Francois Le Faucheur, Editor
Thomas Nadeau
Cisco Systems, Inc.

Jim Boyle
PDNets

Kireeti Kompella
Juniper Networks

William Townsend
Tenor Networks

Darek Skalecki
Nortel Networks

IETF Internet Draft
Expires: December, 2002
Document: draft-ietf-tewg-diff-te-proto-01.txt June, 2002

Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering

Status of this Memo

Abstract

This document specifies the IGP and signaling extensions and procedures (beyond those already specified for existing MPLS Traffic Engineering) for support of Diff-Serv-aware MPLS Traffic Engineering.

A default Bandwidth Constraints model for Diff-Serv-aware Traffic Engineering (the Russian Dolls model) is also specified.

1. Introduction

[DSTE-REQ] presents the Service Providers requirements for support of Diff-Serv-aware MPLS Traffic Engineering (DS-TE). This includes the fundamental requirement to be able to enforce different bandwidth constraints for different classes of traffic.

This document specifies:

- the IGP and signaling extensions and procedures (beyond those already specified for existing MPLS Traffic Engineering [OSPF-TE][ISIS-TE][RSVP-TE][CR-LDP]) for support of the DS-TE requirements [DSTE-REQ] in environments relying on distributed Constraint Based Routing (i.e. path computation involving Head-end LSRs).
- A default Bandwidth Constraint Model for DS-TE called the Russian Dolls model. While DS-TE implementations may support other Bandwidth Constraints model, they must all support the Russian Dolls model to ensure interoperability across all implementations.

2. Definitions

[DSTE-REQ] discusses how a Head-end LSR may split the set of Ordered Aggregates from the traffic to a given Tail-end into multiple Traffic Trunks. Each Traffic Trunk is transported over a separate LSP which is Constraint Based Routed individually.

For readability a number of definitions from [DSTE-REQ] are repeated here:

Traffic Trunk: an aggregation of traffic flows of the same class [i.e. which are to be treated equivalently from the DS-TE perspective] which are placed inside a Label Switched Path.

Class-Type (CT): the set of Traffic Trunks crossing a link that is governed by a specific set of Bandwidth constraints. CT is used for the purposes of link bandwidth allocation, constraint based routing and admission control. A given Traffic Trunk belongs to the same CT on all links. A given Traffic Trunk belongs to the same CT on all links.

TE-Class: A pair of:

i. a Class-Type
ii. a preemption priority allowed for that Class-Type. This means that an LSP transporting a Traffic Trunk from that Class-Type can use that preemption priority as the set-up priority, as the holding priority or both.

3. Configurable Parameters

This section only discusses the differences with the configurable parameters supported for MPLS Traffic Engineering as per [TE-REQ], [ISIS-TE], [OSPF-TE], [RSVP-TE] and [CR-LDP]. All other parameters are unchanged.

3.1. Link Parameters

3.1.1. Bandwidth Constraints (BCs)

[DSTE-REQTS] states that "Regardless of the Bandwidth Constraint Model, the DS-TE solution must allow support for up to 8 BCs."

For DS-TE, the existing "Maximum Reservable link bandwidth" parameter is retained but its semantic is generalized and interpreted as BC0.

Additionally, on every link, a DS-TE implementation MUST provide for configuration of up to 7 additional link parameters which are the seven other potential Bandwidth Constraints i.e. BC1, BC2 , ... BC7.

The LSR MUST interpret these Bandwidth Constraints in accordance with the supported Bandwidth Constraint Model (i.e. what bandwidth

constraint applies to what Class-Type and how).

Where the Bandwidth Constraint Model imposes some relationship among the values to be configured for these Bandwidth Constraints, the LSR MUST enforce those at configuration time. For example, with the "Russian Doll" Bandwidth Constraints Model defined below in section 9, the LSR must ensure that BCi is configured smaller or equal to BCj, where i is greater than j.

3.1.2. per-CT Local Overbooking Multipliers (LOMs)

DS-TE enables a network administrator to apply different overbooking (or underbooking) ratios for different CTs.

The principal method to achieve this is the same as historically used in existing TE deployment, which is :

(i) to take into account the over-booking/underbooking ratio appropriate for the OA/CT associated with the considered LSP at the time of establishing the bandwidth size of a given LSP, AND/OR

(ii) to take into account the overbooking/underbooking ratio at the time of configuring the Maximum Reservable Bandwidth/Bandwidth Constraints and/or the Maximum Link Bandwidth (which effectively controls the maximum size of an individual LSP) and use values which are larger(overbooking) or smaller(underbooking) than the actual link.

We refer to this method as the "LSP/link size overbooking" method.

The "LSP/link size overbooking" method is expected to be often sufficient in many DS-TE environments and requires no additional configurable parameters.

However, in the particular DS-TE environments where, for a given CT, the overbooking ratio needs to be tweaked differently on different links and where a very fine accounting of overbooking cross-effect across Class-Types is required, a DS-TE implementation MAY optionally support the "local overbooking" method as a complement to the "LSP/link size overbooking" method. The "local overbooking" method

relies on optional "per-CT Local Overbooking Multipliers" (LOMs) which are configurable, on every link, for every CT. The per-CT Local Overbooking Multiplier effectively allows the network operator to increase/decrease", on some links, the overbooking ratio already enforced by the "LSP/link size overbooking" method. This is achieved by factoring the per-CT LOM in all local bandwidth accounting for the purposes of admission control and IGP advertisement of unreserved bandwidths. Exact details on how the LOMs need to be factored in depend on the Bandwidth Constraints model. This is discussed for the Russian Dolls model in section 9.

Since the per-CT Local Overbooking Multipliers are factored in the IGP advertisement of unreserved bandwidth by the local LSR, a remote LSR computing a path for a DS-TE tunnel need not be aware that local overbooking is used on the considered link. In fact, the remote LSR does not even need to support the optional local overbooking method. In any case, the remote LSR will compute a path that effectively takes into account the LOMs on the considered link because it bases its computation on advertised unreserved bandwidth which do factor in the LOMs for that link.

3.2. LSR Parameters

3.2.1. TE-Class Mapping

In line with [DSTE-REQ], the preemption attributes defined in [TE-REQ] are retained with DS-TE and applicable across all Class Types. The preemption attributes of setup priority and holding priority retain existing semantics, and in particular these semantics are not affected by the Ordered Aggregate transported by the LSP or by the LSP's Class Type. This means that if LSP1 contends with LSP2 for resources, LSP1 may preempt LSP2 if LSP1 has a higher set-up preemption priority (i.e. lower numerical priority value) than LSP2's holding preemption priority regardless of LSP1's OA/CT and LSP2's OA/CT.

For DS-TE, LSRs MUST allow configuration of a TE-Class mapping whereby the Class-Type and preemption level are configured for each of (up to) 8 TE-Classes.

This mapping is referred to as :

TE-Class[i] <--> < CTc , preemption p >

Where 0 <= i <= 7, 0 <= c <= 7, 0 <= p <= 7

Two TE-Classes must not be identical (i.e. have both the same Class-Type and the same preemption priority).

Where the network administrator uses less than 8 TE-Classes, the DS-TE LSR MUST allow remaining ones to be configured as "Unused".

There are no other restrictions on how any of the 8 Class-Types can be paired up with any of the 8 preemption priorities to form a TE-class. In particular, one given preemption priority can be paired up with two (or more) different Class-Types to form two (or more) TE-classes. Similarly, one Class-Type can be paired up with two (or more) different preemption priorities to form two (or more) TE-Classes. Also, there is no mandatory ordering relationship between the TE-Class index (i.e. "i" above) and the Class-Type (i.e. "c" above) or the preemption priority (i.e. "p" above) of the TE-Class.

To ensure coherent DS-TE operation, the network administrator MUST configure exactly the same TE-Class Mapping on all LSRs of the DS-TE domain.

When the TE-class mapping needs to be modified in the DS-TE domain, care must be exercised during the transient period of reconfiguration during which some DS-TE LSRs may be configured with the new TE-class mapping while others are still configured with the old TE-class mapping. It is recommended that active tunnels do not use any of the TE-classes which are being modified during such a transient reconfiguration period.

3.3. LSP Parameters

3.3.1. Class-Type

With DS-TE, LSRs MUST support, for every LSP, an additional configurable parameter which indicates the Class-Type of the Traffic Trunk transported by the LSP.

There is one and only one Class-Type configured per LSP.

The configured Class-Type indicates, in accordance with the supported Bandwidth Constraint Model, what are the Bandwidth Constraints applicable to that LSP.

3.3.2. Setup and Holding Preemption Priorities

As per existing TE, DS-TE assumes that every DS-TE LSP is configured with a setup and holding priority, each with a value between 0 and 7.

3.3.3. Class-Type/Preemption Relationship

With DS-TE, the preemption priority configured for the setup priority of a given LSP and the Class-Type configured for that LSP must be such that, together, they form one of the (up to) 8 TE-Classes configured in the TE-Class Mapping specified is section 3.2.1 above.

The LSR MUST enforce this rule at configuration time.

The preemption priority configured for the holding priority of a given LSP and the Class-Type configured for that LSP must also be such that, together, they form one of the (up to) 8 TE-Classes configured in the TE-Class Mapping specified is section 3.2.1 above.

