TSVWG   R. Geib, Ed.
Internet-Draft   Deutsche Telekom
Intended status: Informational   D. Black
Expires: January 2, 2016   EMC Corporation
                                                            July 1, 2015

Diffserv interconnection classes and practice
draft-ietf-tsvwg-diffserv-intercon-02

Abstract

This document proposes a limited set of DiffServ PHBs and codepoints to be applied at (inter)connections of two separately administered and operated networks. Many network providers operate MPLS using Treatment Aggregates for traffic marked with different DiffServ PHBs, and use MPLS for interconnection with other networks. This document offers a simple interconnection approach that may simplify operation of DiffServ for network interconnection among providers that use MPLS and apply the Short-Pipe tunnel mode.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

DiffServ has been deployed in many networks. As described by section 2.3.4.2 of RFC 2475, remarking of packets at domain boundaries is a DiffServ feature [RFC2475]. This draft proposes a set of standard QoS classes and code points at interconnection points to which and from which locally used classes and code points should be mapped.

RFC2474 specifies the DiffServ Codepoint Field [RFC2474]. Differentiated treatment is based on the specific DSCP. Once set, it may change. If traffic marked with unknown or unexpected DSCPs is received, RFC2474 recommends forwarding that traffic with default (best effort) treatment without changing the DSCP markings. Many networks do not follow this recommendation, and instead remark unknown or unexpected DSCPs to the zero DSCP upon receipt for consistency with default (best effort) forwarding in accordance with the guidance in RFC 2475 [RFC2474] to ensure that appropriate DSCPs are used within a DiffServ domain. Network providers applying the MPLS Short Pipe model are likely to remark unexpected DSCPs.

This document is motivated by requirements for IP network interconnection with DiffServ support among providers that operate MPLS in their backbones, but is applicable to other technologies. The operational simplifications and methods in this document help align IP DiffServ functionality with MPLS limitations resulting especially from the Short Pipe model of operation [RFC3270]. The latter is widely deployed. Further, limiting DiffServ to a small number of Treatment Aggregates can enable network traffic to leave a network with the same DSCPs that it was received with, even if a different DSCP is used within the network, thus providing an opportunity to extend consistent QoS treatment across network boundaries.

In isolation, use of standard interconnection PHBs and DSCPs may appear to be additional effort for a network operator. The primary offsetting benefit is that the mapping from or to the interconnection PHBs and DSCPs is specified once for all of the interconnections to other networks that can use this approach. Otherwise, the PHBs and DSCPs have to be negotiated and configured independently for each network interconnection, which has poor scaling properties. Further, consistent end-to-end QoS treatment is more likely to result when an interconnection code point scheme is used because traffic is remarked to the same PHBs at all network interconnections. This document envisions one-to-one DSCP remarking at network interconnections (not n DSCP to one DSCP remarking).

In addition to the standard interconnecting PHBs and DSCPs, interconnecting operators need to further agree on the tunneling technology used for interconnection (e.g., MPLS, if used) and control or mitigate the impacts of tunneling on reliability and MTU.

The MPLS Short Pipe tunneling model motivated this work and is its main scope. The approach proposed here may be also be applied for the Pipe tunneling model [RFC2983], [RFC3270]. The uniform model is out of scope of this document.

1.1. Related work

In addition to the activities that triggered this work, there are additional RFCs and Internet-drafts that may benefit from an interconnection PHB and DSCP scheme. RFC 5160 suggests Meta-QoS-Classes to enable deployment of standardized end to end QoS classes [RFC5160]. The authors of that RFC agree that the proposed interconnection class- and codepoint scheme and its enablement of standardised end to end classes would complement their own work.

Work on signaling Class of Service at interconnection interfaces by BGP [I-D.knoll-idr-cos-interconnect], [ID.idr-sla] is beyond the

scope of this draft. When the scheme in this document is used, signaled access to QoS classes may be of interest. These two BGP documents focus on exchanging SLA and traffic conditioning parameters and assume that common PHBs identified by the signaled DSCPs have been established prior to BGP signaling of QoS.

