Georgios Karagiannis
University of Twente / Ericsson

L. Westberg
A. Bader
Ericsson

Hannes Tschofenig
Siemens

October 22, 2006

Resource Unavailability (RU) Per Domain Behavior
draft-karagiannis-ru-pdb-03.txt

Status of this Memo

This Internet-Draft will expire on April 22, 2007.

Copyright Notice

Abstract

This draft specifies a Per Domain Behavior that provides the ability to Diffserv nodes located outside Diffserv domain(s), e.g., receiver or other Diffserv enabled router to detect when the resources provided by the Diffserv domain(s) are not available. The unavailability of resources in the domain is monitored and detected by proportionally marking packets whenever the current link rate exceeds some pre-configured SLS agreed throughput (bandwidth) threshold. It is considered that the SLS agreed throughput threshold is not statically but loosely defined in order to allow a more efficient utilization of the Diffserv domain(s) and a simpler network management operation. This PDB can be applied in association with either a single Diffserv domain or multiple neighboring Diffserv domains, when a trust relationship exist between these multiple Diffserv domains. This specification is denoted as Resource Unavailability (RU) PDB and it follows the guidelines given in [RFC3086].

Table of Contents

1. Introduction

1.1 Applicability

The RU PDB can be applied in the situation where Diffserv nodes located outside Diffserv domain(s) must detect when the resources provided by the Diffserv domain(s) are not available.

This PDB is used when the negotiated SLS is associated to throughput (or bandwidth) and when the SLS agreed throughput threshold is not statically but loosely defined. The main purpose of loosely defining the SLS throughput threshold is to allow a more efficient utilization of the Diffserv domain(s) and a more simple network management operation. This PDB can be applied in association with either a single Diffserv domain or multiple neighboring Diffserv domains, when SLA agreements exist between the operator(s) of these Diffserv domains.

The resource unavailability on the Diffserv nodes within the Diffserv domain(s) can be detected using a DSCP remarking approach where the packet remarking is proportional to the amount of unavailable resources. In particular, the Diffserv nodes mark packets whenever the measured link throughput rate exceeds the SLS pre-configured throughput threshold and the proportion of the marked packets is in proportion to the excess traffic above this SLS pre-configured throughput threshold.

The Diffserv nodes located outside the Diffserv domain(s) can use the remarked DSCP packets to calculate the percentage of throughput (or bandwidth) that does exceed the loosely agreed SLS throughput threshold.

These nodes can then, in combination with the sender of the traffic and the support of the Diffserv domain(s), reduce the generated throughput until the loosely agreed SLS throughput threshold is satisfied. Possible applicability areas on using the remarked DSCP packets/bytes are related to the support of final handling decisions on the admission control and/or rate control of ongoing calls/flows.

In particular, the RU-PDB mechanism can be used in combination with an end-to-end ECN (see, [RFC3168]) congestion control solution, to support any real time application, e.g., video, that use a rate adaptive coding that can adapt the bandwidth, e.g., Datagram Congestion Control Protocol (DCCP) based [RFC4340], but for a Feasible service quality it requires a minimum bandwidth.

A minimum bandwidth has to be allocated because otherwise the real time application service is useless. The minimum bandwidth is allocated by using the DSCP based RU-PDB mechanism in combination with the admission control functionality available at the nodes outside the Diffserv domain(s). If the Diffserv network has more capacity it can utilize that and give the end user a higher quality with better end user experience. The end-to-end ECN method can be used to monitor whether the network has more available capacity. Note that in this case the RU-PDB has to use DSCP marking (and not ECN marking) for the RU notifications. This is because an interoperability problem might occur between the end-to-end ECN marking used by DCCP and the RU PDB marking.

It is important to mention that the RU PDB operation does not require changes to the Diffserv model. The RU PDB is using typical measuring and Diffserv remarking techniques. The remarking procedure remarks packets from an original DSCP value to for example, an experimental or a local use DSCP.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Traffic Conditioning Specification (TCS) and PHB Configuration

Packets using any PHB can receive the RU PDB treatment. Furthermore, the RU PDB can be used in combination with any other defined PDB.

3.1. Diffserv Source End System Configuration

The Diffserv source end systems, which support the RU PDB, ensure that the packets are being marked with the right PHB. Note that the process of marking can be specified by another PDB. In this text, for simplicity reasons, the PDB that is defining the PHB marking is denoted as another_PDB. For each of the chosen PHB, the TCS and PHB configurations associated with the RU PDB are following the rules defined by another_PDB, which MAY use the specifications provided in [RFC2474], [RFC2475], [RFC3246], [RFC2597] and [RFC3290]. Otherwise the PHB configuration follows the rules specified by the PHB specification document, e.g., [RFC3246].