The LSR MUST enforce this rule at configuration time.

3.4. Examples of Parameters Configuration

For illustrative purposes, we now present a few examples of how these configurable parameters may be used. All these examples assume that different bandwidth constraints need to be enforced for different sets of Traffic Trunks (e.g. for Voice and for Data) so that two, or more, Class-Types must be used.

3.4.1. Example 1

The Network Administrator of a first network using two Class Types (CT1 for Voice and CT0 for Data), may elect to configure the following TE-Class Mapping to ensure that Voice LSPs are never driven away from their shortest path because of Data LSPs:

```
TE-Class[0]  <-->  < CT1 , preemption 0 >
TE-Class[1]  <-->  < CT0 , preemption 1 >
TE-Class[i]  <-->  unused,   for 2 <= i <= 7
```

Voice LSPs would then be configured with:
- CT=CT1, set-up priority =0, holding priority=0

Data LSPs would then be configured with:
- CT=CT0, set-up priority =1, holding priority=1

A new Voice LSP would then be able to preempt an existing Data LSP in case they contend for resources. A Data LSP would never preempt a Voice LSP. A Voice LSP would never preempt another Voice LSP. A Data LSP would never preempt another Data LSP.

3.4.2. Example 2

The Network Administrator of another network may elect to configure the following TE-Class Mapping in order to optimize global network

resource utilization by favoring placement of large LSPs closer to their shortest path:

```
TE-Class[0]  <-->  < CT1 , preemption 0 >
TE-Class[1]  <-->  < CT0 , preemption 1 >
TE-Class[2]  <-->  < CT1 , preemption 2 >
TE-Class[3]  <-->  < CT0 , preemption 3 >
TE-Class[i]  <-->  unused,   for 4 <= i <= 7
```

Large size Voice LSPs could be configured with:
- CT=CT1, set-up priority =0, holding priority=0

Large size Data LSPs could be configured with:
- CT=CT0, set-up priority = 1, holding priority=1

Small size Voice LSPs could be configured with:
- CT=CT1, set-up priority = 2, holding priority=2

Small size Data LSPs could be configured with:
- CT=CT0, set-up priority = 3, holding priority=3.

A new large size Voice LSP would then be able to preempt a small size Voice LSP or any Data LSP in case they contend for resources.

A new large size Data LSP would then be able to preempt a small size Data LSP or a small size Voice LSP in case they contend for resources, but it would not be able to preempt a large size Voice LSP.

3.4.3. Example 3

The Network Administrator of another network may elect to configure the following TE-Class Mapping in order to ensure that Voice LSPs are never driven away from their shortest path because of Data LSPs while also achieving some optimization of global network resource utilization by favoring placement of large LSPs closer to their shortest path:

```
     TE-Class[0]  <-->  < CT1 , preemption 0 >
     TE-Class[1]  <-->  < CT1 , preemption 1 >
     TE-Class[2]  <-->  < CT0 , preemption 2 >
     TE-Class[3]  <-->  < CT0 , preemption 3 >
     TE-Class[i]  <-->  unused,   for 4 <= i <= 7
```

Large size Voice LSPs could be configured with:
- CT=CT1, set-up priority = 0, holding priority=0.

Small size Voice LSPs could be configured with:
- CT=CT1, set-up priority = 1, holding priority=1.

Large size Data LSPs could be configured with:
- CT=CT0, set-up priority = 2, holding priority=2.

Small size Data LSPs could be configured with:
- CT=CT0, set-up priority = 3, holding priority=3.

A Voice LSP could preempt a Data LSP if they contend for resources. A Data LSP would never preempt a Voice LSP. A Large size Voice LSP could preempt a small size Voice LSP if they contend for resources. A Large size Data LSP could preempt a small size Data LSP if they contend for resources.

3.4.4. Example 4

The Network Administrator of another network may elect to configure the following TE-Class Mapping in order to ensure that no preemption occurs in the DS-TE domain:

```
     TE-Class[0]  <-->  < CT1 , preemption 0 >
     TE-Class[1]  <-->  < CT0 , preemption 0 >
     TE-Class[i]  <-->  unused,   for 2 <= i <= 7
```

Voice LSPs would then be configured with:

- CT=CT1, set-up priority =0, holding priority=0

Data LSPs would then be configured with:

- CT=CT0, set-up priority =0, holding priority=0

No LSP would then be able to preempt any other LSP.

3.4.5. Example 5

The Network Administrator of another network may elect to configure the following TE-Class Mapping in view of increased network stability through a more limited use of preemption:

```
     TE-Class[0]  <-->  < CT1 , preemption 0 >
     TE-Class[1]  <-->  < CT1 , preemption 1 >
     TE-Class[2]  <-->  < CT0 , preemption 1 >
     TE-Class[3]  <-->  < CT0 , preemption 2 >
     TE-Class[i]  <-->  unused,   for 4 <= i <= 7
```

Large size Voice LSPs could be configured with:

- CT=CT1, set-up priority = 0, holding priority=0.

Small size Voice LSPs could be configured with:

- CT=CT1, set-up priority = 1, holding priority=0.

Large size Data LSPs could be configured with:

- CT=CT0, set-up priority = 2, holding priority=1.

Small size Data LSPs could be configured with:

- CT=CT0, set-up priority = 2, holding priority=2.

A new large size Voice LSP would be able to preempt a Data LSP in case they contend for resources, but it would not be able to preempt any Voice LSP even a small size Voice LSP.

A new small size Voice LSP would be able to preempt a small size Data LSP in case they contend for resources, but it would not be able to preempt a large size Data LSP or any Voice LSP.

A Data LSP would not be able to preempt any other LSP.

4. IGP Advertisement

This section only discusses the differences with the IGP advertisement supported for MPLS Traffic Engineering as per [OSPF-TE] and [ISIS-TE]. The rest of the IGP advertisement is unchanged.

4.1. Bandwidth Constraints

As detailed above in section 3.1.1, up to 8 Bandwidth Constraints (BCb, 0 <= b <= 7) are configurable on any given link.

With DS-TE, the existing "Maximum Reservable Bw" sub-TLV is retained with a generalized semantic so that it is now interpreted as Bandwidth Constraint 0 (BC0).

DS-TE also defines the following optional sub-TLV to advertise the eight potential Bandwidth Constraints (BC0 to BC7):

"Bandwidth Constraints" sub-TLV:
TBD - Bandwidth Constraint Model Id (1 octet)
- Bandwidth Constraints (Nx4 octets)

Where:

- Bandwidth Constraint Model Id: 1 octet identifier for the Bandwidth Constraints Model currently in use by the LSR initiating the IGP advertisement.
  Values 0 to 127 are to be allocated by the TEWG to identify Bandwidth Constraints Models defined in the TEWG. Value 0 identifies the Russian Doll Bandwidth Constraint Model defined in section 9.
  Values 128 to 255 are for experimental use.

- Bandwidth Constraints: contains BC0, BC1,... BCN-1.
  Each Bandwidth Constraint is encoded in 32 bits in IEEE floating point format. The units are bytes (not bits!) per second. It is recommended that only the Bandwidth Constraints corresponding to active CTs be advertised in order to

minimize the impact on IGP scalability.

A DS-TE LSR MAY optionally advertise Bandwidth Constraints.

A DS-TE LSR which does advertise Bandwidth Constraints MUST use the new "Bandwidth Constraints" sub-TLV to do so. For example, where a Service Provider deploys DS-TE with two active CTs, only two Bandwidth Constraints per link would be meaningful (assuming, for instance, the Russian Doll Bandwidth Constraint Model defined in section 9). A DS-TE LSR which does advertise Bandwidth Constraint would include the "Bandwidth Constraints" sub-TLV in the IGP advertisement and this should contain BC0 and BC1.

A DS-TE LSR which does advertise Bandwidth Constraints, MAY also include the existing "Maximum Reservable Bw" sub-TLV. This may be useful in migration situations where some LSRs in the network are not DS-TE capable (see Appendix G) and thus do not understand the new "Bandwidth Constraints" sub-TLV. IN that case, the DS-TE LSR MUST set the value of the "Maximum Reservable Bw" sub-TLV to the same value as the one for BC0 encoded in the "Bandwidth Constraints" sub-TLV.

A DS-TE LSR receiving both the old "Maximum Reservable Bw" sub-TLV and the new "Bandwidth Constraints" sub-TLV for a given link MAY ignore the "Maximum Reservable Bw" sub-TLV.

A DS-TE LSR receiving the "Bandwidth Constraints" sub-TLV with a Bandwidth Constraint Model Id which does not match the Bandwidth Constraint Model it currently uses, MAY generate a warning to the operator reporting the inconsistency between Bandwidth Constraint Models used on different links. If the DS-TE LSR does not support the Bandwidth Constraint Model designated by the Bandwidth Constraint Model Id, or if the DS-TE LSR does not support operations with multiple simultaneous Bandwidth Constraint Models, the DS-TE LSR MAY ignore the corresponding TLV.