1.2. Applicability Statement

This document is primarily applicable to use of Differentiated Services for interconnection traffic between networks, and in particular to interconnection of MPLS-based networks. The approach described in this document is not intended for use within the interconnected (or other) networks, where the approach specified in RFC 5127 [RFC5127] is among the possible alternatives; see Section 3 for further discussion.

The Diffserv-Intercon approach described in this document simplifies IP based interconnection to domains operating the MPLS Short Pipe model to transport plain IP traffic terminating within or transiting through the receiving domain. Transit traffic is received and sent with the same PHB and DSCP. Terminating traffic maintains the PHB with which it was received, however the DSCP may change.

1.3. Document Organization

This document is organized as follows: section 2 reviews the MPLS Short Pipe tunnel model for DiffServ Tunnels [RFC3270]; effective support for that model is a crucial goal of this document. Section 3 provides background on RFC 5127's approach to traffic class aggregation within a DiffServ network domain and explains why this document uses a somewhat different approach. Section 4 introduces DiffServ interconnection Treatment Aggregates, plus the PHBs and DSCPs that are mapped to these Treatment Aggregates. Further, section 4 discusses treatment of non-tunneled and tunneled IP traffic and MPLS VPN QoS aspects. Finally Network Management PHB treatment is described. Appendix A describes the impact of the MPLS Short Pipe model (penultimate hop popping) on QoS for related IP interconnections.

2. MPLS and the Short Pipe tunnel model

The Pipe and Uniform models for Differentiated Services and Tunnels are defined in [RFC2983]. RFC3270 adds the MPLS Short Pipe model in order to support penultimate hop popping (PHP) of MPLS Labels, primarily for IP tunnels and VPNs. The Short Pipe model and PHP have become popular with many network providers that operate MPLS networks and are now widely used to transport non-tunneled IP traffic, not

just traffic encapsulated in IP tunnels and VPNs. This has important implications for DiffServ functionality in MPLS networks.

RFC 2474's recommendation to forward traffic with unrecognized DSCPs with Default (best effort) service without rewriting the DSCP has proven to be a poor operational practice. Network operation and management are simplified when there is a 1-1 match between the DSCP marked on the packet and the forwarding treatment (PHB) applied by network nodes. When this is done, CS0 (the all-zero DSCP) is the only DSCP used for Default forwarding of best effort traffic, so a common practice is to use CS0 to remark traffic received with unrecognized or unsupported DSCPs at network edges.

MPLS networks are more subtle in this regard, as it is possible to encode the provider's DSCP in the MPLS Traffic Class (TC) field and allow that to differ from the PHB indicated by the DSCP in the MPLS-encapsulated IP packet. That would allow an unrecognized DSCP to be carried edge-to-edge over an MPLS network, because the effective DSCP used by the MPLS network would be encoded in the MPLS label TC field (and also carried edge-to-edge); this approach assumes that a provider MPLS label with the provider's TC field is present at all hops within the provider's network. But this is only true for the Pipe tunnel model.

The Short Pipe tunnel model and PHP violate that assumption because PHP pops and discards the MPLS provider label carrying the provider's TC field. That discard occurs one hop upstream of the MPLS tunnel endpoint (which is usually at the network edge), resulting in no provider TC info being available at tunnel egress. To ensure consistent handling of traffic at the tunnel egress, the DSCP field in the MPLS-encapsulated IP header has to contain a DSCP that is valid for the provider's network; propagating another DSCP edge-to-edge requires an IP or MPLS tunnel of some form. See Appendix A for a more detailed discussion.

If transport of a large number (much greater than 4) DSCPs is required across a network that supports this DiffServ interconnection scheme, a tunnel or VPN can be provisioned for this purpose, so that the inner IP header carries the DSCP that is to be preserved not to be changed. From a network operations perspective, the customer equipment (CE) is the preferred location for tunnel termination, although a receiving domains Provider Edge router is another viable option.

3. Relationship to RFC 5127

This document draws heavily upon RFC 5127's approach to aggregation of DiffServ traffic classes for use within a network, but there are some important differences caused by the characteristics of network interconnects.