3.2 Common Diffserv node configurations

The Diffserv nodes, which are supporting the RU-PDB, must perform the following functionalities:

(1) Meter + (2) Marking Action: the Diffserv nodes must be configured with a meter and marking function that measures and remarks bytes that are out of a configured traffic profile (e.g., bandwidth threshold) for a corresponding PHB traffic class, to provide and indication of a potential resource limitation to a Diffserv node outside the domain. The traffic profile can be set according to an engineered bandwidth limitation based SLA or based on a capacity limitation of specific links. By using an algorithm that calculates the rate of bytes that are out of profile, say rate_out_profile_bytes, a number of bytes, i.e., rate_out_profile_bytes/N, are remarked to a second DSCP, denoted in this example as local_DSCP, that receives the same PHB as the original DSCP. "N" is a pre-configured parameter used to indicate the proportionality between the measured out of profile bytes and the remarked bytes. If "N" is used in the algorithm, then it must have the same value in all Diffserv nodes that use this mechanism.

(3) Packet Classification + (4) Scheduling: The Diffserv node SHOULD be configured to consider that the packets marked either with the original_DSCP or with the local_DSCP SHOULD receive the same per hop behavior treatment. However, packets that are marked with the local_DSCP, may be classified to enter a different and larger virtual queue than the packets marked with original_DSCP. This can ensure that the dropping probability of local_DSCP remarked packets is lower than the dropping probability of original_DSCP remarked packets. This classification can be accomplished by using the packet classification function, while the way of how the packets are treated in the virtual queues is accomplished using the scheduling function. Note that the original_DSCP marked packets and their associated local_DSCP packets get the same forwarding behavior. The main difference is related to the fact that the local_DSCP packets get a lower dropping probability compared to the original_DSCP packets. This is because the marking information carried by the local_DSCP packets has a higher significance for the operation of the resource unavailability algorithm compared to the marking information carried by the original_DSCP packets.

The two virtual queues, one for the original_DSCP and another one for local_DSCP marked packets can, for example, be implemented by using one Drop Tail physical queue and by maintaining queuing information and also one queuing threshold for each of the virtual queues. The physical queue uses the same scheduling algorithm, but the length of each of the virtual queue defines the packet dropping probability of a virtual queue.

The classification of packets SHOULD be based on either the DSCP or on a combination of IP header fields including the DSCP.
When a packet is received by the edge router of another domain (new Diffserv domain, that might be managed by another operator), remarking of the original_DSCP and local_DSCP to other DSCPs, say original_new_DSCP and local_new_DSCP might be necessary. This is because the neighbor DSCP operator may use different Diffserv mapping schemes. It is however, considered that SLA agreements exist between the operator(s) of these Diffserv domains, thus also the remarking rules followed in each Diffserv domain are known. Note that the Diffserv nodes used in the neigbouring Diffserv domains should use the same classification, meter & marking actions as described above.

3.3. Configuration of nodes outside the Diffserv domain(s)

When the Diffserv nodes located outside Diffserv domain(s), e.g., receiver Diffserv enabled end systems, receive the remarked packets, the rate of the received marked bytes, per each flow aggregate, is measured. Note that the calculated rate has to be corrected and multiplied with the parameter "N", see above, in order to calculate the real rate of overload, say real_rate_overload. This rate can be use to provide handling decisions on the resource unavailability functionality. Two types of handling decisions could be supported.

When only one pre-configured bandwidth threshold is maintained by this Diffserv node, say Threshold1, then if the calculated rate of remarked bytes is higher than Threshold1, i.e., real_rate_overload > Threshold1, then the Diffserv node can use this information to provide the basis of call admission decisions for new flows. Note that how the admission decision process on call level operates is out of the scope of this draft.

When two pre-configured bandwidth thresholds are used, i.e., Threshold1 and Threshold2, with Threshold2 > Threshold1, then the Diffserv node should operate in the following way. When the calculated rate, real_rate_overload > Threshold1 then the same procedure as described above is used (situation that only one threshold is used). When the calculated rate is higher than Threshold 2, then the Diffserv node can use procedures that are out the scope of this draft, to send notifications to ongoing sessions to enforce rate control. Note that Threshold2 is used in the case that a persistent congestion situation occurs and ongoing calls have to be notified about it.

Note that the flow aggregates are defined by source IP address ranges The size of the aggregates should be large enough to ensure that new calls belong to aggregates where ongoing calls provide feedback for admission control decisions.

4. Attributes of this PDB

The new attributes that are related to this PDB are related to the agreed SLS traffic profiles, e.g., bandwidth thresholds. Different agreed SLS throughput thresholds, see Section 3, might be used. Each of these throughput (bandwidth) thresholds are compared to the calculated rate of remarked packets/bytes..

5. Parameters

The used parameters are the SLS traffic profiles and bandwidth thresholds.