4.2. Unreserved Bandwidth

With DS-TE, the existing "Unreserved Bandwidth" sub-TLV is retained

as the only vehicle to advertise dynamic bandwidth information necessary for Constraint Based Routing on Head-ends, except that it is used with a generalized semantic. The Unreserved Bandwidth sub-TLV still carries eight bandwidth values but they now correspond to the unreserved bandwidth for each of the TE-Class (instead of for each preemption as per existing TE).

More precisely, a DS-TE LSR MUST support the Unreserved Bandwidth sub-TLV with a definition which is generalized into the following:

The Unreserved Bandwidth sub-TLV specifies the amount of bandwidth not yet reserved for each of the eight TE-classes, in IEEE floating point format arranged in increasing order of TE-Class index, with unreserved bandwidth for TE-Class [0] occurring at the start of the sub-TLV, and unreserved bandwidth for TE-Class [7] at the end of the sub-TLV. The unreserved bandwidth value for TE-Class [i] (0 <= i <=

7) is referred to as "Unreserved TE-Class [i]". It indicates the bandwidth that is available, for reservation, to an LSP which :

- transports a Traffic Trunk from the Class-Type of TE-Class[i], and
- has a setup priority corresponding to the preemption priority of TE-Class[i].

The units are bytes per second.

Since the bandwidth values are now ordered by TE-class index and thus can relate to different CTs with different bandwidth constraints and can relate to any arbitrary preemption priority, a DS-TE LSR MUST NOT assume any ordered relationship among these bandwidth.

With existing TE, since all preemption priorities reflect the same (and only) bandwidth constraints and since bandwidth values are advertised in preemption priority order, the following relationship is always true, and is often assumed by TE implementations:

If i < j , then "Unreserved Bw [i]" >= "Unreserved Bw [j]"

With DS-TE, no relationship is to be assumed so that:

```
     If i < j , then
              "Unreserved TE-Class [i]" = "Unreserved TE-Class [j]"
                  OR
              "Unreserved TE-Class [i]" > "Unreserved TE-Class [j]"
                  OR
              "Unreserved TE-Class [i]" < "Unreserved TE-Class [j]".
```

Since some Bandwidth Constraints Models are such that a given Class-Type is constrained by multiple Bandwidth Constraints (as in the case of the Russian Doll Bandwidth Constraint Model specified in section 9), the value to be advertised by the IGP in "Unreserved TE-Class [i]" MUST reflect all of the Bandwidth Constraints relevant to the CT associated with TE-Class [i].

If TE-Class[i] is unused, the value advertised by the IGP in "Unreserved TE-Class [i]" MUST be set to zero.

4.3. Local Overbooking Multiplier

The following additional optional sub-TLV is defined for DS-TE:

"Local Overbooking Multiplier" sub-TLV:

TBD - per-CT Local Overbooking Multipliers (N x 2 octets). Each LOM is encoded as an integer in the range [0 , (2^16 -1)] that represents LOM expressed in percentage. For example, a LOM of 1 (i.e. 100%) is encoded as 100, and represents no overbooking/underbooking. A LOM of 2 (i.e. 200%) is encoded as 200, and represents an overbooking of 2. A LOM of 0.5 (i.e. 50%) is encoded as 50 and represents an underbooking of 2.

where N is the number of per-CT Local Overbooking Multipliers actually advertised. It is recommended that only the LOMs corresponding to active CTs be included, in order to minimize the impact on IGP scalability.

A DS-TE LSR supporting the optional local overbooking method MAY optionally advertise LOMs. A DS-TE LSR which does advertise LOMs MUST use the "Local Overbooking Multiplier" sub-TLV to do so.

For example, where a Service Provider only deploys DS-TE with two CTs and makes use of the Local Overbooking method, the "Local Overbooking Multiplier" sub-TLV may optionally be used and would then contain only LOM[0] and LOM[1.

Note that the use of this sub-TLV is only optional even when the optional Local Overbooking method is actually used (and thus when the Local Overbooking Multipliers parameters are actually configured locally on some or all links). Its use may assist in head-end prediction of network response to LSP establishment.

5. LSP Signaling

This section only describes the signaling extensions beyond those already specified for MPLS Traffic Engineering as per [RSVP-TE] and [CR-LDP] and for Diff-Serv over MPLS as per [DIFF-MPLS].

The Class-Type of the LSP is signaled in RSVP-TE and CR-LDP for DS-TE in order for LSRs to enforce the appropriate bandwidth constraint(s) for admission control and bandwidth accounting.

Protocol and procedure extensions for signaling of the Class-Type are specified in details in Appendix A and B respectively for RSVP-TE and CR-LDP.

A DS-TE implementation based on RSVP-TE MUST support the protocol and procedure extensions specified in Appendix A.

A DS-TE implementation based on CR-LDP MUST support the protocol and procedure extensions specified in Appendix B.

6. Constraint Based Routing

Let us consider the case where a path needs to be computed for an LSP whose Class-Type is configured to CTc and whose set-up preemption priority is configured to p.

Then the pair of CTc and p will map to one of the TE-Classes defined in the TE-Class mapping. Let us assume that this is the i-th TE-Class i.e. TE-Class[i].

The Constraint Based Routing algorithm of a DS-TE LSR is still only required to perform path computation satisfying a single bandwidth constraint which is to fit in "Unreserved TE-Class [i]" as advertised by the IGP for every link. Thus, no changes are required to the existing TE Constraint Based Routing algorithm itself.

The Constraint Based Routing algorithm MAY also optionally take into account, when used, the optional information advertised in IGP which are the Bandwidth Constraints and the Local Overbooking Multipliers. As an example, the Bandwidth Constraints MIGHT be used as a tie-breaker criteria in situations where multiple paths, otherwise equally attractive, are possible.

7. Diff-Serv scheduling

The Class-Type signaled at LSP establishment MAY optionally be used by DS-TE LSRs to dynamically adjust the resources allocated to the Class-Type by the Diff-Serv scheduler. In addition, the Diff-Serv information (i.e. the PSC) signaled by the TE-LSP signaling protocols as specified in [DIFF-MPLS], if used, MAY optionally be used by DS-TE LSRs to dynamically adjust the resources allocated to a PSC/OA within a Class Type by the Diff-Serv scheduler.

8. Existing TE as a Particular Case of DS-TE

We observe that existing TE can be viewed as a particular case of DS-TE where:

(i) a single Class-Type is used, all 8 preemption priorities are allowed for that Class-Type and the following TE-Class Mapping is used:

TE-Class[i] <--> < CT0 , preemption i >
Where 0 <= i <= 7.

(ii) optional per-CT Local Overbooking Multipliers are not used.

In that case, DS-TE behaves as existing TE.

As with existing TE, the IGP advertises:

- Unreserved Bandwidth for each of the 8 preemption priorities
- Max Link Bandwidth

As with existing TE, the IGP may also optionally advertise BC0 = Maximum Reservable Bandwidth.

Note that, unlike with existing TE, the IGP will also advertise the Bandwidth Constraint sub-TLV (which will contain BC0).

Since all LSPs transport traffic from CT0, LSP Signaling is done without explicit signaling of the Class-Type (which is only used for other Class-Types than CT0 as explained in Appendix A and B).

9. Russian Doll Bandwidth Constraints Model

9.1. Definition

[DSTE-REQ] introduces the concept of Bandwidth Constraints Models to characterize the Bandwidth Constraints associated with CTs, but it does not actually select one particular Model.

[DSTE-REQ] also requires that a default Bandwidth Constraints Model be specified. The purpose of such a default Model is to ensure that there is at least one common Bandwidth Constraints model supported across all DS-TE implementations to allow for easier deployment of DS-TE.

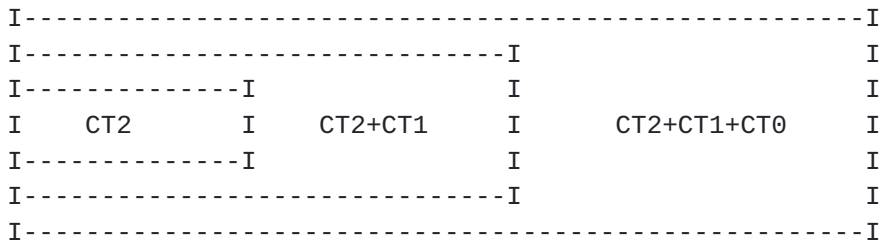
This section specifies a default Bandwidth Constraints Model which is referred to as the "Russian Dolls" Bandwidth Constraints model. DS-TE implementations MUST support the Russian Dolls models and MAY also optionally support other Bandwidth Constraints Models.

