NSIS Working Group
INTERNET-DRAFT

Expires: 23 December 2006

Attila Bader
Lars Westberg
Ericsson
Georgios Karagiannis
University of Twente
Cornelia Kappler
Siemens
Tom Phelan
Sonus

June 23, 2006

RMD-QOSM - The Resource Management in Diffserv QOS Model
<draft-ietf-nsis-rmd-07.txt>

Status of this Memo

Copyright Notice

Abstract

This document describes an NSIS QoS Model for networks that use the Resource Management in Diffserv (RMD) concept. RMD is a technique for adding admission control and preemption function to Differentiated Services (Diffserv) networks. The RMD QoS Model allows devices external to the RMD network to signal reservation requests to edge nodes in the RMD network. The RMD Ingress edge nodes classify the incoming flows into traffic classes and signals resource

requests for the corresponding traffic class along the data path to the Egress edge nodes for each flow. Egress nodes reconstitute the original requests and continue forwarding them along the data path towards the final destination. In addition, RMD defines notification functions to indicate overload situations within the domain to the edge nodes.

Table of Contents

1. Introduction

This document describes a Next Steps In Signaling (NSIS) QoS model for networks that use the Resource Management in Diffserv (RMD) framework ([RMD1], [RMD2], [RMD3], [RMD4]). RMD adds admission control to Diffserv networks and allows nodes external to the networks to dynamically reserve resources within the Diffserv domains.

The Quality of Service NSIS Signaling Layer Protocol (QoS-NSLP) [QoS-NSLP] specifies a generic model for carrying Quality of Service (QoS) signaling information end-to-end in an IP network. Each network along the end-to-end path is expected to implement a specific QoS Model (QOSM) that interprets the requests and installs the necessary mechanisms, in a manner that is appropriate to the technology in use in the network, to ensure the delivery of the requested QoS.

This document specifies an NSIS QoS Model for RMD networks (RMD-QOSM), and an RMD-specific QSpec (RMD-QSPec) for expressing reservations in a suitable form for simple processing by internal nodes. They are used in combination with the QoS-NSLP to provide QoS signaling service in an RMD network. Figure 1 shows an RMD network with the respective entities.

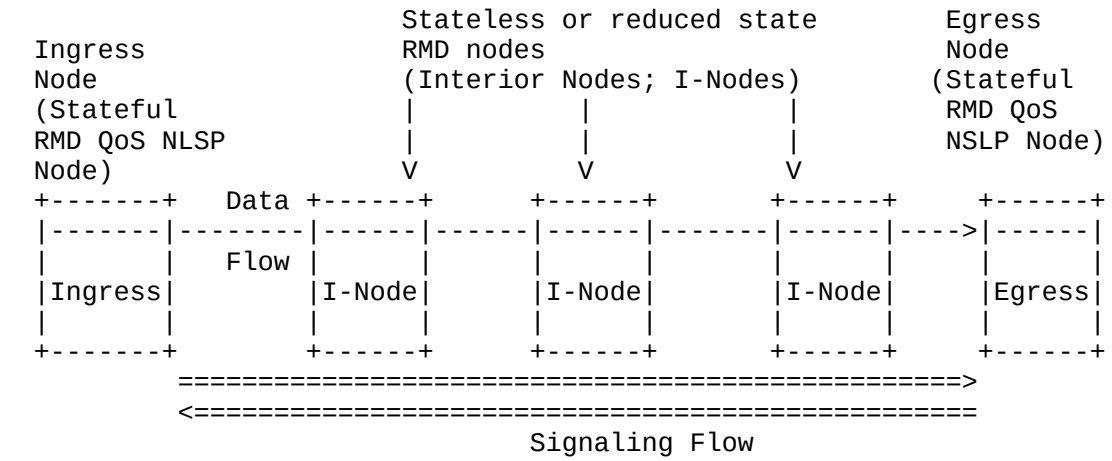

FIGURE 1: Actors in the RMD-QOSM

Internally to the RMD network, RMD-QOSM defines a scalable QoS signaling model in which per-flow QoS-NSLP and NTLP states are not stored in Interior nodes but per-flow signaling is performed (see [QoS-NSLP]).

In the RMD-QOSM, only routers at the edges of a Diffserv domain (Ingress and Egress nodes) support the QoS-NSLP stateful operation. Interior nodes support either the QoS-NSLP stateless operation, or a reduced-state operation with coarser granularity than the edge nodes.

The remainder of this draft is structured following the suggestions in Appendix B of [QSP-T] for the description of QoS Signaling Policies.

After the terminology in Section 2, we give an overview of RMD and the RMD-QOSM in Section 3. In Section 4 we give a detailed description of the RMD-QOSM, including the role of QNEs, the definition of the QSpec, mapping of QSpec generic parameters onto RMD-QOSM parameters, state management in QNEs, and operation and sequence of events. Section 5 discusses security issues.

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 RFC 2119.

The terminology defined by GIST [GIST] and QoS-NSLP [QoS-NSLP] applies to this draft.

In addition, the following terms are used:

Edge node: an (NSIS-capable) node on the boundary of some administrative domain.

Ingress node: An edge node that handles the traffic as it enters the domain.

Egress node: An edge node that handles the traffic as it leaves the domain.

Interior nodes: the set of (NSIS-capable) nodes which form an administrative domain, excluding the edge nodes.

3. Overview of RMD and RMD-QOSM

3.1. RMD

The Differentiated Services (Diffserv) architecture ([RFC2475], [RFC2638]) was introduced as a result of efforts to avoid the scalability and complexity problems of Intserv [RFC1633]. Scalability is achieved by offering services on an aggregate rather than per-flow basis and by forcing as much of the per-flow state as possible to the edges of the network. The service

differentiation is achieved using the Differentiated Services (DS) field in the IP header and the Per-Hop Behavior (PHB) as the main building blocks. Packets are handled at each node according to the PHB indicated by the DS field in the message header.

The Diffserv architecture does not specify any means for devices outside the domain to dynamically reserve resources or receive indications of network resource availability. In practice, service providers rely on subscription-time Service Level Agreements (SLAs) that statically define the parameters of the traffic that will be accepted from a customer.

RMD was introduced as a method for dynamic reservation of resources within a Diffserv domain. It describes a method that is able to provide admission control for flows entering the domain and a congestion handling algorithm that is able to terminate flows in case of congestion due to a sudden failure (e.g., link, router) within the domain.

In RMD, scalability is achieved by separating a fine-grained reservation mechanism used in the edge nodes of a Diffserv domain from a much simpler reservation mechanism needed in the Interior nodes. In particular, it is assumed that edge nodes support per-flow QoS states in order to provide QoS guarantees for each flow. Interior nodes use only one aggregated reservation state per traffic class or no states at all. In this way it is possible to handle large numbers of flows in the Interior nodes. Furthermore, due to the limited functionality supported by the Interior nodes, this solution allows fast processing of signaling messages.

In RMD two basic admission control modes are described: reservation-based and measurement-based admission control.

In the reservation-based method, each Interior node maintains only one reservation state per traffic class. The Ingress edge nodes aggregate individual flow requests into classes, and signal changes in the class reservations as necessary. The reservation is quantified in terms of resource units. These resources are requested dynamically per PHB and reserved on demand in all nodes in the communication path from an Ingress node to an Egress node.

The measurement-based algorithm continuously measures traffic levels and the actual available resources, and admits flows whose resource needs are within what is available at the time of the request. Once an admission decision is made, no record of the decision need be kept. The advantage of measurement-based resource management protocols is that they do not require pre-reservation state nor explicit release of the reservations. Moreover, when the user traffic is variable, measurement based admission control could provide higher network utilization than, e.g., peak-rate reservation. However, this can introduce an uncertainty in the availability of the resources.

Two types of measurement based admission control schemes are possible:

* Congestion notification function based on probing:

This method can be used to implement a simple measurement-based admission control within a Diffserv domain. In this scenario the interior nodes are not NSIS aware nodes. In these interior nodes thresholds are set for the traffic belonging to different PHBs in the measurement based admission control function. In this scenario an end-to-end NSIS message are used as a probe packet, meaning that the DSCP field in the header of the IP packet that carries the NSIS message is re-marked when the predefined congestion threshold is exceeded. In this way the edges can admit or reject flows that are requesting resources. Note that in this situation, in addition to the probe packet, also ordinary data packets passing though the congested node are re-marked.The rate of the re-marked data packets is used to detect a congestion situation that can influence the admission control decissions.

* NSIS measurement-based admission control:

In this case the measurement-based admission control functionality is implemented in NSIS aware stateless routers. The main difference between this type of admission control and the congestion notification based on probing is related to the fact that this type of admission control is applied mainly on NSIS aware nodes, giving the possibility to apply measuring techniques, see e.g., [JaSh97], [GrTs03], that are using current and past information on NSIS sessions that requested resources from an NSIS aware interior node. The admission decision is positive if the currently carried traffic, as characterized by the measured statistics, plus the requested resources for the new flow exceeds the system capacity with a probability smaller than some alpha. Otherwise, the admission decision is negative.

