Internet Engineering Task Force Jim Boyle
Internet Draft
draft-boyle-tewg-ds-nop-00.txt
July 4, 2001
Expires: December, 2001

Accomplishing Diffserv TE needs with Current Specifications

STATUS OF THIS MEMO

Abstract

The emerging requirements for differentiated service control in MPLS based traffic engineered networks [DSTE-REQ] can be met by existing specifications. Current implementations could be adapted with a relaxed perception of semantics on the syntax outlined in current specifications [OSPF-TE][ISIS-TE][RSVP-TE]. This would provide for a solution that fulfills the scenarios and functionality requirements called for in DSTE-REQ.

Comments should be sent to tewg@ops.ietf.org

1. Introduction

A common theme in the scenarios spelled out in section 2 of DSTE-REQ is that current approaches to traffic engineering do not allow an operator to properly control either to what extent a class of traffic can use a link, or what minimal amount of traffic should be held in reserve for a "lower" class of traffic to prevent total starvation. This memo asserts that this is more a function of current traffic engineering implementations commonly in use today

than it is of the current specifications governing the extensions of the base protocols. Further, by rethinking the semantics possible with the current specifications, it would be not only possible to meet the requirements in DSTE-REQ, but it would also be the most operationally optimal approach to meeting those requirements.

The word "priority" was perhaps unfortunate, as it carries a rather strong perception by most operators of the ability of one type of traffic to preempt another. A more generalized term might have been desirable, and general approaches have proven useful in the affinities attributes where intent is not inferable from the name or value seen in the signaled messages or usually in the configuration. A link is colored with one or more of 32 bits, and LSPs are told to stick with links thus marked, or avoid them, but the reasoning is not inferable. With priority, it is natural to assume that a connection at one priority should be able to preempt one of lower perceived importance. In fact, this is how at least two prominent implementations behave in effect today. Nonetheless, even with the current objects referred to as priority, and current implementations leaning heavily toward preemtability of one priority to another (in direct relationship to numeric value of priority), the current specifications allow for a semantic view which generalizes the interrelationship of bandwidths at different "priorities", as will be spelled out in this memo.

2. Altered Perceptions

2.1 Review of current specifications

The signaling specification RSVP-TE allows for use of setup and hold priority objects which allow for "the capability to preempt an established LSP tunnel under administrative policy control". However, in section 4.7.3 of RSVP-TE it clearly states that what determines the acceptance of a session is availability of bandwidth "at the priority specified in the Setup Priority". This correlates to the less semantic definitions of what is advertised in the IGP. For instance section 2.5.8 of OSPF-TE states that the unreserved bandwidth "not yet reserved at each of the eight priorities" is advertised. There is no language in either which states that the bandwidth at the highest priority (0) must be higher than that at the lowest priority (7). Nor is there language in RSVP-TE that states that in path computation if one can not find a path at the setup priority of a to-be-established LSP, it should then try lower priorities to see if it might benefit by preempting traffic. Thus, what is advertised by the IGP could easily be used to control what is available to higher priorities or to provide a limited amount of resources to lower priority sessions.

2.2 Current Basis

Most current implementations in fact do advertise bandwidths available in a manner which has a relationship of:

{Effect 1}

bw-link[n] >= bw-link[n+1] where n is the numeric priority

To the best knowledge of the public, current implementations also follow rules during setup:

{Rule 1}

topology for bw-lsp[n] is the topology consisting of all bw-link[n], not that of max(bw-link[n..7]).

And during call admission control:

{Rule 2}

succeed if (bw-lsp[n] <= bw-link[n])

The latter is likely followed by a bandwidth update procedure which propagates through bandwidths n to 7, thus causing {Effect 1} above as seen by inspection of the IGP's TE database.

This is where some modification would be required in implementations (though not in protocols or even specifications). Priorities would need to be implemented in a way that doesn't necessarily involve relationships with other priorities except as inferred by default or direct configuration.

2.3 Pseudo configuration

In the below pseudo configurations, I hope the intent is clear. I know it falls short of professional cli definition. Syntax is intended to be unix shell/man like.

[]	denotes an optional configuration
<>	a list where selection of one parameter is mandatory. For optional commands the first in the list is the default.