6. Assumptions

The negotiated SLA may include either one pre-configured loosely agreed SLS throughput (bandwidth) threshold or two pre-configured loosely agreed SLS throughput (bandwidth) thresholds (bound). It is assumed that the network operator communicates these throughput (bandwidth) thresholds from the location of where the SLS parameters are maintained up to the Diffserv nodes within the Diffserv domain(s).

The RU PDB can be applied on more than one neighboring Diffserv Domains, when SLA agreements exist between the operator(s) of these Diffserv domains. Therefore, it is possible that that a marked packet can be received by the edge router of a new neighboring Diffserv domain (and thus new domain operator). The new Diffserv domain may use another type of Diffserv remarking scheme. Thus the original_DSCP and local_DSCP, may be remarked to other DSCP. However, the network operator MUST configure the Diffserv remarking scheme such that the semantics and relations between the original_DSCP and local_DSCP remain even when packets using the RU PDB are passing via multiple neighboring Diffserv domains.

Furthermore, a network operator may configure Diffserv nodes located outside Diffserv domain(s) to provide final handling decisions on the resource unavailability and/or overload situation process, see Section 3.3.

If the parameter "N" is used in the algorithm, then it must have the same value in all Diffserv nodes that use this mechanism. "N" is a pre-configured parameter used to indicate the proportionality between the measured out of profile bytes and the remarked bytes.

A domain that does not support the remarking procedures described in this document should convey DSCP information without any modification.

A domain that applies tunneling techniques (MPLS or IP) and does not support the remarking procedures described in this document, should use the Diffserv marking of the inner header when the header of the tunnel is removed. Furthermore, the domain should set the outer header at the entry of the tunnel based on the DS field of the IP packet and map the outer header to the DS field of the IP packet at the end of the tunnel. In MPLS the original_DSCP or local_DSCP should be mapped to different EXP codepoint (Experimental field of MPLS header). In IP, original_DSCP or local_DSCP should be written to the DS field of outer header.

7. Example uses

This section gives an example of how the RU-PDB can be used in a mobile system, and in particular in the wired parts of cellular systems, such as the IP Diffserv based backbone.

Note however, that this example can be applied in any IP based Diffserv scenario that supports real-time applications, which use media codecs, and that do not have the benefit of being totally free in selecting/producing bit-rates. Many media codecs are only able to produce certain steps in bit-rate. They are also commonly looked to certain packet rates or transmission intervals to achieve good performance. There is also the issue that the actual change of bit-rate, and thus quality level, is noticeable and disturbing to the consumer of the media. Which combined results in that bit-rate changes usually needs to be done as seldom as possible and may only be done in steps, sometimes quite large steps. In this situation, in order to achieve the best possible behavior it would be very beneficial if the sending application would know with a relatively high probability that the higher bit-rate would be supported before even trying to utilize it. The real time application can then make use of two mechanisms.

First, during admission control, the real time application claims a high percentage, say 70%, of the bandwidth required to operate at a minimum acceptable level from the point of view of the end user. This mechanism can be accomplished by using the described in this document RU-PDB.

Second, for the rest of the bandwidth, say 30%, required by the application, the sending application could increase the bit-rate and thus the quality level, when it will know with a high probability that the higher bit-rate would be end to end supported. This mechanism can be accomplished by using ECN response specified in [RFC3168], [RFC4340] which will provide real-time media applications with a basic tool for adaptation. If the receiver detects packets marked with Congestion Experienced (CE), it can use that in its adaptation mechanism to reduce bandwidth, by reporting the event to the sender. This is accomplished in the same way as a packet loss would have been notified. However with the benefit that the payload carried by the packet was not lost, which improves the media quality by reducing the number of lost payloads.

A possible way of achieving this is that the sender will increase the transmitted rate, step-by step. For example, during each step the bandwidth could be increased by 5%. If during each step the receiver detects packets marked with Congestion Experienced (CE), it can then use the information in its adaptation mechanism to reduce the increased bandwidth, by reporting the event to the sender.

In the remaining part of this section, more details are given on how the RU-PDB can perform the first mechanism described above, which can be used in a mobile system, and in particular in the IP Diffserv based backbone of cellular systems. It is considered however, that also the second mechanism, which uses the ECN response is also applied, but it will not be explained in this example.

Usually in such a system, a Media Gateway is used between a sender and a receiver of a real time media application, see Figure 1. Note that such an entity can usually perform media transcoding functions.

```
           (SLA)                        (SLA)     Media
             |                            |      Gateway
Source Edge  V  Edge   Interior     Edge  V  Edge  (CAC)  Receiver
|       |        |        |        |        |        |        |
|       C        |  data  | data   |        |        |        |
| data  C#marked |        |        |        |        |        |
|------>C------->|------->|--------|------->|------->|        |
|       #unmarked|        |        |        |        |        |
|       |------->|------->|------->|------->|------->|------->|
```

Figure: 1 Admission control (CAC) using RU-PDB

It is considered in this example that the assumptions described in Section 6 are valid. In particular, it is assumed that the negotiated SLA includes one pre-configured loosely agreed SLS throughput (bandwidth) threshold required by the real time application to operate at a minimum acceptable level from the point of view of the end user.