3.1. RFC 5127 Background

Many providers operate MPLS-based backbones that employ backbone traffic engineering to ensure that if a major link, switch, or router fails, the result will be a routed network that continues to meet its Service Level Agreements (SLAs). Based on that foundation, [RFC5127] introduced the concept of DiffServ Treatment Aggregates, which enable traffic marked with multiple DSCPs to be forwarded in a single MPLS Traffic Class (TC) based on robust provider backbone traffic engineering. This enables differentiated forwarding behaviors within a domain in a fashion that does not consume a large number of MPLS Traffic Classes.

RFC 5127 provides an example aggregation of DiffServ service classes into 4 Treatment Aggregates. A small number of aggregates are used because:

- The available coding space for carrying QoS information (e.g., DiffServ PHB) in MPLS (and Ethernet) is only 3 bits in size, and is intended for more than just QoS purposes (see e.g. [RFC5129]).

- There should be unused codes for interconnection purposes. This leaves space for future standards, for private bilateral agreements and for local use PHBs and DSCPs.

- Migrations from one code point scheme to another may require spare QoS code points.

RFC 5127 also follows RFC 2474 in recommending transmission of DSCPs through a network as they are received at the network edge.

3.2. Differences from RFC 5127

Like RFC 5127, this document also uses four traffic aggregates, but differs from RFC 5127 in three important ways:

- It follows RFC 2475 in allowing the DSCPs used within a network to differ from those to exchange traffic with other networks (at network edges), but provides support to restore ingress DSCP values if one of the recommended interconnect DSCPs in this draft is used. This results in DSCP remarking at both network ingress

and network egress, and this draft assumes that such remarking at network edges is possible for all interface types.

- It treats network control traffic as a special case. Within a network, the CS6 DSCP is used for local network control traffic (routing protocols and OAM traffic that is essential for network operation administration, control and management) that may be destined for any node within the network. In contrast, network control traffic exchanged between networks (e.g., BGP traffic) usually terminates at or close to a network edge, and is not forwarded through the network because it is not part of internal routing or OAM for the receiving network. In addition, such traffic is unlikely to be covered by standard interconnection agreements; it is more likely to be specifically configured (e.g., most networks impose on exchange of BGP for obvious reasons). See Section 4.2 for further discussion.

- Because network control traffic is treated as a special case, a fourth traffic aggregate is defined for use at network interconnections to replace the Network Control aggregate in RFC 5127. Network Control traffic may still be exchanged across network interconnections as further discussed in Section 4.2

4. The DiffServ-Intercon Interconnection Classes

At an interconnection, the networks involved need to agree on the PHBs used for interconnection and the specific DSCP for each PHB. This may involve remarking for the interconnection; such remarking is part of the DiffServ Architecture [RFC2475], at least for the network edge nodes involved in interconnection. This draft proposes a standard interconnection set of 4 Treatment Aggregates with well-defined DSCPs to be aggregated by them. A sending party remarks DSCPs from internal schemes to the interconnection code points. The receiving party remarks DSCPs to her internal scheme. The set of DSCPs and PHBs supported across the two interconnected domains and the treatment of PHBs and DSCPs not recognized by the receiving domain should be part of the interconnect SLA.

RFC 5127's four treatment aggregates include a Network Control aggregate for routing protocols and OAM traffic that is essential for network operation administration, control and management. Using this aggregate as one of the four in RFC 5127 implicitly assumes that network control traffic is forwarded in potential competition with all other network traffic, and hence DiffServ must favor such traffic (e.g., via use of the CS6 codepoint) for network stability. That is a reasonable assumption for IP-based networks where routing and OAM protocols are mixed with all other types of network traffic; corporate networks are an example.

In contrast, mixing of all traffic is not a reasonable assumption for MPLS-based provider or carrier networks, where customer traffic is usually segregated from network control (routing and OAM) traffic via other means, e.g., network control traffic use of separate LSPs that can be prioritized over customer LSPs (e.g., for VPN service) via other means. This segregation of network control traffic from customer traffic is also used for MPLS-based network interconnections. In addition, many customers of a network provider do not exchange Network Control traffic (e.g., routing) with the network provider. For these reasons, a separate Network Control traffic aggregate is not important for MPLS-based carrier or provider networks; when such traffic is not segregated from other traffic, it may reasonably share the Assured Elastic treatment aggregate (as RFC 5127 suggests for a situation in which only three treatment aggregates are supported).