2.3.1 Pseudo configuration of current implementations

In most network elements that have traffic engineering capabilities, there is the ability to configure the maximum reservable bandwidth of a link, which then propagates into the unreserved bandwidths based on current state of reservations. LSP configuration usually allows a wide variety of parameters including the setup priority and the bandwidth. Some even allow

one to configure whether preemption is supported or not.

For discussion, suppose that there exist the capability to get the desired traffic into an LSP, and that these are configured as follows:

```
lsp foo priority 2 bandwidth 64k

link sonet1 bandwidth 2.5G

[mpls preempt <yes|no>]
```

2.3.1 Pseudo configuration of proposed implementations

It is probably desirable, though not necessarily necessary, to decouple the numeric priority from the class of traffic. It is necessary to allow more control on what limits are now placed on a link.

```
class voice use priority 2
class data use priority 4
[mpls preempt <yes|limited|map|no>]
[preempt map voice over data]
[class mute <list>]

lsp foo class voice bandwidth 64k

link sonet1 bandwidth 2.5G
[link sonet1 class-bandwidth voice max 1G]
[link sonet1 class-bandwidth voice max-percent 40]
[link sonet1 class-bandwidth data min 500M]
```

3. Scenario Rundown

The following scenarios are from DSTE-REQ.

3.1 Scenario 1: High Proportion of Voice

The scenario is one in which voice is a class of traffic at a high priority. The operator would like for it to use the shortest administrative path possible under the constraint that no link exceed a certain threshold (ratio wise) of voice traffic.

```
class voice use priority 2
class data use priority 4
class mute 0-1,3,5-7
mpls preempt limited

lsp foo class voice bandwidth 10M
lsp foo class data bandwidth 40M

link sonet1 bandwidth 2.5G
```

```
link sonet1 class-bandwidth voice max-percent 50
```

At initial advertisement, the link sonet1 will advertise that it has the following available:

```
[0, 0, 1.25G, 0, 2.5G, 0, 0, 0] for priorities [0..7]
```

As the voice traffic fills this link, the bandwidth advertisements will be debited from voice and data, so that should no data LSPs be established, the voice LSPs will be limited to 1.25G and the final advertisement would have been:

```
[0, 0, 0, 0, 1.25G, 0, 0, 0]
```

One might wonder what would have been the advertisements should three waves of LSPs come, in (1) 1G data, (2) another 1G data and (3) 1G voice.

```
[0, 0, 1.25G, 0, 2.5G, 0, 0, 0] initially
[0, 0, 1.25G, 0, 1.5G, 0, 0, 0] wave (1)
[0, 0, 1.25G, 0, 500M, 0, 0, 0] wave (2)
[0, 0, 250M, 0, 0, 0, 0, 0] wave (3)
```

Wave 3 requires preemption of 500 Mbs of data LSPs.

3.2 **Scenario 2:** Rerouting on Lower Speed facilities

Appears to be very similar to scenario 1, except perhaps that scenario 1 applies more to non-failure and scenario 2 is directly in context of a failure response. In that case, the discussion section 3.1 applies here as well.

3.3 **Scenario 3:** Maintain relative proportion of traffic classes

Or not. This scenario points out that many router implementations of MPLS TE don't correlate their queuing configurations to the bandwidth "reserved" on the link. The scenario states that this being the case, it might be good to attempt to keep the traffic proportions similar to the queuing proportions. In the case of the scenario, one would have the following configuration.

```
class one use priority 1
class two use priority 2
class three use priority 3
class mute 4-7
mpls preempt limited

lsp foo class one bandwidth 10M
lsp foo class two bandwidth 20M
lsp foo class three bandwidth 30M

link sonet1 bandwidth 2.5G
```

```
link sonet1 class-bandwidth one max-percent 45
link sonet1 class-bandwidth two max-percent 35
link sonet1 class-bandwidth three max-percent 20
```

Initial link advertisement would be:

```
[1125M, 875M, 500M, 0, 0, 0, 0, 0]
```

This is likely to not be the most efficient way to operate a network so it would be probable that one would notch up the efficiency by systematically overstating the link bandwidths or understating the LSP bandwidths.