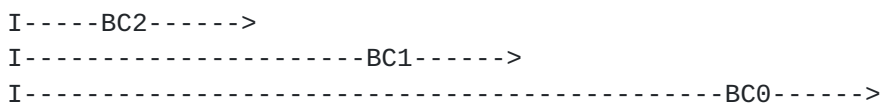
The "Russian Dolls" model is defined in the following manner (assuming that the optional per-CT Local Overbooking Multipliers are not used - i.e. LOM[c]=1 , 0<=c<=7):

- Maximum Number of Bandwidth Constraints (MaxBC)= Maximum Number of Class-Types (MaxCT) = 8

o All LSPs supporting Traffic Trunks from CTc (with b<=c<=7) use no more than BCb i.e.:
  - All LSPs from CT7 use no more than BC7
  - All LSPs from CT6 and CT7 use no more than BC6
  - All LSPs from CT5, CT6 and CT7 use no more than BC5
  - etc.
  - All LSPs from CT0, CT1,... CT7 use no more than BC0

Purely for illustration purposes, the diagram below represents the Russian Doll Bandwidth Constraints model in a pictorial manner when 3 Class-Types are active:





While simpler or, conversely, more flexible/sophisticated Bandwidth Constraints models can be defined, the Russian Dolls model is an attractive trade-off for the following reasons:

- Network administrators generally find it superior to the most basic model of a single independent BC per CT (which, in typical deployment scenarios, results in either capacity wastage, low priority Traffic Trunk starvation and/or degradation of QoS objectives)
- network administrators generally find it sufficient for the real life deployments currently anticipated (e.g. it addresses all the scenarios described in [DSTE-REQ])
- it remains simple and only requires limited protocol extensions, while more sophisticated Bandwidth Constraints

model may require more complex extensions.

More details on the properties of the Russian Dolls model can be found in [RUSSIAN] and [BCMODEL].

As an example usage of the "Russian Doll" Bandwidth Constraints Model, a network administrator using one CT for Voice (CT1) and one CT for data (CT0) might configure on a given link:

- BC0 = 2.5 Gb/s (i.e. Voice + Data is limited to 2.5 Gb/s)
- BC1= 1.5 Gb/s (i.e. Voice is limited to 1.5 Gb/s).

Another (or other) Bandwidth Constraints Model(s) may be specified in another (or other) document(s) to address other potential requirements which may emerge from Service Providers real life deployment and which cannot be efficiently addressed by the Russian Dolls model. This is outside the scope of the current specification.

The Russian Doll Bandwidth Constraints Model can be supported with the extensions defined earlier in this document for DS-TE. Note that a number of other Bandwidth Constraints could also be supported with these same extensions. Note also that not all Bandwidth Constraints models could be supported with these extensions and some may require additional or different extensions. Both of these situations are beyond the scope of this specification.

Note that as per [DSTE-REQTS], while deployment of DS-TE is expected to be easier when a single Bandwidth Constraints Model is used on all the DS-TE LSRs of a DS-TE domain, the DS-TE solution itself does not prevent a network operator to activate different Bandwidth Constraints models on different links in a network, if he/she wishes to do so and if the particular DS-TE LSR implementations allow this.

9.2. Computing "Unreserved TE-Class [i]"

We first observe that, for existing TE, details on admission control algorithms for TE LSPs, and consequently details on formulas for computing the unreserved bandwidth, are outside the scope of the current IETF work. This is left for vendor differentiation. Note that

this does not compromise interoperability across various implementations since the TE schemes rely on LSRs to advertise their local view of the world in terms of Unreserved Bw to other LSRs. This way, regardless of the actual local admission control algorithm used on one given LSR, Constraint Based Routing on other LSRs can rely on advertised information to determine whether an additional LSP will be accepted or rejected by the given LSR. The only requirement is that an LSR advertises unreserved bandwidth values which are consistent with its specific local admission control algorithm and take into account the holding preemption priority of established LSPs.

In the context of DS-TE, again, details on admission control algorithms are left for vendor differentiation and formulas for computing the unreserved bandwidth for TE-Class[i] are outside the scope of this specification. However, DS-TE places the additional requirement on the LSR that the unreserved bandwidth values advertised MUST reflect all of the Bandwidth Constraints relevant to the CT associated with TE-Class[i], as discussed in section 4.2.

As with existing TE, DS-TE assumes that the holding preemption priority of established LSPs is the one considered (as opposed to their set-up preemption priority) for the purpose of computing the unreserved bandwidth for TE-Class [i].

Example formulas for computing "Unreserved TE-Class [i]" are provided in Appendix C.

9.3. Admission Control Rules

A DS-TE LSR MUST support the following admission control rule:

Regardless of how the admission control algorithm actually computes the unreserved bandwidth for TE-Class[i] for one of its local link, an LSP of bandwidth B, of set-up preemption priority p and of Class-Type CTc is admissible on that link iff:

B <= unreserved bandwidth for TE-Class[i], AND
B <= Max Link Bandwidth

Where

- TE-Class [i] maps to < CTc , p > in the LSR's configured TE-Class mapping
- Max Link Bandwidth is the maximum link bandwidth configured on the link and advertised in IGP.

Note that this admission control rule assumes that the optional per-CT Local Overbooking Multipliers are not used (i.e. LOM[c]=1, 0<=c<= 7).

9.4. Support of Optional Local Overbooking Method

We remind the reader that, as discussed in section 3.1.2, the "LSP/link size overbooking" method (which does not use the Local Overbooking Multipliers) is expected to be sufficient in many DS-TE environments. It is expected that the optional Local Overbooking method (and LOMs) would only be used in specific environments, in particular where different overbooking ratios need to be enforced on different links of the DS-TE domain and cross-effect of overbooking across CTs needs to be accounted for very accurately.

This section discusses the impact of the optional local overbooking method on the Russian Dolls model and associated rules and formula. This is only applicable in the cases where the optional local overbooking method is indeed supported by the DS-TE LSRs and actually deployed.

9.4.1. Russian Dolls Model Definition

Let us define "Reserved(CTc)" as the sum of the bandwidth reserved by all established LSPs which belong to CTc.

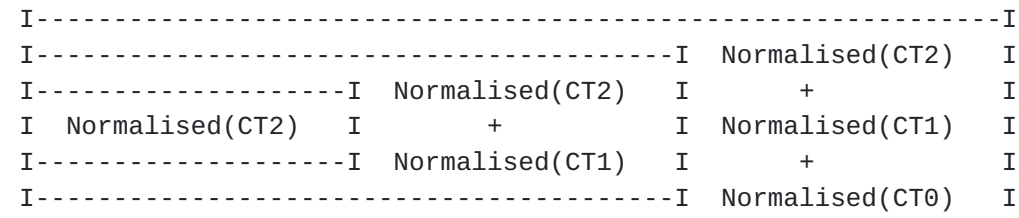
Let us define "Normalised(CTc)" as "Reserved(CTc)/LOM(c)".

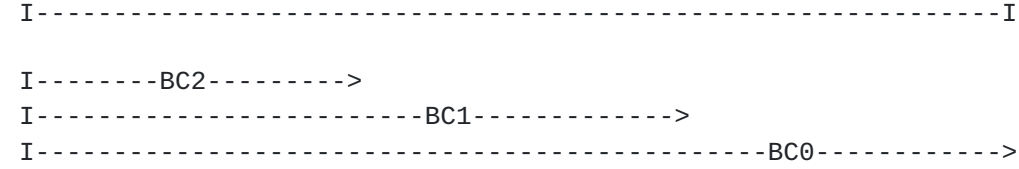
When the optional Local Overbooking method is supported, the "Russian Doll" model definition MUST be extended in the following manner:

- Maximum Number of Bandwidth Constraints (MaxBC)= Maximum Number of Class-Types (MaxCT) = 8
- SUM [Normalised(CTc), for b<=c<=7] <= BCb i.e.:
  - o [Normalised(CT7)] <= BC7
  - o [Normalised(CT6) + Normalised(CT7)] <= BC6
  - o [Normalised(CT5) + Normalised(CT6) + Normalised(CT7)] <= BC5
  - o etc.
  - o [Normalised(CT0) + Normalised(CT1) + ... Normalised(CT6) + Normalised(CT7)] <= BC0

Purely for illustration purposes, the diagram below represents the

Russian Doll Bandwidth Constraints model in a pictorial manner when 3 Class-Types are active and the local overbooking method is used:





9.4.2. Bandwidth Constraints

When the optional Local Overbooking method is supported, the relationship among the Bandwidth Constraints values to be enforced by the LSR is not modified. The LSR MUST still ensure that BCi is configured smaller or equal to BCj, where i is greater than j.

When the Bandwidth Constraints are optionally advertised in the Bandwidth Constraints sub-TLV, the values advertised are the BC values as configured (i.e. LOMs do not affect the advertisement of BCs).

9.4.3. Computing "Unreserved TE-Class [i]"

As discussed in section 9.2, details on formulas for computing the unreserved bandwidth, are outside the scope of the current IETF work. However, when the Local Overbooking method is supported, the advertised unreserved bandwidth values MUST reflect the per-CT Local Overbooking Multipliers. This MUST be consistent with how the LOMs are taken into account in the specific local admission control algorithm.

Note that this may result in the Unreserved Bandwidth values advertised for a particular CT being larger than one or more of the BCs relevant to this CT (since BCs are not affected by LOMs).

Example formulas for computing "Unreserved TE-Class [i]" when local overbooking is supported are provided in section 2 of Appendix C.

9.4.4. Max Link Bandwidth

The Max Link bandwidth parameter effectively controls the maximum size of a single LSP. The value advertised in the Max Link Bandwidth sub-TLV is the value as configured locally (i.e. LOMs do not affect the advertisement of the Max Link Bandwidth).