In contrast, VoIP is emerging as a valuable and important class of network traffic for which network-provided QoS is crucial, as even minor glitches are immediately apparent to the humans involved in the conversation.

Similar approaches to use of a small number of traffic aggregates (including recognition of the importance of VoIP traffic) have been taken in related standards and recommendations from outside the IETF, e.g., Y.1566 [Y.1566], GSMA IR.34 [IR.34]and MEF23.1 [MEF23.1].

The list of the four DiffServ Interconnect traffic aggregates follows, highlighting differences from RFC 5127 and the specific traffic classes from RFC 4594 that each class aggregates.

Telephony Service Treatment Aggregate: PHB EF, DSCP 101 110 and VOICE-ADMIT, DSCP 101100, see [RFC3246] , [RFC4594][RFC5865]. This Treatment Aggregate corresponds to RFC 5127s real time Treatment Aggregate definition regarding the queuing, but it is restricted to transport Telephony Service Class traffic in the sense of RFC 4594.

Bulk Real-Time Treatment Aggregate: This Treatment Aggregate is designed to transport PHB AF41, DSCP 100 010 (the other AF4 PHB group PHBs and DSCPs may be used for future extension of the set of DSCPs carried by this Treatment Aggregate). This Treatment Aggregate is designed to transport the portions of RFC 5127's Real Time Treatment Aggregate, which consume large amounts of bandwidth, namely Broadcast Video, Real-Time Interactive and Multimedia Conferencing. The treatment aggregate should be configured with a rate queue (which is in line with RFC 4594 for the mentioned traffic classes). As compared to RFC 5127, the number of DSCPs has been reduced to

one (initially). The proposed queuing mechanism is in line with RFC4594 definitions for Broadcast Video and Real-Time Interactive. If need for three-color marked Multimedia Conferencing traffic arises, AF42 and AF43 PHBs may be added.

Assured Elastic Treatment Aggregate This Treatment Aggregate consists of the entire AF3 PHB group AF3, i.e., DSCPs 011 010, 011 100 and 011 110. As compared to RFC5127, just the number of DSCPs, which has been reduced. This document suggests to transport signaling marked by AF31. RFC5127 suggests to map Network Management traffic into this Treatment Aggregate, if no separate Network Control Treatment Aggregate is supported (for a more detailed discussion of Network Control PHB treatment see section 3.2). GSMA IR.34 proposes to transport signaling traffic by AF31 too.

Default / Elastic Treatment Aggregate: transports the default PHB, CS0 with DSCP 000 000. RFC 5127 example refers to this Treatment Aggregate as Aggregate Elastic. An important difference as compared to RFC5127 is that any traffic with unrecognized or unsupported DSCPs may be remarked to this DSCP. If a Lower Effort PHB, as proposed by e.g. [RFC3662], is standardised, Diffserv-Intercon support is suggested by marking this traffic with a DSCP 000 xx0 at interconnection interfaces. Note that this requires standardisation (this document is informational only).

RFC 4594's Multimedia Streaming class has not been mapped to the above scheme. By the time of writing, the most popular streaming applications use TCP transport and adapt picture quality in the case of congestion. A possible deployment of Diffserv-Intercon may be to move all quality content to the Bulk Real Time Treatment Aggregate. The Assured Elastic Treatment Aggregate may be used to carry signaling traffic. Another option is to carry Multimedia Streaming as part of the Assured Elastic Treatment aggregate, as its properties are suitable for protocols applying automated retransmit requests.

The overall approach to DSCP marking at network interconnections is illustrated by the following example. Provider O and provider W are peered with provider T. They have agreed upon a QoS interconnection SLA.