In order to provide admission control, a Call Admission Control (CAC) function has to be supported by a node outside the Diffserv domain, that in this example is the Media Gateway, see Figure 1. Note that the way of how the CAC function is applied to admit or reject a flow is out of the scope of this example.

The example will be described in steps:

(1) An IP packet is sent from the source towards the receiver. Note that all packets will pass through the Media Gateway. The packets are Diffserv marked to receive a relatively high level of QoS, e.g., with Expedited Forwarding (EF). The packet is received by the first edge router. The edge router monitors to see if this packet is out-of-profile. If the edge, see Figure 1, realizes that a packet is out-of-profile, then the packet is marked to a second (local) DSCP, say local_EF DSCP. Note that the traffic profile of the meter & marking function available in each node should be lower than the bandwidth agreed in the SLA or the bandwidth available for EF traffic on links. The bandwidth between profiles provides an interval where feedback on the resource availability is already sent but the actual resource limitation is not reached, which allows the Media Gateway CAC function to interpret the resource unavailability notification and block new calls before reaching congestion. Furthermore, note that the Diffserv scheduling in routers is configured to use the same physical queue for packets marked with the original DSCP (EF) and with the local DSCP (local_EF DSCP). Note however, that different virtual queues might be used, see Section 3. The packet is then forwarded further.

(2) The packet is received by the ingress edge router of the next domain. This domain may be new Diffserv domain, which is administrated by a new backbone operator. The new Diffserv domain may use a different Diffserv mapping scheme, so remarking local_EF DSCP packets to another DSCP may be necessary. However, due to the SLA agreements that exist between the two backbone operators, the semantics of interpreting the new value of the local_EF DSCP will remain. Furthermore, the same meter & marking function that was explained in step 1 will also be applied. The packet is then forwarded further.

(3) The packet is processed by the interior node(s) in the domain in a similar way as described in step (1).

(4) The packet is received by the egress of the same domain. It is processed as described in step (1).

(5) The packet is received by an ingress. It is processed as described in step (2).

(6) Marked and remarked packets are received by the CAC function of the Media Gateway. The amount of remarked packets (local_EF DSCP) is counted in this node to provide the basis of call admission decisions for new flows, see Section 3.3. Note that the amount of remarked packets is counted separately for flow aggregates, which are defined by source IP address ranges, see Section 3.3.

(7) The packet is marked back to the original DSCP (EF) and forwarded towards its destination.

8. Environmental concerns

There are no environmental concerns specific to this PDB. However, depending on the the area wherein the RU PDB is applied (one Diffserv domain or multiple neigboring domains), different requirements have to be fulfilled by the network operators, see Section 6.

9. Security considerations for RU PDB

There are no specific security exposures for this PDB. See the general security considerations in [RFC2474] and [RFC2475]. Note that when multiple Diffserv domains are using the RU-PDB, SLA agreements should exist between the operator(s) of these multiple neighboring Diffserv domains and therefore, it is considered that a trust relationship exist between these domains.

10. IANA Considerations

[Editor's Note: A future version of this document will provide instructions to IANA.]

11. Acknowledgements

We thank Kathie Nichols for reviewing this draft and providing Useful comments.

12. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, 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, and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules For their Specification", RFC 3086, April 2001.

13 Informative References

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

[RFC3168] Ramakrishnan, K., Floyd, S., Black, D., "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC3290] Bernet, Y., Blake, S., Grossman, D., Smith, A., "An Informal Management Model for Diffserv Routers", RFC 3290, May 2002.

[RFC3246] B. Davie, et al., "An Expedited Forwarding PHB (Per-Hop Behavior) ", RFC 3246, March 2002.

[RFC4340] Kohler, E., Handley, M., Floyd, S., "Datagram Congestion Control Protocol (DCCP)", RFC 4340, March 2006.

Authors' Addresses

Georgios Karagiannis
University of Twente
P.O. BOX 217
7500 AE Enschede, The Netherlands
EMail: g.karagiannis@ewi.utwente.nl

Lars Westberg
Ericsson Research
Kistagangen 26
SE-164 80 Stockholm
Sweden
EMail: Lars.Westberg@ericsson.com

Attila Bader
Ericsson Research
Ericsson Hungary Ltd.
Laborc 1, Budapest, H-1037
Hungary
EMail: Attila.Bader@ericsson.com

Hannes Tschofenig
Siemens
Otto-Hahn-Ring 6
Munich, Bavaria 81739
Germany
EMail: Hannes.Tschofenig@siemens.com

Intellectual Property Statement

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.