9.4.5. Admission Control Rules

When the optional Local Overbooking method is supported, the DS-TE LSR MUST support the same admission control rules as without LOM:

Regardless of how the admission control algorithm actually computes the unreserved bandwidth for TE-Class[i] for one of its local link,

an LSP of bandwidth B, of set-up preemption priority p and of Class-Type CTc is admissible on that link iff:

(i) B <= unreserved bandwidth for TE-Class[i], AND
(ii) B <= Max Link Bandwidth

Where

- TE-Class [i] maps to < CTc , p > in the LSR's configured TE-Class mapping
- Max Link Bandwidth is the maximum link bandwidth configured on the link and advertised in IGP.

Condition (i) above applies to B [and not B/LOM(c)] because the LOM is already factored in the unreserved bandwidth values.

Condition (ii) above also applies to B [and not B/LOM(c)]. Condition (ii) controls the maximum bandwidth that a single LSP can grab on its

own, by limiting it to the Maximum Link Bandwidth. Since "overbooking" is usually only (fully) achievable when a sufficient number of LSPs share the Link Bandwidth, it is not appropriate to factor in the LOM when assessing the maximum bandwidth that a single LSP can grab on a link.

9.4.6. Example Usage of LOM

To illustrate usage of the local overbooking method, let's consider a DS-TE deployment where two CTs (CT0 for data and CT1 for voice) and a single preemption priority are used.

The TE-Class mapping is the following:

TE-Class	<-->	CT, preemption
0		CT0, 0
1		CT1, 0
rest		unused

Let's assume that on a given link, BCs and LOMs are configured in the following way:

BC0 = 200
BC1 = 100
LOM(0) = 4 (i.e. = 400%)
LOM(1) = 2 (i.e. = 200%)

Let's further assume that the DS-TE LSR uses the example formulas presented in section 2 of Appendix C for computing unreserved bandwidth values.

If there is no established LSP on the considered link, the LSR will advertise for that link in IGP :

Unreserved TE-Class [0] = 4 x (200 - 0/4 - 0/2)= 800
Unreserved TE-Class [1] = 2 x (100- 0/2) = 200

Note again that these values advertised for Unreserved Bandwidth are larger than BC1 and BC0.

If there is only a single established LSP, with CT=CT0 and BW=100, the LSR will advertise:

Unreserved TE-Class [0] = 4 x (200 - 100/4 - 0/2)=700
Unreserved TE-Class [1] = 2 x (100- 0/2) = 200

If there is only a single established LSP, with CT=CT1 and BW=100, the LSR will advertise:

Unreserved TE-Class [0] = 4 x (200 - 0/4 - 100/2)= 600
Unreserved TE-Class [1] = 2 x (100- 100/2) = 100

Note that the impact of an LSP on the unreserved bandwidth of a CT does not depend only on the LOM for that CT: it also depends on the LOM for the CT of the LSP. This can be seen in our example. A BW=100 tunnel affects Unreserved
CT0 twice as much if it is a CT1 tunnel, than if it is a CT0 tunnel. It reduces Unreserved CTO by 200 (800->600) rather than by 100 (800->700). This is because LOM(1) is half as big as LOM(0). This illustrates why the local overbooking method allows very fine accounting of cross-effect of overbooking across CTs, as compared with the LSP/link size overbooking method.

If there are two established LSPs, one with CT=CT1 and BW=100 and one with CT=CT0 and BW=100, the LSR will advertise:

Unreserved TE-Class [0] = 4 x (200 - 100/4 - 100/2) = 500
Unreserved TE-Class [1] = 2 x (100 - 100/2) = 100

If there are two LSPs established, one with CT=CT1 and BW=100, and one with CT=CT0 and BW=480, the LSR will advertise:

```
Unreserved TE-Class [0] = 4 x (200 - 480/4 - 100/2) = 120
Unreserved TE-Class [1] = 2 x MIN [ (200 - 480/4 - 100/2),
                                    (100 - 100/2) ]
                        = 2 x MIN [ 30, 50 ]
                        = 60
```

10. Security Considerations

The solution is not expected to add specific security requirements beyond those of Diff-Serv and existing TE. The security mechanisms currently used with Diff-Serv and existing TE can be used with this solution.

11. Acknowledgments

We thank Martin Tatham, Angela Chiu and Pete Hicks for their earlier contribution in this work. We also thank Sanjaya Choudhury for his thorough reviews and numerous suggestions.

References

[DSTE-REQ] Le Faucheur et al, Requirements for support of Diff-Serv-aware MPLS Traffic Engineering, draft-ietf-tewg-diff-te-reqts-05.txt, June 2002.

[OSPF-TE] Katz, Yeung, Traffic Engineering Extensions to OSPF, draft-katz-yeung-ospf-traffic-06.txt, October 2001.

[ISIS-TE] Smit, Li, IS-IS extensions for Traffic Engineering, draft-ietf-isis-traffic-04.txt, August 2001.

[RSVP-TE] Awduche et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[CR-LDP] Jamoussi et al, "Constraint-Based LSP Setup using LDP", RFC 3212, January 2002.

[DIFF-MPLS] Le Faucheur et al, "MPLS Support of Diff-Serv", RFC3270, May 2002.

[RUSSIAN] Le Faucheur, "Considerations on Bandwidth Constraints Model for DS-TE", draft-lefaucheur-tewg-russian-dolls-00.txt, June 2002.

[BCMODEL] Lai, "Bandwidth Constraints Models for DS-TE", draft-wlai-tewg-bcmodel-00.txt, June 2002.

Authors' Address:

Francois Le Faucheur
Cisco Systems, Inc.
Village d'Entreprise Green Side - Batiment T3
400, Avenue de Roumanille
06410 Biot-Sophia Antipolis
France
Phone: +33 4 97 23 26 19
Email: flefauch@cisco.com

Jim Boyle
Protocol Driven Networks
1381 Kildaire Farm Road #288

Cary, NC 27511
Phone: +1 919 852-5160
Email: jboyle@pdnets.com

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94099
Email: kireeti@juniper.net

William Townsend
Tenor Networks
100 Nagog Park
Acton, MA 01720
Phone: +1-978-264-4900
Email: btownsend@tenornetworks.com

Thomas D. Nadeau
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA 01824
Phone: +1-978-244-3051
Email: tnadeau@cisco.com

Darek Skalecki
Nortel Networks
3500 Carling Ave,
Nepean K2H 8E9
Phone: +1-613-765-2252
Email: dareks@nortelnetworks.com

Appendix A - RSVP Extensions for Diff-Serv-aware TE

In this section we describe extensions to RSVP for support of Diff-Serv-aware MPLS Traffic Engineering. These extensions are in addition to the extensions to RSVP defined in [RSVP-TE] for support of (aggregate) MPLS Traffic Engineering and to the extensions to RSVP

defined in [DIFF-MPLS] for support of Diff-Serv over MPLS.

1. Diff-Serv-aware TE related RSVP Messages Format

One new RSVP Object is defined in this document: the CLASSTYPE Object. Detailed description of this Object is provided below. This new Object is applicable to Path messages. This specification only defines the use of the CLASSTYPE Object in Path messages used to establish LSP Tunnels in accordance with [RSVP-TE] and thus containing a Session Object with a C-Type equal to LSP_TUNNEL_IPv4 and containing a LABEL_REQUEST object.

Restrictions defined in [RSVP-TE] for support of establishment of LSP Tunnels via RSVP are also applicable to the establishment of LSP Tunnels supporting Diff-Serv-aware Traffic Engineering. For instance, only unicast LSPs are supported and Multicast LSPs are for further study.

This new CLASSTYPE object is optional with respect to RSVP so that general RSVP implementations not concerned with MPLS LSP set up do not have to support this object.

An LSR supporting Diff-Serv-aware Traffic Engineering in compliance with this specification MUST support the CLASSTYPE Object.

1.1. Path Message Format

The format of the Path message is as follows:

```
<Path Message> ::=      <Common Header> [ <INTEGRITY> ]
                         <SESSION> <RSVP_HOP>
                         <TIME_VALUES>
                         [ <EXPLICIT_ROUTE> ]
                         <LABEL_REQUEST>
                         [ <SESSION_ATTRIBUTE> ]
                         [ <DIFFSERV> ]
                         [ <CLASSTYPE> ]
                         [ <POLICY_DATA> ... ]
```

```
                          [ <sender descriptor> ]

  <sender descriptor> ::=  <SENDER_TEMPLATE> [ <SENDER_TSPEC> ]
                           [ <ADSPEC> ]
                           [ <RECORD_ROUTE> ]
```

2. CLASSTYPE Object

The CLASSTYPE object format is shown below.

2.1. CLASSTYPE object

class = TBD, C_Type = 1 (need to get an official class num from the IANA with the form 0bbbbbbb)

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Reserved                                               |  CT |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Reserved : 29 bits
This field is reserved. It must be set to zero on transmission and must be ignored on receipt.

CT : 3 bits
Indicates the Class-Type. Values currently allowed are 1, 2, ... , 7.