Traffic of provider O terminates within provider Ts network, while provider W's traffic transits through the network of provider T to provider F. Assume all providers run their own internal codepoint schemes for a PHB group with properties of the DiffServ-Intercon Assured Treatment Aggregate.

```
           Provider-O             Provider-W
           RFC5127                GSMA 34.1
               |                      |
          +----------+           +----------+
          |AF21, AF22|           | CS3, CS2 |
          +----------+           +----------+
               |                      |
               V                      V
           +++++++++              +++++++++
           |Rtr PrO|              |Rtr PrW|  Rtr Pr:
           +++++++++              +++++++++  Router Peering
               |        DiffServ      |
          +----------+           +----------+
          |AF31, AF32|           |AF31, AF32|
          +----------+           +----------+
               |        Intercon      |
               V                      V
           +++++++++                  |
           |RtrPrTI|------------------+
           +++++++++
               |     Provider-T domain
          +-----------+
          | MPLS TC 2 |
          | DSCP rew. |           rew. -> rewrite
          | AF21, AF22|
          +-----------+
             |      |  Local DSCPs Provider-T
             |      |  +----------+   +++++++++
             V      +->|AF21, AF22|->-|RtrDstH|
             |         +----------+   +++++++++
         +----------+                 RtrDst:
         |AF21, AF22|                 Router Destination
         +----------+
             |
          +++++++++
          |RtrPrTE|
          +++++++++
             |        DiffServ
         +----------+
         |AF31, AF32|
         +----------+
             |        Intercon
          +++++++++
          |RtrPrF|
          +++++++++
             |
         +----------+
         | CS4, CS3 |
```

```
         +----------+
              |
         Provider-F
         GSM IR.34
```

DiffServ-Intercon example

Figure 1

Providers only need to deploy internal DSCP to DiffServ-Intercon DSCP mappings to exchange traffic in the desired classes. Provider W has decided that the properties of his internal classes CS3 and CS2 are best met by the Diffserv-Intercon Assured Elastic Treatment Aggregate, PHBs AF31 and AF32 respectively. At the outgoing peering interface connecting provider W with provider T the former's peering router remarks CS3 traffic to AF31 and CS2 traffic to AF32. The domain internal PHBs of provider T that meet the requirements of Diffserv-Intercon Assured Elastic Treatment Aggregate are AF2x. Hence AF31 traffic received at the interconnection with provider T is remarked to AF21 by the peering router of domain T, and domain T has chosen to use MPLS TC value 2 for this aggregate. Traffic received with AF32 is similarly remarked to AF22, but uses the same MPLS TC for the Treatment Aggregate, i.e. TC 2. At the penultimate MPLS node, the top MPLS label is removed. The packet should be forwarded as determined by the incoming MPLS TC. The peering router connecting domain T with domain F classifies the packet by it's domain T internal DSCP AF21 for the Diffserv-Intercon Assured Elastic Treatment Aggregate. As it leaves domain T on the interface to domain F, this causes the packet to be remarked to AF31. The peering router of domain F classifies the packet for domain F internal PHB CS4, as this is the PHB with properties matching DiffServ-Intercon's Assured Elastic Treatment Aggregate. Likewise, AF21 traffic is remarked to AF32 by the peering router od domain T when leaving it and from AF32 to CS3 by domain F's peering router when receiving it.

This example can be extended. Suppose Provider-O also supports a PHB marked by CS2 and this PHB is supposed to be transported by QoS within Provider-T domain. Then Provider-O will remark it with a DSCP other than the AF31 DSCP in order to preserve the distinction from CS2; AF11 is one possibility that might be private to the interconnection between Provider-O and Provider-T; there's no assumption that Provider-W can also use AF11, as it may not be in the SLA with Provider-W.

Now suppose Provider-W supports CS2 for internal use only. Then no DiffServ- Intercon DSCP mapping may be configured at the peering

router. Traffic, sent by Provider-W to Provider-T marked by CS2 due to a misconfiguration may be remarked to CS0 by Provider-T.

See section 4.1 for further discussion of this and DSCP transparency in general.