RMD describes the following procedures:

* Classification of an individual resource reservation or a resource query into Per Hop Behavior (PHB) groups at the Ingress node of the domain,

* Hop-by-hop admission control based on a PHB within the domain. There are two possible modes of operation for internal nodes to admit requests. One mode is the stateless or measurement-based mode, where the resources within the domain are queried. Another mode of operation is the reduced-state reservation or reservation based mode, where the resources within the domain are reserved.

* a method to forward the original requests across the domain up to the Egress node and beyond.

* a congestion control algorithm that notifies the egress edge nodes about congestion. It is able to terminate the appropriate number of flows in case a of congestion due to a sudden failure (e.g., link or router failure) within the domain.

3.2. Basic features of RMD-QOSM

3.2.1 Role of the QNEs

The protocol model of the RMD-QOSM is shown in Figure 2. The figure shows QNI and QNR nodes, not part of the RMD network, that are the ultimate initiator and receiver of the QoS reservation requests. It also shows QNE nodes that are the Ingress and Egress nodes in the RMD domain (QNE Ingress and QNE Egress), and QNE nodes that are Interior nodes (QNE Interior).

All nodes of the RMD domain are mainly QoS-NSLP aware nodes. Edge nodes store and maintain QoS-NSLP and NTLP states and therefore are stateful nodes. The NSIS aware Interior nodes are NTLP stateless. Furthermore they are either QoS-NSLP stateless (for NSIS measurement-based operation), or are reduced state nodes storing per PHB aggregated QoS-NSLP states (for reservation-based operation).

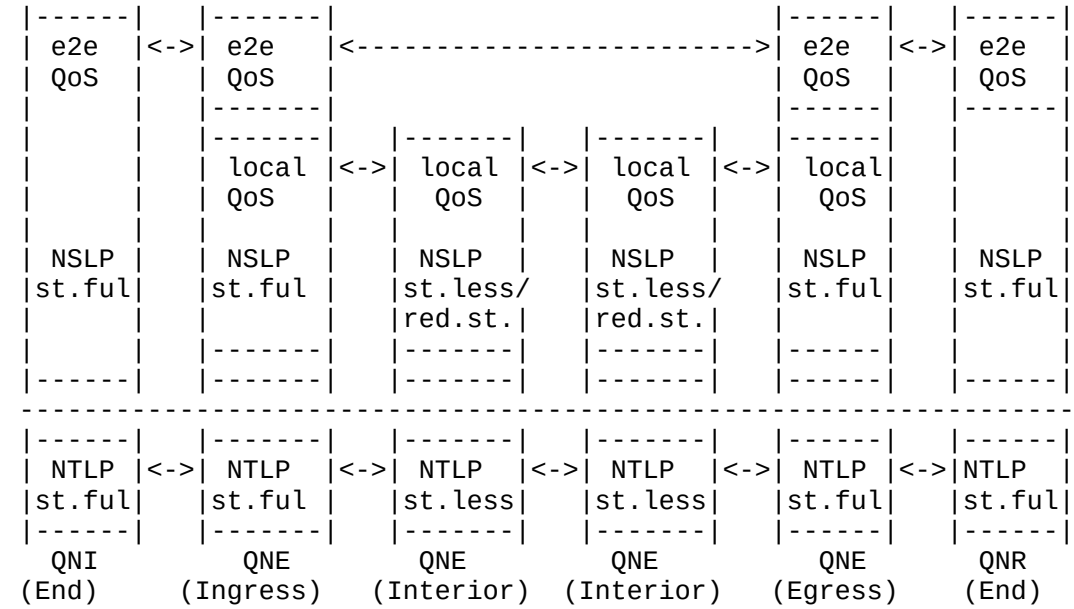

Figure 2: Protocol model of stateless/reduced state operation

Note that the RMD domain may contain Interior nodes that are not NSIS aware nodes (not shown in the figure). These nodes are assumed to have sufficient capacity for flows that might be admitted. Furthermore, some of these NSIS unaware nodes may be used for measuring the traffic congestion level on the data path. These measurements can be used by RMD-QOSM in the congestion control based on probing operation and/or severe congestion operation (see Section 4.6.1.6).

3.2.2 RMD-QOSM Signaling

The basic RMD-QOSM signaling is shown in Figure 3. A RESERVE message is created by a QNI with an Initiator QSpec describing the reservation and forwarded along the path towards the QNR. When the original RESERVE message arrives at the Ingress node, an RMD-QSpec is constructed based on the top-most QSPEC in the message (usually the Initiator QSPEC). The RMD-QSpec is sent in a local, independent RESERVE message through the Interior nodes towards the QNR. This local RESERVE message uses the NTLP hop-by-hop datagram signaling mechanism. Meanwhile, the original RESERVE message is sent to the Egress node on the path to the QNR using the reliable transport mode of NTLP.

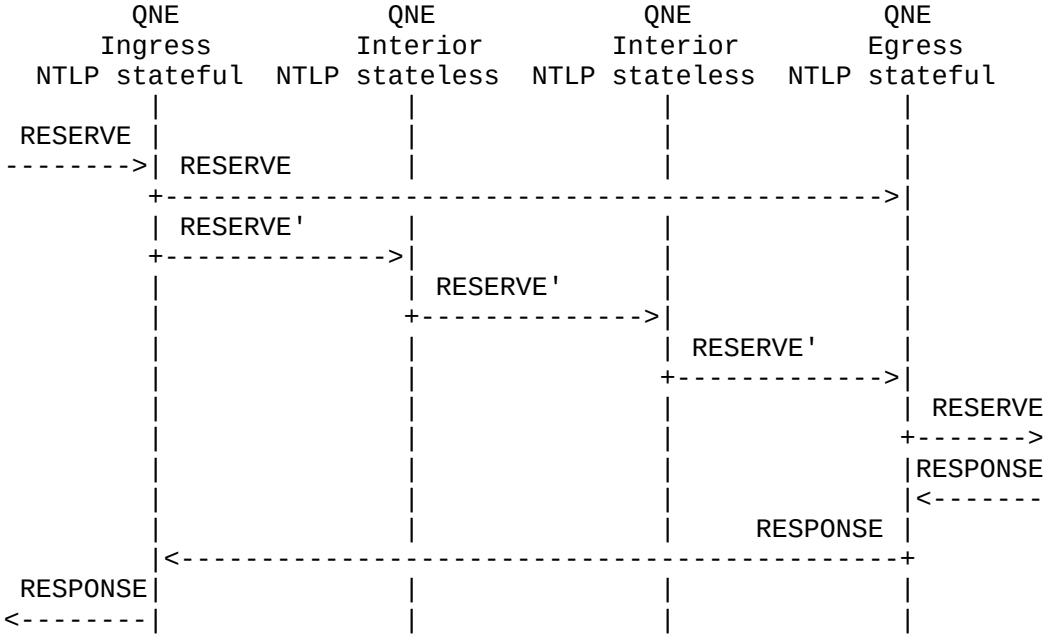

Figure 3: Sender-initiated reservation with Reduced State Interior Nodes

Each QoS-NSLP node on the data path processes the local RESERVE message and checks the availability of resources with either the reservation-based or the measurement-based method. When the message reaches the Egress node, and the reservation is successful in each Interior nodes, the original RESERVE message is forwarded to the next domain. When the Egress node receives a RESPONSE message from the downstream end, it is forwarded directly to the Ingress node.

If an intermediate node cannot accommodate the new request, it indicates this by marking a single bit in the message, and continues forwarding the message until the Egress node is reached. From the Egress node a RESPONSE message is sent directly the Ingress node.

As a consequence in the stateless/reduced state domain only sender-initiated reservation can be performed and functions requiring per flow NTLP or QoS-NSLP states, like summary refreshes, cannot be used. If per flow
identification, is needed, i.e., associating the flow IDs for the reserved resources, Edge nodes act on behalf of Interior nodes.

3.2.3 RMD-QOSM Applicability and considerations

The RMD-QOSM is a Diffserv-based bandwidth management methodology that is not able to provide a full Diffserv support. The reason of this is that the RMD-QOSM concept can only support the (Expedited Forwarding) EF-like functionality behavior, where the use bandwidth as a signaled <QoS Desired> parameter is required. The RMD-QOSM is not able to support the full set of (Assured Forwarding) AF-like functionality where multiple PHBs/DSCPs are used. This is because the signaled <QoS Desired> parameter should contain two token buckets needed to signal AF in full generality. Note however, that RMD-QOSM could also support a single AF PHB, as far as the traffic or the upper limit of the traffic can be characterized by a single bandwidth parameter.

A very important consideration on using RMD-QOSM is that within one RMD domain only one of the following RMD-QOSM schemes can be used at a time. Thus a RMD router can never process and use two different RMD-QOSM signaling schemes at the same time.