This scenario hints at directly coupling the bandwidth reserved at a given priority/class to the parameters used in queuing. That's probably worth a try. A configuration might look like:

```
class one use priority 1
class two use priority 2
class three use priority 3

queues 1 by mpls-class one  # queue parameters associated with
queues 2 by mpls-class two  # reserved bandwidth
queues 3 by mpls-class three

[mpls queue by <cos|class>] # queue by cos bits, or by lsp class

lsp foo class one bandwidth 10M
lsp foo class two bandwidth 20M
lsp foo class three bandwidth 30M

link sonet1 bandwidth 2.5G
```

3.4 Scenario 4: Guaranteed Bandwidth Services

This appears to be related to the above scenarios, particularly in that it requires limiting the amount of traffic ultimately available at a given class (as in Scenario 1) and the interaction this has with queuing (scenario 3). The additional pieces include that this class is likely best served with some form of priority queuing and that traffic entering an LSP would be policed. So:

```
queues 1 by mpls-class 1
queues 1 priority

lsp foo class one bandwidth 10M police
```

4. Functionality Rundown

These correlate to the functionality requirements in DSTE-REQ section 2.

4.1 DS-TE Compatibility

If there is a network with a mix of routers where some have the legacy implementations and some the implementations outlined in this memo, there is a potential for implementation compatibility problems. Areas of concern might include the following:

- a router's inability to accept TE IGP advertisements that do not follow {Effect 1} of section 2.2 above. This is not specified in OSPF-TE nor ISIS-TE. Any "sanity checks" of this sort should not be done.

- a router which does not follow {Rule 1} above might attempt to signal an bw-lsp[n] against an insufficient bw-link[n] because it found sufficient bw-link[n+1..7] This should fail for routers which follow {Rule 2} above, in accordance with RSVP-TE.

Implementation nuances aside, there should be no compatibility issues between routers with legacy and new implementations.

The most difficult part of the transition would be that of router syntax reconfiguration. However, as the on-the-wire signaling would be indistinguishable between the two implementations, it would be possible to even migrate a small portion of a network and try out the reconfiguration and code stability. If successful, migration through the rest of the network could proceed.

This approach carries no additional information in the IGP nor are any new or additional objects required in the signaling or routing protocol. It is suspected that the allowed variability of configuration is not without any additional implementation complexity, and that single solution approaches can be coded more optimally than general approaches. So it is assumed that there would likely be some impact to stability or scalability with this approach. However, assuming some typical configurations, it is likely that coding and testing could be optimized for speed and robustness around those configurations.

That said, current implementations can not accomplish the scenarios outlined in DSTE-REQ, and this memo is an attempt to meet those in a way that is with minimal impact on coding, scalability, stability and operations.

4.2 Separate Bandwidth Constraints

This is achievable with current specifications, as shown in the above examples.

4.3 Number of Class-Types

Eight classes are currently supported, and these may be grouped together in sets, or Class-Types, as required and supported by implementations.

4.4 Number of Classes

See 4.3.

4.5 Preemption

4.5.1 Preemption Within a Class-Type

This is achievable. It is expected that implementations would allow one to follow current (original) approach which can be thought of as 8 classes in one Class-Type. Additional approaches would include original approach with limits on higher priorities as well as arbitrary associations as shown above.

4.5.2 Preemption across Class-Types

This is achievable. Also see 4.5.1, though this might be a non-requirement.

4.6 Resource Class Affinity

This appears to be a requirement for no new requirement. Current encoding specification allows for Resource Class Affinity.

4.7 Traffic Mapping

This is achievable. It is not precluded by current specifications.

4.8 Dynamic Adjustment of Diff-serv PHBs

This is achievable. This could potentially be a key approach in diffserv enabled networks taking advantage of traffic engineering technologies.

4.9 Multiple TE Metrics

No requirements are specified.

5. Acknowledgments

This idea has benefitted from private discussions with many of the usual suspects. Namely Francois, William, Don, Luca, Vijay, Kireeti and most especially Dave. Thanks.

6. References

[DSTE-REQ] draft-ietf-tewg-diff-te-reqts-01.txt

[ISIS-TE] draft-ietf-isis-traffic-03.txt

[OSPF-TE] draft-katz-yeung-ospf-traffic-05.txt

[RSVP-TE] draft-ietf-mpls-rsvp-lsp-tunnel-08.txt

7. Author's Address

Jim Boyle
email: jimpb@nc.rr.com