RFC2575 states that Ingress nodes must condition all other inbound traffic to ensure that the DS codepoints are acceptable; packets found to have unacceptable codepoints must either be discarded or must have their DS codepoints modified to acceptable values before being forwarded. For example, an ingress node receiving traffic from a domain with which no enhanced service agreement exists may reset the DS codepoint to the Default PHB codepoint. As a consequence, an interconnect SLA needs to specify not only the treatment of traffic that arrives with a supported interconnect DSCP, but also the treatment of traffic that arrives with unsupported or unexpected DSCPs.

The proposed interconnect class and code point scheme is designed for point to point IP layer interconnections among MPLS networks. Other types of interconnections are out of scope of this document. The basic class and code point scheme is applicable on Ethernet layer too, if a provider e.g. supports Ethernet priorities like specified by IEEE 802.1p.

4.1. End-to-end QoS: PHB and DS CodePoint Transparency

This section briefly discusses end-to-end QoS approaches related to the Uniform, Pipe and Short Pipe tunnel model.

- With the Uniform model, neither DCSP nor PHB change when an interconnected network is passed. This would mean that a packet received with syntax network management, marked by CS6 is, if MPLS is applicable, forwarded with an MPLS label marked TC6. The uniform model is not within scope of this document.

- With the Pipe model, the inner tunnel DCSP remains unchanged, but an outer tunnel DSCP and the PHB may change when an interconnected network is passed. This would mean that a packet received with (private) syntax scavenger marked by DSCP CS1, is transported by default PHB and if MPLS is applicable, forwarded with an MPLS label marked TC0. CS1 is not rewritten. The Pipe model is not within scope of this document.

- With the Short Pipe model, the DCSP likely changes and the might PHB change when an interconnected network is passed. This draft describes a method to speed up and simplify QoS interconnection if a DSCP rewrite can't be avoided. It offers a set of PHBs and

treatment aggregates as well as a set of interconnection DSCPs allowing straightforward rewriting to domain-internal DSCPs as well as defined forwarding and markings to the next domain. DiffServ-Intercon supports the Short Pipe model. The solution described here can be used in other contexts benefitting from a defined interconnection QoS interface.

The basic idea is that traffic sent with a DiffServ interconnect PHB and DSCP is restored to that PHB and DSCP at each network interconnection, even though a different PHB and DSCP may be used by each network involved. The key requirement is that the network ingress interconnect DSCP be restored at network egress, and a key observation is that this is only feasible in general for a small number of DSCPs.

4.2. Treatment of Network Control traffic at carrier interconnection interfaces

As specified by RFC4594, section 3.2, Network Control (NC) traffic marked by CS6 is to be expected at some interconnection interfaces. This document does not change RFC4594, but observes that network control traffic received at network ingress is generally different from network control traffic within a network that is the primary use of CS6 envisioned by RFC 4594. A specific example is that some CS6 traffic exchanged across carrier interconnections is terminated at the network ingress node, e.g. if BGP is running between two routers on opposite ends of an interconnection link; in this case the operators would enter into a bilateral agreement to use CS6 for that BGP traffic.

The end-to-end QoS discussion in the previous section (4.1) is generally inapplicable to network control traffic - network control traffic is generally intended to control a network, not be transported across it. One exception is that network control traffic makes sense for a purchased transit agreement, and preservation of the CS6 DSCP marking for network control traffic that is transited is reasonable in some cases, although it is generally inappropriate to use CS6 for transiting traffic, including transiting network control traffic. Use of an IP tunnel is suggested in order to reduce the risk of CS6 markings on transiting network control traffic being interpreted by the network providing the transit. In this case, the CS6 marked traffic is forwarded based on the Uniform or Pipe model, Short Pipe doesn't apply.

If the MPLS Short Pipe model is deployed for non-tunneled IPv4 traffic, an IP network provider should limit access to the CS6 and CS7 DSCPs so that they are only used for network control traffic for the provider's own network.

Interconnecting carriers should specify treatment of CS6 marked traffic received at a carrier interconnection which is to be forwarded beyond the ingress node. An SLA covering the following cases is recommended when a provider wishes to send CS6 marked traffic across an interconnection link which isn't terminating at the interconnected ingress node:

- classification of traffic which is network control traffic for both domains. This traffic should be classified and marked for the NC PHB.