The available RMD-QOSM signaling schemes are:

* per flow congestion notification based on probing (see Sections 4.3.2, 4.6.1.7, 4.6.2.6). Note that this scheme uses for severe congestion handling the Severe congestion handling by proportional data packet marking, see Section 4.6.1.6.2, 4.6.2.5.2)

* per flow RMD NSIS measurement based admission control (see Sections 4.3.2, 4.6.1, 4.6.2). Note that this scheme uses for severe congestion handling the Severe congestion handling by proportional data packet marking, see Section 4.6.1.6.2, 4.6.2.5.2)

* per aggregate RMD NSIS measurement based admission control (see Sections 4.3.1, 4.3.2, 4.6.1, 4.6.2). Note that this scheme uses for severe congestion handling the Severe congestion handling by proportional data packet marking, see Section 4.6.1.6.2, 4.6.2.5.2)

* per flow RMD reservation based in combination with severe congestion handling by the RMD-QOSM refresh procedure (see Sections 4.3.1, 4.3.3, 4.6.1, 4.6.1.6.1, 4.6.2.5.1). Note that this scheme uses for severe congestion handling the Severe congestion handling by the RMD-QOSM refresh procedure, see Section 4.6.1.6.1, 4.6.2.5.1)

* per flow RMD reservation based in combination with severe congestion handling by proportional data packet marking procedure (see Sections 4.3.1, 4.3.3, 4.6.1, 4.6.1.6.2, 4.6.2.5.2). Note that this scheme uses for severe congestion handling the Severe congestion handling by proportional data packet marking procedure, see Section 4.6.1.6.2, 4.6.2.5.2)

* per aggregate RMD reservation based in combination with severe congestion handling by the RMD-QOSM refresh procedure (see Sections 4.3.1, 4.3.3, 4.6.1, 4.6.1.6.1, 4.6.2.5.1). Note that this scheme uses for severe congestion handling the Severe congestion handling by the RMD-QOSM refresh procedure, see Section 4.6.1.6.1, 4.6.2.5.1)

* per aggregate RMD reservation based in combination with severe congestion handling by proportional data packet marking procedure (see Sections 4.3.1, 4.3.3, 4.6.1, 4.6.1.6.2, 4.6.2.5.2). Note that this scheme uses for severe congestion handling the Severe congestion handling by proportional data packet marking procedure, see Section 4.6.1.6.2, 4.6.2.5.2)

4. RMD-QOSM, Detailed Description

This section describes the RMD-QOSM in more detail. In particular, it defines the role of stateless and reduced-state QNEs, the RMD-QOSM QSpec Object, the format of the RMD-QOSM QoS-NSLP messages and how QSpecs are processed and used in different protocol operations.

4.1. RMD-QSpec Definition

The QSPEC format is specified in [QSP-T] and is as follows:

QSPEC = <QSPEC Version> <QOSM ID> <QSPEC Control Information> <QoS Description>

The <QSPEC Version> and <QoSM ID> used by the RMD-QOSM are assigned by IANA, see Section 6. The <QSPEC Control Information> contains the following fields:

<QSPEC Control Information> = <PHR container> <PDR container>

The Per Hop Reservation container (PHR container) and the Per Domain Reservation container (PDR container) are specified in Section 4.1.2 and 4.1.3, respectively. The <PHR container> contains the QoS specific control information for intra-domain communication and reservation. The <PDR container> contains additional control information that is needed for edge-to-edge communication.

The <QoS Description> contains the <RMD-QOSM QoS description field> that is specified in Section 4.1.1. The <RMD-QOSM QoS Description> field, the <PHR container> are used and processed by the Edge and Interior nodes. The <PDR container> field is only processed by Edge nodes.

4.1.1. RMD-QOSM QoS Description

The RMD-QOSM QoS Description carried by the RESERVE message only contains the QoS Desired object [QSP-T]. The QoS Reserved object is carried by the RESPONSE message.

<RMD-QOSM QoS Description> = <QoS Desired> for RESERVE

<RMD-QOSM QoS Description> = <QoS Reserved> for RESPONSE

<QoS Desired> = <Bandwidth> <PHB Class> <Admission Priority>

<QoS Reserved> = <Bandwidth> <PHB Class> <Admission Priority>
The bit format of the <Bandwidth> (see Figure 4), <PHB Class> (see Figure 5) and <Admission Priority> complies to the bit format specified in [QSP-T]. Note that for the RMD-QOSM a reservation established without an <Admission Priority> parameter is equivalent to a reservation established with an <Admission Priority> whose value is 1.

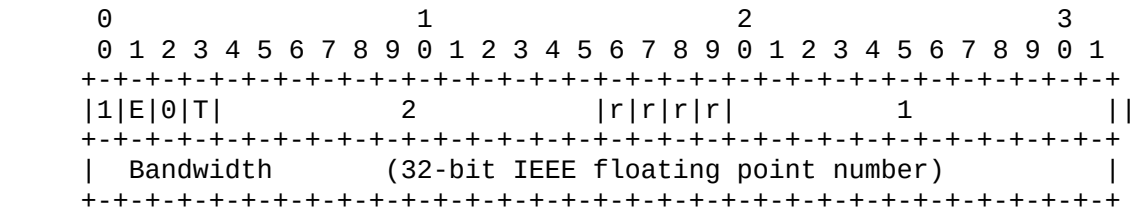

Figure 4: Bandwidth parameter

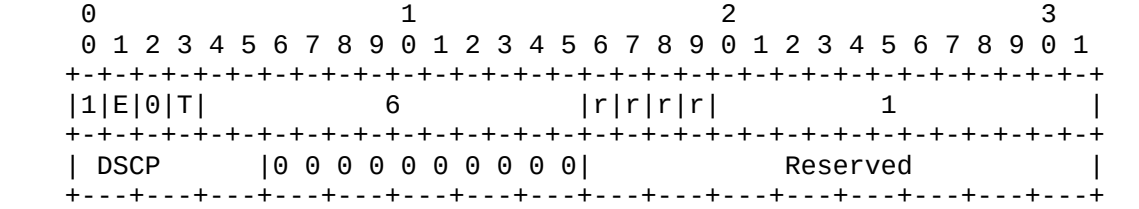

Figure 5: PHB_Class parameter

4.1.2. PHR Container

This section describes the parameters used by the PHR container.

<PHR container> = <Overload %>, <S>,<M>,
<Admitted Hops>, <B>, <Hop_U> <Time Lag>

The bit format of the PHR container can be seen in Figure 6. Note that in Figure 6 <Hop U> is represented as <U>.

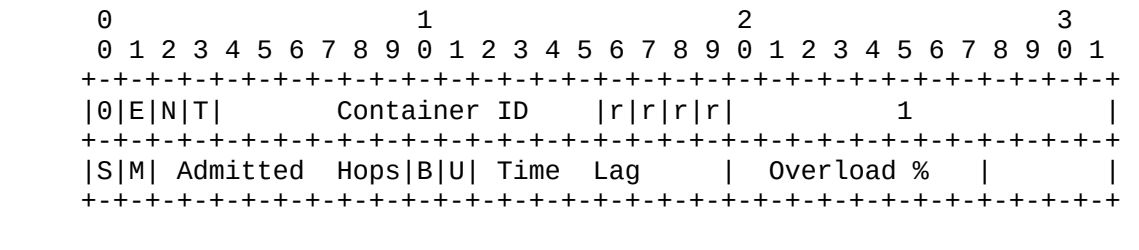

Figure 6: PHR container

Parameter/Container ID:
8 bit field, indicating the PHR type: PHR_Resource_Request, PHR_Release_Request, PHR_Refresh_Update. It is used to further specify QoS-NSLP RESERVE and RESPONSE messages.

"PHR_Resource_Request" (Container ID = 1): initiate or update the traffic class reservation state on all nodes located on the communication path between the QNE(Ingress) and QNE(Egress) nodes.

"PHR_Refresh_Update" (Container ID = 2): refresh the traffic class reservation soft state on all nodes located on the communication path between the QNE(Ingress) and QNE(Egress) nodes according to a resource reservation request that was successfully processed during a previous refresh period.

"PHR_Release_Request" (Container ID = 3): explicitly release, by subtraction, the reserved resources for a particular flow from a traffic class reservation state.

<S> (Severe Congestion):
1 bit. In case of a route change refreshing RESERVE messages follow the new data path, and hence resources are requested there. If the resources are not sufficient to accommodate the new traffic sever congestion occurs. Congested Interior nodes SHOULD notify Edge QNEs about the congestion by setting the S bit.

<Overload %>:
8 bits In case of severe congestion the level of overload is indicated by the Overload %. Overload % SHOULD be higher than 0 if S bit is set. If overload in a node is greater than the overload in a previous node then Overload % SHOULD be updated.

<M>:
1 bit. In case of unsuccessful resource reservation or resource query in an Interior QNE, this QNE sets the M bit in order to notify the Egress QNE.

<Admitted Hops>:
8 bit field. The <Admitted Hops> counts the number of hops in the RMD domain where the reservation was successful. The <Admitted Hops> is set to "0" when a RESERVE message enters a domain and it is increased by one at each Interior QNE. However when a QNE that does not have sufficient resources to admit the reservation is reached, the M Bit is set, and the <Admitted Hops> value is frozen.

<Hop_U> (NSLP_Hops unset):
1-bit. The QNE(Ingress) node MUST set the <Hop_U> parameter to 0. This parameter SHOULD be set to "1" by a node when the node does not increase the <Admitted Hops> value. This is the case when an RMD-QOSM reservation-based node is not admitting the reservation request. When <Hop_U> is set "1" the <Admitted Hops> SHOULD NOT be changed.