3. Handling CLASSTYPE Object

To establish an LSP tunnel with RSVP, the sender LSR creates a Path message with a session type of LSP_Tunnel_IPv4 and with a LABEL_REQUEST object as per [RSVP-TE]. The sender LSR may also include the DIFFSERV object as per [DIFF-MPLS].

If the LSP is associated with Class-Type 0, the sender LSR MUST NOT

include the CLASSTYPE object in the Path message.

If the LSP is associated with Class-Type N (1 <= N <=7), the sender LSR MUST include the CLASSTYPE object in the Path message with the Class-Type (CT) field set to N.

If a path message contains multiple CLASSTYPE objects, only the first one is meaningful; subsequent CLASSTYPE object(s) MUST be ignored and MUST not be forwarded.

Each LSR along the path MUST record the CLASSTYPE object, when present, in its path state block.

If the CLASSTYPE object is not present in the Path message, the LSR MUST associate the Class-Type 0 to the LSP.

The destination LSR responding to the Path message by sending a Resv message MUST NOT include a CLASSTYPE object in the Resv message (whether the Path message contained a CLASSTYPE object or not).

During establishment of an LSP corresponding to the Class-Type N, the LSR MUST perform admission control over the bandwidth available for that particular Class-Type.

An LSR that recognizes the CLASSTYPE object and that receives a path message which contains the CLASSTYPE object but which does not contain a LABEL_REQUEST object or which does not have a session type of LSP_Tunnel_IPv4, MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'Unexpected CLASSTYPE object'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object, which recognizes the CLASSTYPE object but does not support the particular Class-Type, MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'Unsupported Class-Type'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object, which recognizes the CLASSTYPE object but determines that the Class-Type value is not valid (i.e. Class-Type value 0), MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'Invalid Class-Type value'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object, which:

- recognizes the CLASSTYPE object,
- supports the particular Class-Type, but
- determines that the tuple formed by (i) this Class-Type and (ii) the set-up priority signaled in the same Path message, is not one of the eight TE-classes configured in the TE-class mapping,

MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'CT and setup priority do not form a configured TE-Class'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object, which:
- recognizes the CLASSTYPE object,
- supports the particular Class-Type, but
- determines that the tuple formed by (i) this Class-Type and (ii) the holding priority signaled in the same Path message, is not one of the eight TE-classes configured in the TE-class mapping,

MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'CT and holding priority do not form a configured TE-Class'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object and with the DIFFSERV object for an L-LSP, which:
- recognizes the CLASSTYPE object,
- has local knowledge of the relationship between Class-Types and PSC (e.g. via configuration)
- based on this local knowledge, determines that the PSC signaled in the DIFFSERV object is inconsistent with the Class-Type signaled in the CLASSTYPE object,

MUST send a PathErr towards the sender with the error code 'Diff-Serv-aware TE Error' and an error value of 'Inconsistency between signaled PSC and signaled CT'. Those are defined below in section 5.

An LSR receiving a Path message with the CLASSTYPE object and with the DIFFSERV object for an E-LSP, which:
- recognizes the CLASSTYPE object,
- has local knowledge of the relationship between Class-Types and PHBs (e.g. via configuration)
- based on this local knowledge, determines that the PHBs signaled in the MAP entries of the DIFFSERV object are inconsistent with the Class-Type signaled in the CLASSTYPE object,

MUST send a PathErr towards the sender with the error code 'Diff-

Serv-aware TE Error' and an error value of 'Inconsistency between signaled PHBs and signaled CT'. Those are defined below in section 5.

An LSR MUST handle the situations where the LSP can not be accepted for other reasons than those already discussed in this section, in accordance with [RSVP-TE] and [DIFF-MPLS] (e.g. a reservation is rejected by admission control, a label can not be associated).

4. Non-support of the CLASSTYPE Object

An LSR that does not recognize the CLASSTYPE object Class-Num MUST behave in accordance with the procedures specified in [RSVP] for an unknown Class-Num whose format is 0bbbbbbb (i.e. it must send a

PathErr with the error code 'Unknown object class' toward the sender).

An LSR that recognizes the CLASSTYPE object Class-Num but does not recognize the CLASSTYPE object C-Type, MUST behave in accordance with the procedures specified in [RSVP] for an unknown C-type (i.e. it must send a PathErr with the error code 'Unknown object C-Type' toward the sender).

In both situations, this causes the path set-up to fail. The sender SHOULD notify management that a LSP cannot be established and possibly might take action to retry reservation establishment without the CLASSTYPE object.

5. Error Codes For Diff-Serv-aware TE

In the procedures described above, certain errors must be reported as a 'Diff-Serv-aware TE Error'. The value of the 'Diff-Serv-aware TE Error' error code is (TBD).

The following defines error values for the Diff-Serv-aware TE Error:

Value	Error
1	Unexpected CLASSTYPE object
2	Unsupported Class-Type

3	Invalid Class-Type value
4	CT and setup priority do not form a configured TE-Class
5	CT and holding priority do not form a configured TE-Class
6	Inconsistency between signaled PSC and signaled CT
7	Inconsistency between signaled PHBs and signaled CT

Appendix B - CR-LDP Extensions for Diff-Serv-aware TE

CR-LDP, defined in [CR-LDP], is an extension to LDP, defined in [LDP], for support of (aggregate) MPLS Traffic Engineering. In this section we describe extensions to CR-LDP for support of Diff-Serv-aware MPLS Traffic Engineering. These extensions are in addition to the extensions to LDP defined in [DIFF-MPLS] for support of Diff-Serv over MPLS. They closely resemble the extensions to RSVP defined in the previous section.

Note that extensions of this section for support of Diff-Serv-aware Traffic Engineering are not applicable to LDP due to the fact that LDP does not support MPLS Traffic Engineering and bandwidth reservation in particular.

1. Diff-Serv-aware TE related CR-LDP Messages Encoding

One new CR-LDP TLV is defined in this document: the Class Type TLV.

Detailed description of this TLV is provided below. This new TLV is applicable to Label Request messages.

Restrictions defined in [CR-LDP] for support of establishment of LSPs via CR-LDP are also applicable to the establishment of LSPs supporting Diff-Serv-aware Traffic Engineering: for instance, only unicast LSPs are supported and multicast LSPs are for further study.

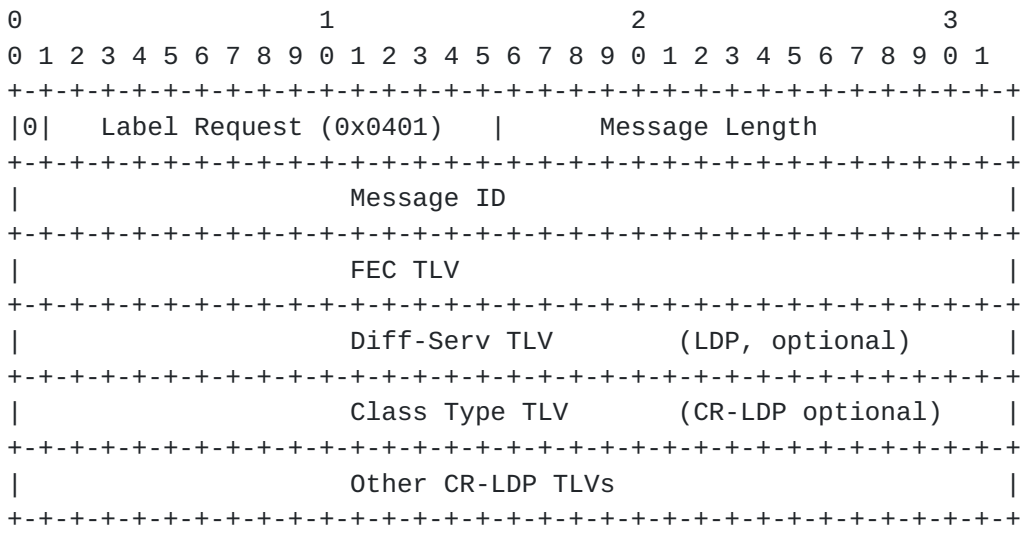
This new Class Type TLV is optional with respect to CR-LDP so that general CR-LDP implementations not concerned with Diff-Serv-aware Traffic Engineering are not required to support this TLV.

An LSR supporting Diff-Serv-aware Traffic Engineering in compliance

with this specification MUST support the Class Type TLV.

1.1. Label Request Message Encoding

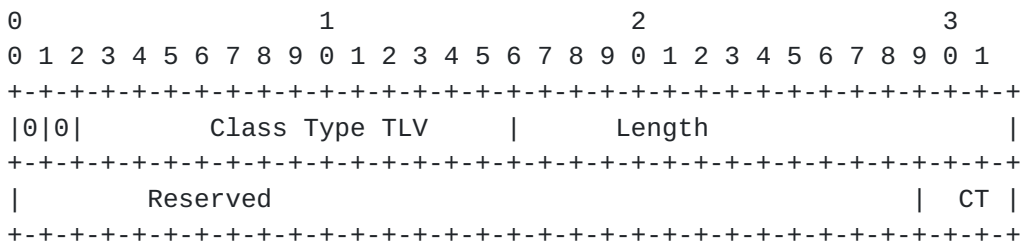
The encoding for the CR-LDP Label Request message is extended as follows, to optionally include the Class Type TLV:



The extension is based on a related LDP extension, defined in [DIFF-MPLS], for support of Diff-Serv TLV but further extended for CR-LDP with CR-LDP TLVs.