- classification of traffic which is network control traffic for the sending domain only. This traffic should be classified for a PHB offering similar properties as the NC class (e.g. AF31 as specified by this document). As an example GSMA IR.34 proposes an Interactive class / AF31 to carry SIP and DIAMETER traffic. While this is service control traffic of high importance to the interconnected Mobile Network Operators, it is certainly not Network Control traffic for a fixed network providing transit between such operators, and hence should not receive CS6 treatment in such a network.

- any other CS6 marked traffic should be remarked or dropped.

5. Acknowledgements

Bob Brisoe reviewed the draft and provided rich feedback. Fred Baker and Brian Carpenter provided intensive feedback and discussion. Al Morton and Sebastien Jobert provided feedback on many aspects during private discussions. Mohamed Boucadair and Thomas Knoll helped adding awareness of related work. James Polks discussion during IETF 89 helped to improve the relation of this draft to RFC 4594 and RFC 5127.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document does not introduce new features, it describes how to use existing ones. The security considerations of RFC 2475 [RFC2475] and RFC 4594 [RFC4594] apply.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

[RFC5865] Baker, F., Polk, J., and M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010.

8.2. Informative References

[I-D.knoll-idr-cos-interconnect]
Knoll, T., "BGP Class of Service Interconnection", draft-knoll-idr-cos-interconnect-14 (work in progress), May 2015.

[ID.idr-sla]
IETF, "Inter-domain SLA Exchange", IETF, http://datatracker.ietf.org/doc/draft-ietf-idr-sla-exchange/, 2013.

[IEEE802.1Q]
IEEE, "IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks", 2005.

[IR.34] GSMA Association, "IR.34 Inter-Service Provider IP Backbone Guidelines Version 7.0", GSMA, GSMA IR.34 http://www.gsma.com/newsroom/wp-content/uploads/2012/03/ir.34.pdf, 2012.

[MEF23.1] MEF, "Implementation Agreement MEF 23.1 Carrier Ethernet Class of Service Phase 2", MEF, MEF23.1 http://metroethernetforum.org/PDF_Documents/technical-specifications/MEF_23.1.pdf, 2012.

[RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.

[RFC3662] Bless, R., Nichols, K., and K. Wehrle, "A Lower Effort Per-Domain Behavior (PDB) for Differentiated Services", RFC 3662, December 2003.

[RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

[RFC5127] Chan, K., Babiarz, J., and F. Baker, "Aggregation of Diffserv Service Classes", RFC 5127, February 2008.

[RFC5160] Levis, P. and M. Boucadair, "Considerations of Provider-to-Provider Agreements for Internet-Scale Quality of Service (QoS)", RFC 5160, March 2008.

[Y.1566] ITU-T, "Quality of service mapping and interconnection between Ethernet, IP and multiprotocol label switching networks", ITU, http://www.itu.int/rec/T-REC-Y.1566-201207-I/en, 2012.

Appendix A. Appendix A The MPLS Short Pipe Model and IP traffic

The MPLS Short Pipe Model (or penultimate Hop Label Popping) is widely deployed in carrier networks. If non-tunneled IPv4 traffic is transported using MPLS Short Pipe, IP headers appear inside the last section of the MPLS domain. This impacts the number of PHBs and DSCPs that a network provider can reasonably support . See Figure 2 (below) for an example.

For tunneled IPv4 traffic, only the outer tunnel header is relevant for forwarding. If the tunnel does not terminate within the MPLS network section, only the outer tunnel DSCP is involved, as the inner DSCP does not affect forwarding behavior. In this case, the Pipe model applies.

Non-tunneled IPv6 traffic as well as Layer 2 and Layer 3 VPN traffic all use an additional MPLS label; in this case, the MPLS tunnel follows the Pipe model. Classification and queuing within an MPLS network is always based on an MPLS label, as opposed to the outer IP header.

Carriers often select QoS PHBs and DSCP without regard to interconnection. As a result PHBs and DSCPs typically differ between network carriers. PHBs may be mapped. With the exception of best effort traffic, a DSCP change should be expected at an interconnection at least for plain IP traffic, even if the PHB is suitably mapped by the carriers involved.