<B>: 1 bit. Indicates bi-directional reservation.

<Time Lag>: 8 bit field. The time lag used in a sliding window over the refresh period.

4.1.3. PDR container

This section describes the parameters of the PDR container.

The bit format of the PDR container can be seen in Figure 7.

<PDR container> = <Overload %> <S> <M> <Max Admitted Hops> <B> [<PDR Reverse Requested Resources>]

Note that in Figure 7 <Max Admitted Hops> is represented as <Max Adm Hops>.

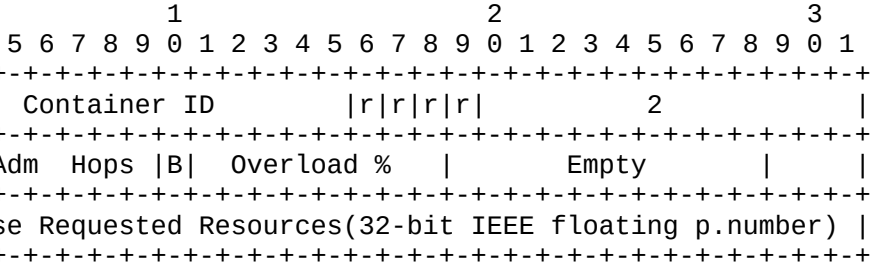

Figure 7: PDR container

Parameter/Container ID:

8-bit field identifying the type of PDR container field.

"PDR_Reservation_Request" (Parameter/Container ID = 4): generated by the QNE(Ingress) node in order to initiate or update the QoS-NSLP per domain reservation state in the QNE(Egress) node

"PDR_Refresh_Request" (Parameter/Container ID = 5): generated by the QNE(Ingress) node and sent to the QNE(Egress) node to refresh, in case needed, the QoS-NSLP per domain reservation states located in the QNE(Egress) node

"PDR_Release_Request" (Parameter/Container ID = 6): generated and sent by the QNE(Ingress) node to the QNE(Egress) node to release the per domain reservation states explicitly

"PDR_Reservation_Report" (Parameter/Container ID = 7): generated and sent by the QNE(Egress) node to the QNE(Ingress) node to report that a "PHR_Resource_Request" and a "PDR_Reservation_Request" control information fields have been received and that the request has been admitted or rejected

"PDR_Refresh_Report" (Parameter/Container ID = 8) generated and sent by the QNE(Egress) node in case needed, to the QNE(Ingress) node to report that a "PHR_Refresh_Update" control information field has been received and has been processed

"PDR_Release_Report" (Parameter/Container ID = 9) generated and sent by the QNE(Egress) node in case needed, to the QNE(Ingress) node to report that a "PHR_Release_Request" and a "PDR_Release_Request" control information fields have been received and have been processed.

"PDR_Congestion_Report" (Parameter/Container ID = 10): generated and sent by the QNE(Egress) node to the QNE(Ingress) node and used for congestion notification

<S> (PDR Severe Congestion):
1-bit. Specifies if a severe congestion situation occurred. It can also carry the <S> parameter of the "PHR_Resource_Request" or "PHR_Refresh_Update" fields.

<Overload %>:
8-bit. It includes the Overload % of the "PHR_Resource_Request" or "PHR_Refresh_Update" control information fields, indicating the level of overload to the Ingress node.

<M> (PDR Marked):
1-bit. Carries the <M> value of the "PHR_Resource_Request" or "PHR_Refresh_Update" control information fields.

<B>: 1 bit Indicates bi-directional reservation.

<Max Admitted Hops>:
8-bit. The <Admitted Hops> value that has been carried by the PHR container field used to identify the RMD reservation based node that admitted or process a "PHR_Resource_Request"

<PDR Reverse Requested Resources>:
32 bits. This field only applies when the "B" flag is set to "1". It specifies the requested number of units of resources that have to be reserved by a node in the reverse direction when the intra-domain signaling procedures require a bi-directional reservation procedure.

4.2. Message Format

The format of the messages used by the RMD-QOSM complies with the QoS-NSLP specification. As specified in [QoS-NSLP], for each QoS-NSLP message type, there is a set of rules for the permissible choice of object types. These rules are specified using Backus-Naur Form (BNF) augmented with square brackets surrounding optional sub-sequences. The BNF implies an order for the objects in a message. However, in many (but not all) cases, object order makes no logical difference. An implementation SHOULD create messages with the objects in the order shown here, but accept the objects in any permissible order.

The format of a local (intra-domain) RESERVE message used by the RMD-QOSM is:

```
RESERVE = COMMON_HEADER
        RSN [ RII ] [ REFRESH_PERIOD ]
        [ *BOUND_SESSION_ID ]
        [[ PACKET_CLASSIFIER ] [ RMD-QSPEC ]]
```

The format of an intra-domain Query message that may be used by the RMD-QOSM is as follows:

```
QUERY = COMMON_HEADER
    [ RII ] [ *BOUND_SESSION_ID ]
    [ PACKET_CLASSIFIER ] RMD-QSPEC
```

A QUERY message MUST contain an RII object to indicate a RESPONSE is desired, unless the QUERY is being used to initiate reverse-path state for a receiver-initiated reservation.

The format of a local (intra-domain) RESPONSE message used by the RMD-QOSM is as follows:

```
intra-domain RESPONSE = COMMON_HEADER
              [ RII / RSN ] INFO_SPEC [ RMD-QSPEC ]
```

The format of an end-to-end RESPONSE message that is used by the RMD-QOSM to carry an intra-domain RMD-QSPEC object is as follows:

```
RESPONSE = COMMON_HEADER [RII/RSN] INFO_SPEC [QSPEC] [RMD-QSPEC]
```

The format of an intra-domain NOTIFY message used by the RMD-QOSM is as follows:

```
NOTIFY = COMMON_HEADER INFO_SPEC [ RMD-QSPEC ]
```

The format of an end-to-end NOTIFY message that is used by the RMD-QOSM to carry an intra-domain RMD-QSPEC object is as follows:

```
NOTIFY = COMMON_HEADER INFO_SPEC [QSPEC] [RMD-QSPEC]
```

All objects, except RMD-QSPEC objects, are specified in [QoS-NSLP].

4.3. RMD node state management

The QoS-NSLP state creation and management is specified in [QoS-NSLP]. This section describes the state creation and management functions of the Resource Management Function (RMF) in the RMD nodes.

4.3.1 Aggregated versus per flow reservations at the QNE Edges

The QNE Edges maintain for the RMD QoS model either per flow, or aggregated QoS-NSLP reservation states. Each per flow or aggregated QoS-NSLP reservation state, associated with the RMD-QOS model, is identified by a NTLP SESSION_ID (see [GIST]). In RMD, these states are denoted as PDR states.

When the QNE Edges use aggregated QoS-NSLP reservation states the SESSION_ID of the aggregated state, the IP addresses of the Ingress and Egress nodes, the PHB value and the size of the aggregated reservation, e.g., reserved bandwidth have to be maintained.

The size of the aggregation is specified in Section 1.4.4 of [RFC3175]. The size of the aggregated reservations needs to be greater or equal to the sum of bandwidth of the inter domain (end-to-end) reservations it aggregates. A policy can be used to maintain the amount of required bandwidth on a given aggregated reservation by taking into account the sum of the underlying inter domain (end-to-end) reservations, while endeavoring to change reservation less frequently. This MAY require a trend analysis. If there is a significant probability that in the next interval of time the current aggregated reservation is exhausted, the Ingress router MUST predict the necessary bandwidth and request it. If the Ingress router has a significant amount of bandwidth reserved but has very little probability of using it, the policy MAY predict the amount of bandwidth required and release the excess. To increase or decrease the aggregate, the RMD modification procedures SHOULD be used (see Section 4.6.1.4).

4.3.2 Measurement-based method

The measurement-based method can be classified in two schemes:

* Congestion notification based on probing:

In this scheme the interior nodes are Diffserv aware but not NSIS aware nodes. Each interior node counts the bandwidth that is used by each PHB traffic class. This counter value is stored in an RMD_QOSM state. For each traffic belonging to a PHB traffic class a predefined congestion threshold is set. The predefined congestion notification threshold is set according to, an engineered bandwidth limitation based on e.g. agreed Service Level Agreement or a capacity limitation of specific links. The threshold is usually less than the capacity limit, i.e., admission threshold, in order to avoid congestion due to the error of estimating the actual traffic load. The value of this threshold SHOULD be stored in another RMD_QOSM state.

In this scenario end-to-end NSIS message is used as a probe packet. In this case the DSCP field of the GIST message is re-marked when the predefined congestion notification threshold is exceeded in an interior node. Note that in this situation, in addition to the probe packet, also ordinary data packets passing though the congested node are re-marked. The rate of the re-marked data packets is used to detect a congestion situation that can influence the admission control decissions.