2. Class Type TLV

The Class Type TLV has the following form:



Reserved : 29 bits
This field is reserved. It must be set to zero on transmission and must be ignored on receipt.

CT : 3 bits
Indicates the Class-Type. Values currently allowed are 1, 2, ... , 7.

3. Handling Class Type TLV

To establish an LSP using CR-LDP, an ingress LSR generates a Label Request message as per [CR-LDP]. This Label Request may optionally include the Diff-Serv TLV as defined in [DIFF-MPLS] for LDP but extended to CR-LDP.

If the LSP is associated with Class-Type 0, the ingress LSR MUST NOT include the Class Type TLV in the Label Request message.

If the LSP is associated with Class-Type N (1 <= N <= 7), the ingress LSR MUST include the Class Type TLV in the Label Request message with the Class-Type (CT) field set to N.

If a Label Request message contains multiple Class Type TLVs, only the first one is meaningful; subsequent Class Type TLV(s) MUST be ignored and not forwarded.

If the Class Type TLV is not present in the Label Request message, an LSR MUST associate the Class-Type 0 to the LSP.

A downstream LSR sending a Label Mapping message in response to a Label Request message MUST NOT include the Class-Type TLV (whether the Class-Type TLV was included in the Label Request message or not).

During establishment of an LSP corresponding to the Class-Type N, an LSR MUST perform admission control over the bandwidth available for that particular Class-Type.

An LSR that recognizes the Class Type TLV and receives a Label Request message which contains the Class Type TLV but which does not contain any of the CR-LDP TLVs, MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'Unexpected Class-Type TLV'. This is defined below in section 4. This error can only occur when an LDP LSP as opposed to CR-LDP LSP is being established. As was already mentioned, Class Type TLV extension for Diff-Serv-aware Traffic Engineering is not applicable to LDP.

An LSR receiving a Label Request message with the Class Type TLV, which recognizes the Class Type TLV but does not support the particular Class-Type, MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a

Status Code of 'Unsupported Class-Type'. This is defined below in section 4.

An LSR receiving a Label Request message with the Class Type TLV, which recognizes the Class Type TLV but determines that the Class-Type value is not valid (i.e. Class-Type value 0), MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'Invalid Class-Type value'. This is defined below in section 4.

An LSR receiving a Label Request message with the Class Type TLV, which:

- recognizes the Class Type TLV,
- supports the particular Class-Type, but
- determines that the tuple formed by (i) this Class-Type and (ii) the set-up priority signaled in the same Label Request message, is not one of the eight TE-classes configured in the TE-class mapping,

MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'CT and setup priority do not form a configured TE-Class'. This is defined below in section 4.

An LSR receiving a Label Request message with the Class Type TLV, which:

- recognizes the Class Type TLV,
- supports the particular Class-Type, but
- determines that the tuple formed by (i) this Class-Type and (ii) the holding priority signaled in the same Label Request message, is not one of the eight TE-classes configured in the TE-class mapping,

MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'CT and holding priority do not form a configured TE-Class'. This is defined below in section 4.

An LSR receiving a Label Request message with the Class Type TLV and

with the Diff-Serv TLV for an L-LSP, which:

- recognizes the Class Type TLV,
- has local knowledge of the relationship between Class-Types and PSC (e.g. via configuration)
- based on this local knowledge, determines that the PSC signaled in the Diff-Serv TLV is inconsistent with the Class-Type signaled in the Class-Type TLV,

MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'Inconsistency between signaled PSC and signaled CT'. This is defined below in section 4.

An LSR receiving a Label Request message with the Class Type TLV and with the Diff-Serv TLV for an E-LSP, which:

- recognizes the Class Type TLV,

- has local knowledge of the relationship between Class-Types and PHBs (e.g. via configuration)
- based on this local knowledge, determines that the PHBs signaled in the MAP entries of the Diff-Serv TLV are inconsistent with the Class-Type signaled in the Class-Type TLV,

MUST reject the label request by sending upstream a Notification message which includes the Status TLV with a Status Code of 'Inconsistency between signaled PHBs and signaled CT'. This is defined below in section 4.

An LSR MUST handle the situations where the LSP can not be accepted for other reasons than those already discussed in this section, in accordance with [CR-LDP], [LDP] and [DIFF-MPLS] (e.g. reservation rejected by admission control, a label can not be associated).

4. Status Code Values for Diff-Serv-aware TE

In the procedures described above, certain errors must be reported. The following values are defined for the Status Code field of the Status TLV:

Status Code	E	Status Data

Unexpected Class Type TLV	0	TBD
Unsupported Class-Type	0	TBD
Invalid Class-Type value	0	TBD
CT and setup priority do not form a configured TE-Class	0	TBD
CT and holding priority do not form a configured TE-Class'	0	TBD
Inconsistency between signaled PSC and signaled CT	0	TBD
Inconsistency between signaled PHBs and signaled CT	0	TBD

Appendix C - Example Formulas for Computing "Unreserved TE-Class [i]"

1. Example Formulas Without Local Overbooking

Keeping in mind that details of admission control algorithms as well as formulas for computing "Unreserved TE-Class [i]" are outside the scope of this specification, we provide in this section, for illustration purposes, an example of how values for the unreserved bandwidth for TE-Class[i] might be computed, assuming:

- the Russian Doll Bandwidth Constraints Model is used
- the basic admission control algorithm which simply deducts the exact bandwidth of any established LSP from all of the Bandwidth Constraints relevant to the CT associated with that LSP.

- the optional per-CT Local Overbooking Multipliers are not used (.i.e. LOM[c]=1, 0<= c <=7).

We assume that:

TE-Class [i] <--> < CTc , preemption p>

in the configured TE-Class mapping.

Let us define "Reserved(CTb,q)" as the sum of the bandwidth reserved by all established LSPs which belong to CTb and have a holding

priority of q. Note that if q and CTb do not form one of the 8 possible configured TE-Classes, then there can not be any established LSP which belong to CTb and have a holding priority of q, so in that case Reserved(CTb,q)=0.

For readability, formulas are first shown assuming only 4 CTs are active. The formulas below can be extended trivially to cover the cases where more CTs are used.

```
If CTc = CT0, then "Unreserved TE-Class [i]" =
     [ BC0 - SUM ( Reserved(CTb,q) ) ] for q <= p and 0 <= b <= 3


If CTc = CT1, then "Unreserved TE-Class [i]" =
     MIN  [
     [ BC1 - SUM ( Reserved(CTb,q) ) ] for q <= p and 1 <= b <= 3,
     [ BC0 - SUM ( Reserved(CTb,q) ) ] for q <= p and 0 <= b <= 3
          ]


If CTc = CT2, then "Unreserved TE-Class [i]" =
     MIN  [
     [ BC2 - SUM ( Reserved(CTb,q) ) ] for q <= p and 2 <= b <= 3,
     [ BC1 - SUM ( Reserved(CTb,q) ) ] for q <= p and 1 <= b <= 3,
     [ BC0 - SUM ( Reserved(CTb,q) ) ] for q <= p and 0 <= b <= 3
          ]


If CTc = CT3, then "Unreserved TE-Class [i]" =
     MIN  [
     [ BC3 - SUM ( Reserved(CTb,q) ) ] for q <= p and 3 <= b <= 3,
     [ BC2 - SUM ( Reserved(CTb,q) ) ] for q <= p and 2 <= b <= 3,
     [ BC1 - SUM ( Reserved(CTb,q) ) ] for q <= p and 1 <= b <= 3,
     [ BC0 - SUM ( Reserved(CTb,q) ) ] for q <= p and 0 <= b <= 3
          ]
```

The formula can be generalized to 8 active CTs and expressed in a more compact way in the following:

```
      "Unreserved TE-Class [i]" =
     MIN  [
     [ BCc - SUM ( Reserved(CTb,q) ) ] for q <= p and c <= b <= 7,
```

```
     . . .
     [ BC0 - SUM ( Reserved(CTb,q) ) ] for q <= p and 0 <= b <= 7,
         ]
```

where:

TE-Class [i] <--> < CTc , preemption p>
in the configured TE-Class mapping.

2. Example Formulas With Local Overbooking

When the optional local overbooking method is supported, the above example generalized formula becomes:

```
      "Unreserved TE-Class [i]" =
     LOM(c) x MIN  [
     [ BCc - SUM ( Normalised(CTb,q) ) ] for q <= p and c <= b <= 7,
     . . .
     [ BC0 - SUM ( Normalised(CTb,q) ) ] for q <= p and 0 <= b <= 7,
         ]
```

where:

- TE-Class [i] <--> < CTc , preemption p>
  in the configured TE-Class mapping.
- Normalised(CTb,q) = Reserved(CTb/q)/LOM(b)

Appendix D - Prediction for Multiple Path Computation

There are situations where a Head-End needs to compute paths for multiple LSPs. There are potential advantages for the Head-end in trying to predict the impact of the n-th LSP on the unreserved bandwidth when computing the path for the (n+1)-th LSP, before receiving updated IGP information. One example would be to perform better load-distribution of the multiple LSPs across multiple paths. Another example would be to avoid CAC rejection when the (n+1)-th LSP would no longer fit on a link after establishment of the n-th LSP. While there are also a number of conceivable scenarios where doing such predictions might result in a worse situation, it is more likely to improve the situation. As a matter of fact, a number of network administrators have elected to use such predictions when deploying existing TE.

