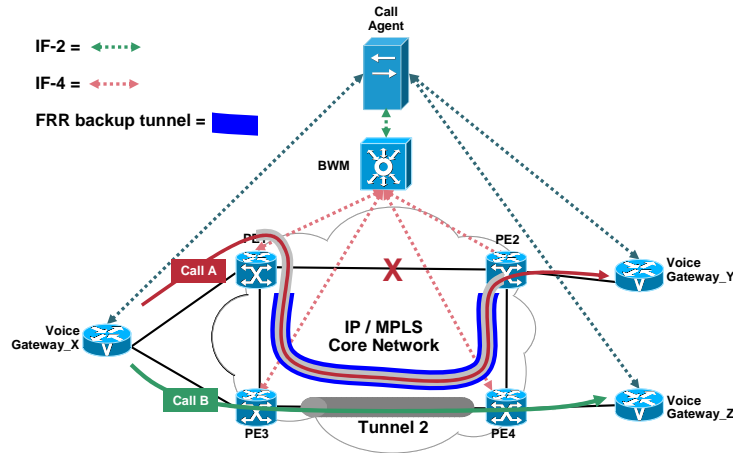


Figure 7. MPLS TE with specified tunnel bandwidth and FRR: example step #4

- 5. Router PE1 uses CSPF to recalculate the path for Tunnel 1;
 - a. If there are no paths available which meet the bandwidth constraint, Tunnel 1 is not able to be re-established and Call A is impacted by the failure but congestion on Link PE3→PE4 is avoided (hence the service to Call B is preserved), and the bandwidth manager is informed.
The bandwidth manager may clear down Call A by signalling to the call agent or Call A may be cleared down by bearer level packet loss detection in the gateways.
 - b. (t3) If there were another path available with sufficient bandwidth for the rerouted tunnel, it would be re-established.

FRR can reduce the loss of connectivity and service disruption caused to those calls directly impacted by the failed network element to in the order of 50ms.

3 Other Network Considerations with Respect to the Bandwidth Manager Technical Report

3.1 Bandwidth Re-use

Section 1.4.1 of [1] raises the issue of bandwidth partitioning when edge-to-edge TE is used; this should be considered in terms of both data plane and control plane behaviours. TE may reserve bandwidth associated with a particular class on a particular link from a control plane perspective, and that reserved bandwidth may align with minimum scheduler bandwidth assurances for that class, however, a work-conserving scheduler will allow unused bandwidth from that class to be re-used by other classes (which are presumably not using TE), i.e. at the data plane the TE bandwidth is assured but not partitioned.

3.2 TE tunnel scalability

Section 1.4.1.3 of [1] refers to a concatenation of TE tunnels from ingress head-end router to egress tail-end router, with a label stack imposed at the ingress head-end as a technique for improving the scalability of a TE tunnel deployment by reducing the number of tunnels needed. This model requires that the bandwidth manager specifies the label stack (i.e. determines the end-to-end path of tunnels) to be used by the tunnel head-end router for a particular call. Two other deployment models may be considered, however, which achieve the same effect:

1. Tunnels may be concatenated without the need to impose a label stack at the ingress head-end. In this case, traffic is switched from one tunnel to another at a combined tail-end / head-end router based upon a routing table lookup at that router.
2. It is possible to aggregate tunnels into tunnels rather than to concatenate tunnels. This does not require that a label stack is imposed at the ingress head-end router, but rather an additional label is added to the stack at the head-end of each aggregating tunnel, which is then popped off the stack at the penultimate hop of that tunnel.

The two deployment models described above are fully supported by existing IETF standards.

3.2.1 Impact on R2 IF-4 Interface

The two deployment models described above for scaling TE tunnel deployments do not require that the bandwidth manager specify to a TE tunnel head-end the end-to-end path of tunnels that should be used for a particular call, by way of specifying a label stack. Rather the head-end determines which tunnel will be used for each call, and the bandwidth manager needs to be able to resolve a particular call to the TE tunnels that it will use, and to track the status of the tunnels.

This potentially leads to a significant simplification of the requirements for the R2 IF-4 interface described in [1] (which is redefined as the TC-2 interface in the R3 architecture), which may be realisable with already standardised interfaces, such as SNMPv3. Hence, as follow on work to this paper, the IF-4 requirements will be reconsidered with the target of potentially demonstrating a simplified IF-4 interface for GMI2006.

4 Conclusions

Network level behaviours can impact the determinism of call admission control decisions for a particular bandwidth management deployment:

- Any of the deployment models described in section 2 could be used to provide a deterministic CAC capability, if only network working case conditions are important
- If CAC is needed to cover network element failure case conditions and transient congestion following network failures is to be avoided then the “MPLS TE with specified tunnel bandwidth” model described in section 2.3 is required
- The loss of connectivity and service disruption caused to those calls directly impacted by a failed network element can be reduced to in the order of 50ms using MPLS TE FRR as

- described in section 2.4
- A model where the bandwidth manager does not specify the label stack to the head-end router for a particular call potentially leads to a significant simplification of the IF-4 interface; such a simplified interface may be demonstrable for GMI2006

5 References

- [1] MSF-TR-ARCH-005-FINAL, "Bandwidth Management in Next Generation Packet Networks"
- [2] J. Moy, "OSPF Version 2", RFC 2328, April 1998
- [3] Callon, R., "OSI IS-IS for IP and Dual Environment", RFC 1195, December 1990.
- [4] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow,, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, December 2001
- [5] D. Katz, K. Kompella, D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC3630, September 2003
- [6] H. Smit, T. Li, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)", RFC3784, June 2004
- [7] Pierre Francois, Clarence Filisfil, John Evans and Olivier Bonaventure, "Achieving subsecond IGP convergence in large IP networks", ACM SIGCOMM Computer Communication Review, Vol. 35, Issue 3 (July 2005), pp. 35-44
- [8] P. Pan, Ed., G. Swallow, Ed., A. Atlas, Ed, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC4090, May 2005