* NSIS measurement-based admission control:

The measurement based admission control is implemented in NSIS aware stateless routers. In particular, the QNE Interior nodes operating in NSIS measurement-based mode are QoS-NSLP stateless nodes, i.e., they do not support any QoS-NSLP or NTLP/GIST states. These measurement-based nodes store two RMD-QOSM states per PHR group. These states reflect the traffic conditions at the node and are not affected by QoS-NSLP signaling. One state stores the measured user traffic load associated with the PHR group and another state stores the maximum traffic load threshold that can be admitted per PHR group. When a measurement-based node receives a local RESERVE message, it compares the requested resources to the available resources (maximum allowed minus current load) for the requested PHR group. If there are insufficient resources, it sets the <M> bit in the RMD-QSpec. No change to the RMD-QSpec is made when there are sufficient resources.

4.3.3 Reservation-based method

QNE Interior nodes operating in reservation-based mode are QoS-NSLP reduced state nodes, i.e., they do not store NTLP/GIST states but they do store per PHB-aggregated QoS-NSLP states.

The reservation-based PHR installs and maintains one reservation state per PHB, in all the nodes located in the communication path from the QNE Ingress node up to the QNE Egress node. This state represents the number of currently reserved resource units. Thus, the QNE Ingress node signals only the resource units requested by each flow. These resource units, if admitted, are added to the currently reserved resources per PHB.

For each PHB a threshold is maintained that specifies the maximum number of resource units that can be reserved. This threshold could, for example, be statically configured. An example of how the admission control and its maintenance process occurs in the interior nodes is described in Section 3 of [CsTa05]. The simplified concept that is used by the per traffic class admission control process, is based on the following equation:

last + p <= T,

where p: requested bandwidth rate, T: admission threshold, which reflects the maximum traffic volume that can be admitted in the traffic class, last: a counter that records the aggregated sum of the signaled bandwidth rates of previous admitted flows.

The per-PHB group reservation states are soft states, which are refreshed by sending periodic refresh local RESERVE messages. If a refresh message corresponding to a number of reserved resource units is not received, the aggregated reservation state is decreased in the next refresh period by the corresponding amount of resources that were not refreshed. The refresh period can be refined using a sliding window algorithm described in [RMD3].

The reserved resources for a particular flow can also be explicitly released from a PHB reservation state by means of a PHR release message. The usage of explicit release enables the instantaneous release of the resources regardless of the length of the refresh period. This allows a longer refresh period, which also reduces the number of periodic refresh messages.

Note that both in case of measurement- and reservation-based methods, the way of how the maximum bandwidth thresholds are maintained is out of the specification of this document. However, when admission priorities are supported, the Maximum Allocation [RFC4125] or the Russian Dolls [RFC4127] bandwidth allocation model may be used. In this case three types of priority traffic classes within the same PHB, e.g., Expedited Forwarding, can be differentiated. These three different priority traffic classes, which are associated to the same PHB, are denoted in this document as PHB_low_priority, PHB_normal_priority and PHB_high_priority.

4.4. Transport of RMD-QOSM messages

The intra-domain (local) messages used by the RMD-QOSM MUST operate in the NTLP/GIST Datagram mode (see [GIST]). Therefore, the NSLP functionality available in all QoS NSLP nodes that are able to support the RMD-QOSM MUST require the intra-domain GIST functionality available in these nodes to operate in the datagram mode, i.e., require GIST to:

* operate in unreliable mode. This can be satisfied by passing this requirement from the QoS-NSLP layer to the GIST layer via the API transfer-attributes.

* do not create a message association state. This requirement can be satisfied by a local policy, e.g., the QNE is configured to do not create a message association state

* do not create any NTLP routing state. This can be satisfied by passing this requirement from the QoS-NSLP layer to the GIST layer via the API.

All the intra-domain local messages are transported using the GIST data messages (see [GIST]). At the ingress the original (end-to-end) RESERVE message is forwarded but ignored by the stateless or reduced-state nodes, see Figure 3. The intermediate (interior) nodes are bypassed using multiple levels of the router alert option (see [QoS-NSLP]. In that case, interior routers are configured to handle only certain levels of router alert (RAO) values. This is accomplished by marking the end-to-end RESERVE message, i.e., modifying the QoS-NSLP default NSLP-ID value to another NSLP-ID predefined value.

The marking MUST be accomplished by the ingress by modifying the QoS_NSLP default NSLP-ID value to a NSLP-ID predefined value. In this way the egress MUST stop this marking process by reassigning the QoS-NSLP default NSLP-ID value to the original (end-to-end) RESERVE message. Note that the assignment of these NSLP-ID values is a QOS-NSLP issue, which should be accomplished via IANA.

4.5 Edge discovery and message addressing

Mainly, the Egress node discovery can be performed either by using the GIST discovery mechanism [GIST], manual configuration or any other discovery technique. The addressing of signaling messages depends on the used GIST transport mode. The RMD QoS signaling messages that are processed only by the Edge nodes use the peer-peer addressing of the GIST connection (C) mode. RMD QoS signaling messages that are processed by all nodes of the Diffserv domain,i.e., Edges and Interior nodes, use the end-end addressing of the GIST datagram (D) mode. RMD QoS signaling messages that are addressed to the data path end nodes are intercepted by the Egress nodes.

4.6. Operation and sequence of events

4.6.1. Basic unidirectional operation

This section describes the basic unidirectional operation and sequence of events of the RMD-QOSM. The following basic operation cases are distinguished:

* Successful reservation (Section 4.6.1.1),
* Unsuccessful reservation (Section 4.6.1.2),
* RMD refresh reservation (Section 4.6.1.3),
* RMD modification of aggregated reservation (4.6.1.4)
* RMD release procedure (Section 4.6.1.5)
* Severe congestion handling (Section 4.6.1.6)
* Admission control using congestion notification based on probing (Section 4.6.1.7).

The QNEs at the Edges of the RMD domain support the RMD QoS Model and end-to-end QoS models, which process the RESERVE message differently. Note that the term end-to-end QoS model applies to any QoS model that is initiated and terminated outside the RMD-QOSM aware domain. However, there might be situations where a QoS model is initiated and/or terminated by the QNE Edges and is considered to be an end-to-end QoS model. This can occur when the QNE Edge can also operate as a QNI or as a QNR. Note that the described functionality applies to the RMD reservation-based and to the NSIS measurement-based admission control methods. The QNE Edge nodes maintain either per flow QoS-NSLP reservation states or aggregated QoS-NSLP reservation states. When the QNE Edges maintain aggregated QoS-NSLP reservation states, the RMD-QOSM functionality may accomplish a RMD modification procedure (see Section 4.6.1.4), instead of the reservation initiation procedure that is described in this subsection.

4.6.1.1. Successful reservation

This section describes the operation of the RMD-QOSM where a reservation is successfully accomplished.

The QNI generates the initial RESERVE message, and it is forwarded by the NTLP as usual [GIST].

4.6.1.1.1. Operation in Ingress node

When an end-to-end reservation request (RESERVE) arrives at the Ingress node (QNE), see Figure 8, it is processed based on the end-to-end QoS model. Note that when the QOSM ID of the end-to-end QoS model is not known to the Ingress node (QNE), the Ingress MUST interpret at least the mandatory parameters (see [QSP-T]). If the QSPEC object contains also optional parameters that are not used by the RMD-QOSM, then the N-flag of each of these objects MUST be set. Subsequently, the RMD QoS Description: <Bandwidth> and <PHB Class> are derived from the QoS Description of the end-to-end QSpec. The Ingress QNE performs then the following functionality.

If the requested <Bandwidth> parameter cannot be satisfied locally, then an end to end RESPONSE message has to be generated. An end-to-end QSpec object MUST be included in the RESPONSE message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values. The "E" flag associated with the QSPEC <QoS Reserved> object and the "E" flag associated with the <Bandwidth> parameter are set. In addition, the INFO-SPEC object is included in the end to end RESPONSE message. The error code used by this INFO-SPEC is:

Error severity class: 0x04 Transient Failure
Error code value: 0x07 Total reservation failure

Furthermore, all the other RESPONSE parameters are set according to the end-to-end QoS model or according to [QoS-NSLP] and [QSP-T].

If the request was satisfied locally (see Section 4.3), the Ingress QNE node generates two RESERVE messages: one intra-domain and one end-to-end RESERVE messages. These are bound together in the following way. The end-to-end RESERVE SHOULD contain in the BOUND_SESSION_ID the SESSION_ID of its bound intra-domain session. Furthermore, if the QNE Edge nodes maintain intra-domain per flow QoS-NSLP reservation states then the value of Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions). If the QNE Edge nodes maintain intra-domain aggregated QoS-NSLP reservation states then the value of Binding_Code MUST be set to 0x03 (Aggregate sessions).

The intra-domain RESERVE SHOULD contain in the BOUND_SESSION_ID the SESSION_ID of its bound end-to-end session. The value of the Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions). Note that the end to end RESERVE is tunneled within the RMD domain. Therefore, the T-flag of the QSPEC parameters has to be processed/set according to the [QSP-T] specification.