Such predictions are local matters, are optional and are outside the scope of this specification.

Where such predictions are not used, the optional Bandwidth Constraint sub-TLV and the optional Local Overbooking Multiplier sub-

TLV need not be advertised in IGP since the information contained in the Unreserved Bw sub-TLV is all that is required by Head-Ends to perform Constraint Based Routing.

Where such predictions are used on Head-Ends, the optional Bandwidth Constraint sub-TLV (and the optional Local Overbooking Multiplier sub-TLV if different overbooking ratios need to be supported on different links) MAY be advertised in IGP. This is in order for the Head-ends to predict as accurately as possible how an LSP affects unreserved bandwidth values for subsequent LSPs.

Remembering that actual admission control algorithms are left for vendor differentiation, we observe that predictions can only be performed effectively when the Head-end LSR predictions are based on the same (or a very close) admission control algorithm as used by other LSRs.

Appendix E - Addressing [DSTE-REQ] Scenarios

This Appendix provides examples of how the DS-TE solution can be used to support each of the scenario described in [DSTE-REQ].

1. Scenario 1: Limiting Amount of Voice

By configuring on every link:

- Bandwidth Constraint 1 (for CT1=Voice) = "certain percentage" of link capacity
- BC0= Max Reservable Link Bandwidth = link capacity

By configuring:

- every CT1/Voice TE-LSP with preemption =0
- every CT0/Data TE-LSP with preemption =1

The proposed solution will address all the requirements:

- amount of Voice traffic limited to desired percentage on every link
- data traffic capable of using all remaining link capacity

- voice traffic capable of preempting other traffic

2. Scenario 2: Maintain Relative Proportion of Traffic Classes

By configuring on every link:
- BC2 for CT2 = e.g. 45%
- BC1 for CT1+CT2 = e.g. 80%
- BC0 for CT0+CT1+CT2= e.g.100%

The proposed DS-TE solution will ensure that the amount of traffic of each Class Type established on a link is within acceptable levels as compared to the resources allocated to the corresponding Diff-Serv PHBs regardless of which order the LSPs are routed in, regardless of which preemption priorities are used by which LSPs and regardless of failure situations. Optional automatic adjustment of Diff-Sev scheduling configuration could be used for maintaining very strict relationship between amount of established traffic of each Class Type and corresponding Diff-Serv resources.

3. Scenario 3: Guaranteed Bandwidth Services

By configuring on every link:
- BC1 for CT1 = "given" percentage of bandwidth (appropriate to achieve the Guaranteed Bandwidth service's QoS objectives)
- BC0 for CT0+CT1 = 100%

The proposed DS-TE solution will ensure that the amount of Guaranteed Bandwidth Trafic established on every link remains below the given percentage so that it will always meet its QoS objectives. AT the same time it will allow traffic engineering of the rest of the traffic such that links can be filled up.

Appendix F - Solution Evaluation

1. Satisfying Detailed Requirements

This DS-TE Solution address all the scenarios presented in [DSTE-REQ] as explained in Appendix E. It also satisfy all the detailed

requirements presented in [DSTE-REQ].

2. Flexibility

This DS-TE solution supports 8 CTs. It is entirely flexible as to how Traffic Trunks are grouped together into a CT.

3. Extendibility

A maximum of 8 CTs is considered by the authors of this document as more than comfortable. However, this solution could be extended to support more CTs if deemed necessary in the future. However, this would necessitate additional IGP extensions beyond those specified in this document.

4. Scalability

This DS-TE solution is expected to have a very small scalability impact compared to existing TE.

From an IGP viewpoint, the amount of mandatory information to be advertised is identical to existing TE. Two additional sub-TLVs have been specified, but their use is optional and those contained a limited amount of static information (at most 8 Bandwidth Constraints and 8 LOMs).

We expect no noticeable impact on LSP Path computation since, as with existing TE, this solution only require CSPF to consider a single unreserved bandwidth value for any given LSP.

From a signaling viewpoint we expect no significant impact due to this solution since it only requires processing of one additional information (the Class-Type) and does not significantly increase the likelihood of CAC rejection. Note that DS-TE has some inherent impact on LSP signaling in the sense that it assumes that different classes of traffic are split over different LSPs so that more LSPs need to be signaled; but this is due to the DS-TE concept itself and not to the actual DS-TE solution discussed here.

5. Backward Compatibility/Migration

This solution is expected to allow smooth migration from existing TE to DS-TE. This is because existing TE can be supported exactly as today as a particular configuration of DS-TE. This means that an "upgraded" LSR with a DS-TE implementation can directly interwork with an "old" LSR supporting existing TE only.

This solution is expected to allow smooth migration when increasing the number of CTs actually deployed since it only requires configuration changes. however, these changes must be performed in a coordinated manner across the DS-TE domain.

Appendix G - Interoperability with non DS-TE capable LSRs

This DSTE solution allows operations in a hybrid network where some LSRs are DS-TE capable while some LSRs and not DS-TE capable, which may occur during migration phases. This Appendix discusses the constraints and operations in such hybrid networks.

We refer to the set of DS-TE capable LSRs as the DS-TE domain. We refer to the set of non DS-TE capable (but TE capable) LSRs as the TE-domain.

Hybrid operations requires that the TE-class mapping in the DS-TE domain is configured so that:

- a TE-class exist for CT0 for every preemption priority actually used in the TE domain
- the index in the TE-class mapping for each of these TE-classes is equal to the preemption priority.

For example, imagine the TE domain uses preemption 2 and 3. Then, DS-TE can be deployed in the same network by including the following TE-classes in the TE-class mapping:

i <--->	CT	preemption
2	CT0	2
3	CT0	3

Another way to look at this is to say that, the whole TE-class mapping does not have to be consistent with the TE domain, but the

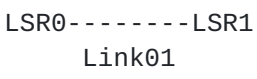
subset of this TE-Class mapping applicable to CT0 must effectively be consistent with the TE domain.

In such hybrid networks :

- CT0 LSPs can be established by both DS-TE capable LSRs and non-DSTE capable LSRs
- CT0 LSPs can transit via (or terminate at) both DS-TE capable LSRs and non-DSTE capable LSRs
- LSPs from other CTs can only be established by DS-TE capable LSRs
- LSPs from other CTs can only transit via (or terminate at) DS-TE capable LSRs

Note that, for such hybrid operation, non DS-TE capable LSRs need to be able to accept Unreserved Bw sub-TLVs containing non decreasing bandwidth values (ie with Unreserved [p] < Unreserved [q] with p <q). Also, the Bandwidth Constraints sub-TLV needs to be advertised in the DS-TE domain and the Maximum Reservable Bandwidth sub-TLV needs to be advertised in the TE-domain and the DS-TE domain.

Let us consider, the following example to illustrate operations:



Where:

LSR0 is a non-DS-TE capable LSR
LSR1 and LSR2 are DS-TE capable LSRs

Let's assume again that preemption 2 and 3 are used in the TE-domain and that the following TE-class mapping is configured on LSR1 and LSR2:

i <--->	CT	preemption
0	CT1	0
1	CT1	1
2	CT0	2
3	CT0	3
rest	unused	

LSR0 is configured with a Max Reservable bandwidth=m for Link01.
LSR1 is configured with a BC0=x0 and a BC1=x1(possibly=0) for Link01.

LSR0 will advertise in IGP for Link01:

- Max Reservable Bw sub-TLV = <m>
- Unreserved Bw sub-TLV =

<CT0/0,CT0/1,CT0/2,CT0/3,CT0/4,CT0/5,CT0/6,CT0/7>

On receipt of such advertisement, LSR1 will:

- understand that LSR0 is not DS-TE capable because it advertised a Max Reservable Bw sub-TLV and no Bandwidth Constraint sub-TLV

- conclude that only CT0 LSPs can transit via LSR0 and that only the values CT0/2 and CT0/3 are meaningful in the Unreserved Bw sub-TLV. LSR1 may effectively behave as if the six other values contained in the Unreserved Bw sub-TLV were set to zero.

LSR1 will advertise in IGP for Link01:

- Max Reservable Bw sub-TLV = <x0>
- Bandwidth Constraint sub-TLV = <BC Model ID=0, x0,x1>
- Unreserved Bw sub-TLV = <CT1/0,CT1/1,CT0/2,CT0/3,0,0,0,0>

On receipt of such advertisement, LSR0 will:

- Ignore the Bandwidth Constraint sub-TLV (unrecognized)
- Correctly process CT0/2 and CT0/3 in the Unreserved Bw sub-TLV and use these values for CTO LSP establishment
- Incorrectly believe that the other values contained in the Unreserved Bw sub-TLV relates to other preemption priorities for CT0, but will actually never use those since we assume that only preemption 2 and 3 are used in the TE domain.