Beyond RFC3270's suggestions that the Short Pipe Model is only applicable to VPNs, current network structures also use it to transport non tunneled IPv4 traffic. This is shown in figure 2.

```
                |
       \|/           IPv4, DSCP_send
        V
        |
   Peering Router
        |
       \|/           IPv4, DSCP_send
        V
        |
   MPLS Edge Router
        |          Mark MPLS Label, TC_internal
       \|/         Remark DSCP to
        V            (Inner: IPv4, DSCP_d)
        |
   MPLS Core Router  (penultimate hop label popping)
        |                        \
        |            IPv4, DSCP_d |  The DSCP needs to be in network-
        |                 ^^^^^^^^|  internal QoS context. The Core
       \|/                         > Router might require or enforce
        V                         |  it. The Edge Router may wrongly
        |                         |  classify, if the DSCP is not in
        |                        /   network-internal DiffServ context.
   MPLS Edge Router
        |                        \   Traffic leaves the network marked
       \|/           IPv4, DSCP_d |  with the network-internal
        V                          > DSCP_d that must be dealt with
        |                         |  by the next network (downstream).
        |                        /
   Peer Router
        |          Remark DSCP to
       \|/           IPv4, DSCP_send
        V
        |
```

Short-Pipe / penultimate hop popping example

Figure 2

The packets IP DSCP must be in a well understood Diffserv context for schedulers and classifiers on the interfaces of the ultimate MPLS link (last link traversed before leaving the network). The necessary Diffserv context is network-internal and a network operating in this mode enforces DSCP usage in order to obtain robust QoS behavior.

Without DiffServ-Intercon treatment, the traffic is likely to leave each network marked with network-internal DSCP. DSCP_send of the figure above is remarked to the receiving network's DiffServ scheme.

It leaves the domain marked by the domains DSCP_d. This structure requires that every carrier deploys per-peer PHB and DSCP mapping schemes.

If Diffserv-Intercon is applied DSCPs for traffic transiting the domain can be mapped from and remapped to an original DSCP. This is shown in figure 3. Internal traffic may continue to use internal DSCPs (e.g, DSCP_d) and those may also be used between a carrier and its direct customers.

```
   Internal Router
        |
        |   Outer Header
       \|/    IPv4, DSCP_send
        V
        |
   Peering Router
        |  Remark DSCP to
       \|/    IPv4, DSCP_ds-int    DiffServ-Intercon DSCP and PHB
        V
        |
   MPLS Edge Router
        |
        |   Mark  MPLS Label, TC_internal
       \|/  Remark DSCP to
        V     (Inner: IPv4, DSCP_d)   domain internal DSCP for
        |                             the PHB
   MPLS Core Router  (penultimate hop label popping)
        |
        |     IPv4, DSCP_d
        |           ^^^^^^
       \|/
        V
        |
        |
   MPLS Edge Router--------------------+
        |                              |
       \|/  Remark DSCP to            \|/  IPv4, DSCP_d
        V     IPv4, DSCP_ds-int        V
        |                              |
        |                              |
   Peer Router              Domain internal Broadband
        |                        Access Router
       \|/  Remark DSCP to            \|/
        V     IPv4, DSCP_send          V  IPv4, DSCP_d
        |                              |
```

Short-Pipe example with Diffserv-Intercon

Figure 3

Appendix B. Change log

00 to 01 Added an Applicability Statement. Put the main part of the RFC5127 related discussion into a separate chapter.

01 to 02 More emphasis on the Short-Pipe tunel model as compared to Pipe and Uniform tunnel models. Further editorial improvements.

Authors' Addresses

Ruediger Geib (editor)
Deutsche Telekom
Heinrich Hertz Str. 3-7
Darmstadt 64295
Germany

Phone: +49 6151 5812747
Email: Ruediger.Geib@telekom.de

David L. Black
EMC Corporation
176 South Street
Hopkinton, MA
USA

Phone: +1 (508) 293-7953
Email: david.black@emc.com