The intra-domain RESERVE message is associated with the (local NTLP) SESSION_ID mentioned above. The selection of the IP source and IP destination address of this message depends on how the different inter-domain (end-to-end) flows are aggregated by the QNE Ingress node (see Section 4.3.1). As described in Section 4.3.1, the QNE Edges maintain either per flow, or aggregated QoS-NSLP reservation states for the RMD QoS model, which are identified by (local NTLP) SESSION_IDs (see [GIST]). Note that this NTLP SESSION ID is a different one than the SESSION_ID associated with the end-to-end RESERVE message.

If no QOS-NSLP aggregation procedure at the QNE Edges is possible then the IP source and IP destination address of this message MUST be equal to the IP Source and IP destination addresses of the data flow. The intra-domain RESERVE message MUST be sent using the NTLP datagram mode (see Section 4.4). In addition, the intra-domain RESERVE (RMD-QSPEC) message MUST include a PHR container (PHR_Resource_Request) and the "RMD QOS Description" field.

The end-to-end RESERVE message includes the end-to-end QSpec and it is sent towards the Egress QNE. If the end-to-end QSpec does not carry an RII object, then an RII object has to be generated and included into the end-to-end RESERVE message.

Note that after completing the initial discovery phase, the GIST connection mode can be used between the QNE Ingress and QNE Egress. The end-to-end RESERVE message is forwarded using the GIST forwarding procedure to bypass the Interior stateless or reduced-state QNE nodes, see Figure 8. The bypassing procedure is described in Section 4.4. At the QNE Ingress the end-to-end RESERVE message is marked, i.e., modifying the QoS-NSLP default NSLP-ID value to another NSLP-ID predefined value, which corresponds to a RAO value that will be used by the GIST message carrying the end-to-end RESPONSE message to bypass the QNE Interior nodes. Note that the QNE Interior nodes, see [GIST], are configured to handle only certain levels of router alert (RAO) values.

Furthermore, note that the initial discovery phase and the process of sending the end-to-end RESERVE message towards the QNE Egress MAY be done simultaneously.

The (initial) intra-domain RESERVE message MUST be sent by the QNE Ingress and it MUST contain the following values:

* the value of the <RSN> object SHOULD be the same as the value of the RSN object of the end-to-end RESERVE message;

* the value of the <BOUND_SESSION_ID> object MUST be the SESSION_ID associated to the end-to-end RESERVE message. Furthermore, if the QNE Edge nodes maintain per flow QoS-NSLP reservation states then the value of Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions).

* the SCOPING flag MUST not be set, meaning that a default scoping of the message is used. Therefore, the QNE Edges MUST be configured as boundary nodes and the QNE Interior nodes MUST be configured as Interior (intermediary) nodes;

* The <RII> object MUST not included in this message;

* The flag REPLACE MUST be set to FALSE = 0;

* the value of the <REFRESH_PERIOD> object MUST be calculated and set by the QNE Ingress node, see also Section 4.6.1.3;

* the value of the <PACKET_CLASSIFIER> object SHOULD be associated with the path-coupled routing MRM. The flag that has to be set is the flag T (traffic class) meaning that the packet classification of packets is based on the DSCP value included in the IP header of the packets. Note that the DSCP value SHOULD be obtained from the MRI values obtained from GIST.

* the PHR resource units MUST be included into the <Bandwidth> parameter of the "RMD QoS Description" field;

* the value of the Parameter/Container ID field of the PHR container MUST be set to 1, (i.e., PHR_Resource_Request;)

* the value of the <Admitted Hops> parameter in the PHR container MUST be set to "1";

* the value of the <Hop_U> parameter in the PHR container MUST be set to "0";

* If the end-to-end RESERVE message carried an <Admission Priority> parameter, then this parameter should be copied and carried by the (initiating) intra-domain RESERVE. Note that for the RMD-QOSM a reservation established without an <Admission Priority> parameter is equivalent to a reservation with Admission Priority value 1. Note that in this case each admission priority is associated with a priority traffic class. The three priority traffic classes (PHB_low_priority, PHB_normal_priority, PHB_high_priority) may be associated with the same PHB.

* In a single-domain case the PDR container MAY not be included into the message.

When an end-to-end RESPONSE(PDR) message is received by the QNE Ingress node, the RMD-QSPEC, see Section 4.6.1.1.3, has to be identified, processed and removed from the end-to-end RESPONSE message. The QoS-NSLP state in the QNE Ingress stores and maintains the binding between each end-to-end session and each intra-domain session. In this way the QNE Ingress can match the PHR container that has been carried by the intra-domain RESERVE with the received PDR container that has been carried by the end-to-end RESPONSE message.

The RMD QoS model functionality is notified by reading the <M> parameter of the "PDR RMD control information" container that the reservation has been successful.

Furthermore, the INFO_SPEC object SHOULD be read by the QoS-NSLP functionality. In case of successful reservation the INFO_SPEC object SHOULD have the following values:

* Error Severity Class: 0x02 Success
* Error Code value: 0x01 Reservation successful

If the end-to-end RESPONSE message has to be forwarded to a node outside the RMD-QOSM aware domain then the non-default values of the objects contained in this message (i.e., <RII/RSN>, <INFO_SPEC>, [*QSPEC]) MUST be set by the QOS-NSLP protocol functions of the QNE.

4.6.1.1.2 Operation in the Interior nodes

Each QNE Interior node MUST use the QoS-NSLP and RMD-QOSM parameters of the intra-domain RESERVE (RMD-QSPEC) message as follows:

* the values of the <RSN>, <RII>, <PACKET_CLASSIFIER>, <REFRESH_PERIOD>, <BOUND_SESSION_ID> objects MUST NOT be changed. The interior node is informed by the <PACKET_CLASSIFIER> object that the packet classification should be done on the DSCP value. The value of the DSCP value SHOULD be obtained via the MRI parameters that the QoS-NSLP receives from GIST.
* The flag REPLACE MUST be set to FALSE = 0;

* the value of <Bandwidth> parameter of the "RMD QoS Description" field is used by the QNE Interior node for admission control, see Section 4.3.2 and Section 4.3.3. Furthermore, if the <Admission Priority> parameter is carried by the "RMD QoS Description" field this parameter is processed as described in the following bullet.

* in case of the RMD reservation-based procedure, and if these resources are admitted (see Section 4.3.3), they are added to the currently reserved resources. Furthermore, the value of the <Admitted Hops> parameter in the PHR container has to be increased by one.

* If the bandwidth allocated for the PHB_high_priority traffic is fully utilized, and a high priority request arrives, other policies can be used, which are beyond the scope of this document. One example for these policies can be that the high priority session is admitted through preemption of ongoing lower priority sessions, when the bandwidth reserved by the lower priority sessions can satisfy the high priority bandwidth request.. When the available bandwidth for the PHB_lower_priority and for the PHB_normal_priority is not enough to support the high priority traffic, then it will generate congestion for these PHB traffic classes. A solution to this congestion problem can be accomplished by using the severe congestion detection mechanism specified in Section 4.6.1.6.2.1. The degree of this congested bandwidth is indicated by using a specific DSCP (see Section 4.6.1.6.2.1) by marking the bytes proportionally to the degree of congestion. Other mechanisms may also be used as queues for the new high priority requests until capacity becomes available for the high priority sessions.

* in case of the RMD measurement based method, and if these resources are admitted (see Section 4.3.2), using a MBAC algorithm, the number of this resources will be used to update the MBAC algorithm.

4.6.1.1.3 Operation in the Egress node

When the end-to-end RESERVE message is received by the egress node, it is only forwarded further, towards QNR, if the processing of the intra-domain RESERVE(RMD-QSPEC) message was successful at all nodes in the RMD domain. In this case, the QNE Egress MUST stop the marking process that was used to bypass the QNE Interior nodes by reassigning the QoS-NSLP default NSLP-ID value to the end-to-end RESERVE message, see Section 4.4. Furthermore the carried BOUND_SESSION_ID object associated with the intra-domain session SHOULD be removed.
Note that the received end to end RESERVE was tunneled within the RMD domain. Therefore, the T-flag of the QSPEC parameters has to be processed/set according to the [QSP-T] specification.

If the the processing of the intra-domain RESERVE(RMD-QSPEC) was not successful at all nodes in the RMD domain then the inter domain (end-to-end) reservation is considered as being failed. Furthermore, note that the Egress should use a timer, that uses a pre-configured value, which can be used to synchronize the arrival of the end to end RESERVE and the intra-domain RESERVE (RMD-QSPEC) messages. If these two messages do not arrive during the time defined by the timer, then the reservation is considered as being failed. In this case a RESPONSE message is sent towards the QNE ingress with the following INFO_SPEC values:

Error Class: 0x04 Transient Failure
Error Code: 0x05 Mismatch synchronization between end-to-end RESERVE and intra-domain RESERVE

When the intra-domain RESERVE(RMD-QSPEC) is received by the QNE Egress node of the session associated with the intra-domain RESERVE(RMD-QSPEC) (the PHB session) with the session included in its <BOUND_SESSION_ID> object MUST be bound. The session included in the <BOUND_SESSION_ID> object is the session associated with the end-to-end RESERVE message.
Note that if the QNE Edge nodes maintain per flow QoS-NSLP reservation states then the value of Binding_Code = 0x01 (Tunnel and end-to-end sessions) is used.

Note that when the interior nodes are using mechanisms to admit high priority session through preemption of ongoing lower priority sessions, the mechanisms of solving the congestion on a low priority traffic PHB may use the solution specified in Section 4.6.1.6.2.2.

The QNE Egress MUST wait for the end-to-end RESPONSE message that has the same SESSION ID and RII object as the end-to-end RESERVE message forwarded towards QNR.

The non-default values of the objects contained in the end-to-end RESPONSE(PDR) message MUST be used and/or set by the QNE Egress as follows:

* the values of the <RII/RSN>, <INFO_SPEC>, [QSPEC] objects are set according to [QoS-NSLP] and/or [QSP-T].. The INFO_SPEC object SHOULD be set by the QoS-NSLP functionality. In case of successful reservation the INFO_SPEC object SHOULD have the following values:
  Error Severity Class: 0x02 Success,
  Error Code value: 0x01 Reservation successful,
  Furthermore, an end-to-end QSpec object MUST be included in the RESPONSE message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values.

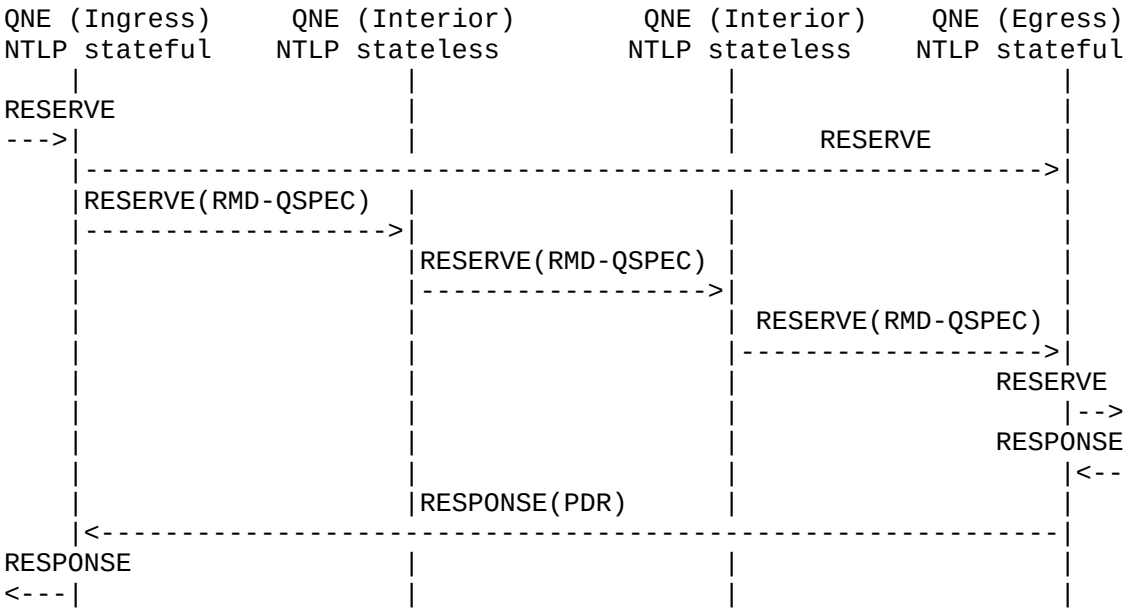

Figure 8: Basic operation of successful reservation procedure used by the RMD-QOSM

In addition to the above, the QNE Egress MUST also generate a RMD-QSPEC object that is carried by the end-to-end RESPONSE(PDR) message, see Section 4.2.

The following parameters of the RMD-QSPEC object MUST be used and/or set in the following way:

* the value of the Parameter/Container ID field of the PDR container MUST be set "7" (i.e., PDR_Reservation_Report);

* the value of the <M> field of the PDR container MUST be equal to the value of the <M> parameter of the PHR container that was carried by its associated intra-domain RESERVE(RMD-QSPEC) message.

The end-to-end RESPONSE(PDR) message is addressed and sent to its upstream QoS-NSLP neighbor, i.e., QNE Ingress node. Note that for all upstream messages the RAO is not set. Therefore, all Interior nodes ignore the end-to-end RESPONSE messages.

4.6.1.2. Unsuccessful reservation

This section describes the operation where a request for reservation cannot be satisfied by the RMD-QOSM.

The QNE Ingress, the QNE Interior and QNE Egress nodes process and forward the end-to-end RESERVE message and the intra-domain RESERVE(RMD-QSPEC) message in the same way as specified in Section 4.6.1.1. The main difference between the unsuccessful operation and successful operation is that one of the QNE nodes does not admit the request due to lack of resources. This also means that the QNE edge node MUST NOT forward the end-to-end RESERVE message towards the QNR node.

Note that the described functionality applies to the RMD reservation-based and to the NSIS measurement-based admission control methods. The QNE Edge nodes maintain either per flow QoS-NSLP reservation states or aggregated QoS-NSLP reservation states. When the QNE edges maintain aggregated QoS-NSLP reservation states, the RMD-QOSM functionality may accomplish a RMD modification procedure (see Section 4.6.1.4), instead of the reservation initiation procedure that is described in this subsection.

4.6.1.2.1 Operation in the Ingress nodes

When an end-to-end RESERVE message arrives at the QNE Ingress and if there are no resources available locally, the QNE Ingress MUST reject this end-to-end RESERVE message and sends a RESPONSE message back to the sender, as described in the QoS-NSLP specification, see [QoS-NSLP] and [QSP-T].

When an end-to-end RESPONSE(PDR) message is received by an Ingress node, see Section 4.6.1.2.3, the following actions take place. The non-default values of the objects contained in the end-to-end RESPONSE (PDR) message MUST be used and/or set by the QNE Ingress node as follows:

* the values of the <RII/RSN>, [<INFO_SPEC>], [QSPEC] objects are set according to the QoS-NSLP procedures. Furthermore, the INFO_SPEC object, generated by the Egress is read by the QoS-NSLP functionality.

* the RMD-QSPEC object, see Section 4.2, has to be processed and removed. The RMD Resource Management Function (RMF) is notified by reading the <M> parameter of the PDR container that the reservation has been unsuccessful. Note that when the QNE edges maintain a per flow QoS-NSLP reservation state the RMD-QOSM functionality, has to start an RMD release procedure (see Section 4.6.1.5). When the QNE edges maintain aggregated QoS-NSLP reservation states the RMD-QOSM functionality MAY start a RMD modification procedures (see Section 4.6.1.4).

4.6.1.2.2 Operation in the Interior nodes

In case of the RMD reservation based scenario, and if the intra-domain reservation request is not admitted by the QNE Interior node then the <Hop_U> and <M> parameters of the PHR container MUST be set to "1". The <Admitted Hops> counter MUST NOT be increased. Furthermore, the "E" flag associated with the QSPEC <QoS Desired> object and the "E" flag associated with the <Bandwidth> parameter SHOULD be set. In case of the RMD measurement based scenario, the <M> parameter of the PHR container MUST be set to "1". Furthermore, the "E" flag associated with the QSPEC <QoS Desired> object and the "E" flag associated with the <Bandwidth> parameter SHOULD be set.

In general, if a QNE Interior node receives a QSpec <Bandwidth> parameter with the "E" flag set and a PHR container type "PHR_Resource_Request", with the <M> parameter set to "1", then this PHR container and the "RMD QoS Description" field MUST NOT be processed.

4.6.1.2.3 Operation in the Egress nodes

In the RMD reservation based and the RMD measurement based scenario, when the <M> marked intra-domain RESERVE(RMD-QSPEC) is received by the QNE Egress node (see Figure 9) the session associated with the intra-domain RESERVE(RMD-QSPEC) (the PHB session) and the session included in its BOUND_SESSION_ID object MUST be bound. The session in the <BOUND_SESSION_ID> object is the session associated with the end-to-end RESERVE.

The QNE Egress node MUST generate an end-to-end RESPONSE message that has to be sent to its previous stateful QoS-NSLP hop.

* the values of the <RII/RSN>, <INFO_SPEC> objects are set by the standard QoS-NSLP protocol functions. In case of unsuccessful reservation the INFO_SPEC object SHOULD have the following values:
  Error Severity Class: 0x04, Transient Failure
  Error Code value: 0x07 Total reservation failure

The QSpec that was carried by the end to end RESERVE belonging to the same session as this end-to-end RESPONSE is included in this message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values. The "E" flag associated with the QSPEC <QoS Reserved> object and the "E" flag associated with the <Bandwidth> parameter are set.

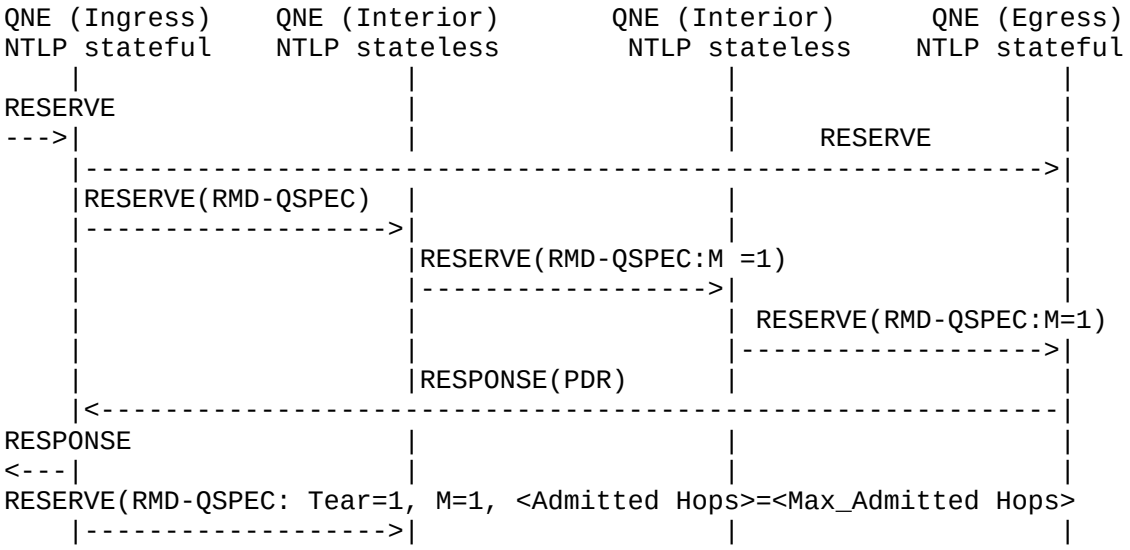

Figure 9: Basic operation during unsuccessful reservation initiation used by the RMD-QOSM

In addition to the above, similarly to the successful operation, see Section 4.6.1.1.3, the QNE Egress MUST also generate an RMD-QSPEC object that is carried by the end-to-end RESPONSE message.

The following fields of the RMD-QSPEC object MUST be used and/or set in the following way:

* the value of the <PDR Control Type> of the PDR container MUST be set to "7" (PDR_Reservation_Report);

* the value of the <Admitted Hops> parameter of the PHR container included in the received <M> marked PDR container MUST be included in the <Max_Admitted Hops> parameter of the PDR container;

* the value of the <M> parameter of the PDR container MUST be set to "1".

4.6.1.3 RMD refresh reservation

In case of RMD measurement-based method, QoS-NSLP states in the RMD domain are not maintained, therefore, the end-to-end RESERVE (refresh) message is sent directly to the QNE Egress.

The refresh procedure in case of RMD reservation-based method follows a similar scheme as the reservation process, shown in Figure 3. If the RESERVE messages arrive within the soft state time-out period, the corresponding number of resource units are not removed. However, the transmission of the intra-domain and end-to-end (refresh) RESERVE message are not necessarily synchronized. Furthermore, the generation of the end-to-end RESERVE message, by the QNE edges, depends on the locally maintained refreshed interval (see [QoS-NSLP]).

4.6.1.3.1 Operation in the Ingress node

The Ingress node MUST be able to generate an intra-domain (refresh) RESERVE(RMD-QSpec) at any time. Before generating this message, the RMD QoS signaling model functionality is using the RMD traffic class (PHR) resource units for refreshing the RMD traffic class state.

Note that the RMD traffic class refresh periods MUST be equal in all QNE edge and QNE Interior nodes and SHOULD be smaller (default: more than two times smaller) than the refresh period at the QNE Ingress node used by the end-to-end RESERVE message. The intra-domain RESERVE (RMD-QSPEC) message MUST include a "RMD QoS Description" field and a PHR container (i.e., PHR_Refresh_Update).

The selection of the IP source and destination address of this message depends on if and how the different inter domain (end-to-end) flows can be aggregated by the QNE Ingress node (see Section 4.3.1). Note that this QoS-NSLP aggregation procedure is different than the RMD traffic class aggregation procedure. One example is the approach used by the RSVP aggregation scenario ([RFC3175]), where the IP source address of this message is the IP address of the aggregator (i.e., QNE Ingress) and the IP destination address of this message is the IP address of the De-aggregator (i.e., QNE Egress). An alternative approach is the one used in "RSVP Refresh Overhead Reduction Extensions" ([RFC2961]). If no QOS-NSLP aggregation procedure at the QNE edges is possible then the IP source and IP destination address of this message MUST be equal to the IP source and IP destination addresses of the data flow.

An example of this RMD specific refresh operation can be seen in Figure 10.

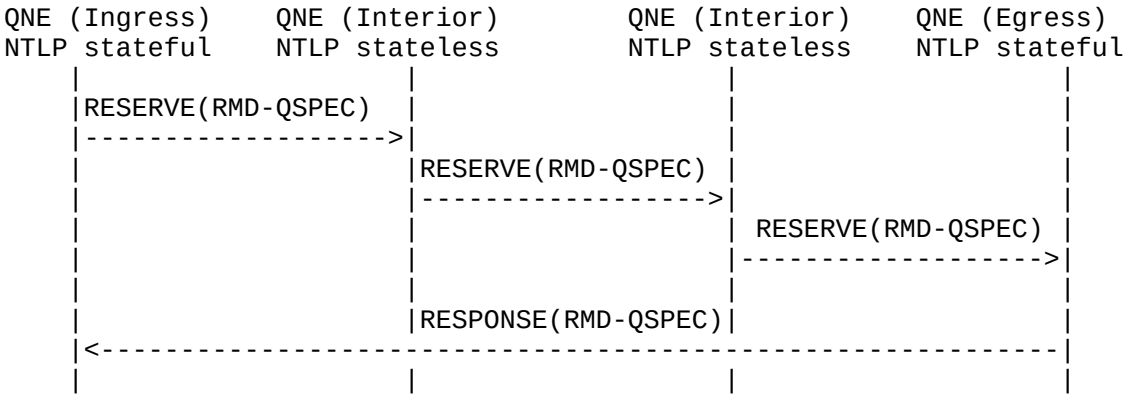

Figure 10: Basic operation of RMD specific refresh procedure

Most of the non-default values of the objects contained in this message MUST be used and set by the QNE Ingress in the same way as described in Section 4.6.1.1. The following objects are used and/or set differently:

* The flag REPLACE MUST be set to FALSE = 0;

* the PHR resource units MUST be included into the <Bandwidth> parameter. The value of the <Bandwidth> parameter depends on how the different inter domain (end-to-end) flows are aggregated by the QNE Ingress node (e.g., the sum of all the PHR requested resources of the aggregated flows). If no QOS-NSLP aggregation is accomplished by the QNE Ingress node, the value of the <Bandwidth> parameter SHOULD be equal to the <Bandwidth> parameter of its associated new (initial) intra-domain RESERVE (RMD-QSPEC) message;

* the value of the Parameter/Container field of the "PHR RMD-QOSM control information" container MUST be set to "2", i.e., "PHR_Refresh_Update";

* In a single-domain case the PDR container field is not needed in the message.

* the value of the <RII> object MUST contain the Response Identification Information value of the Ingress QNE, that is unique within a session and different for each message (see [QoS-NSLP]).

When the intra-domain RESPONSE (RMD-QSPEC) message, see Section 4.6.1.3.3., is received by the QNE Ingress node, then:

* the values of the <RII/RSN>, <INFO_SPEC>, [*QSPEC] objects are processed by the standard QoS-NSLP protocol functions (see Section 4.6.1.1);

* the PDR has to be processed and removed by the RMD-QOSM functionality in the QNE Ingress node. The RMD-QOSM functionality is notified by the <PDR M> parameter of the PDR container that the refresh procedure has been successful or unsuccessful. All session(s) (in case of the flow aggregation procedure there will be more than one sessions) associated with this RMD specific refresh session MUST be informed about the success or failure of the refresh procedure. In case of failure, the QNE Ingress node has to generate (in a standard QoS-NSLP way) an error end-to-end RESPONSE message that will be sent towards QNI.

4.6.1.3.2 Operation in the Interior node

The intra-domain RESERVE (RMD-QSPEC) message is received and processed by the QNE Interior nodes. Any QNE edge or QNE Interior node that receives a "PHR_Refresh_Update" control information field MUST identify the traffic class state (PHB) (using the <PHB Class> parameter). Most of the parameters in this refresh intra-domain RESERVE (RMD-QSPEC) message MUST be used and/or set by a QNE Interior node in the same way as described in Section 4.6.1.1.

The following objects are used and/or set differently:

* the value of <Bandwidth> parameter of the "RMD QoS Description" field is used by the QNE Interior node for refreshing the RMD traffic class state. These resources (included in <Bandwidth>), if reserved, are added to the currently reserved resources per PHB and therefore they will become a part of the per traffic class (per-PHB) reservation state, see Section 4.3.3. If the refresh procedure cannot be fulfilled then the <M> parameter of the PHR container MUST be set to "1". Furthermore, the "E" flag associated with <QoS Desired> object and the "E" flag associated with the <Bandwidth> parameter SHOULD be set.

Any PHR container of type "PHR_Refresh_Update", and its associated "RMD QoS Description" field (i.e., <Bandwidth>), whether it is marked or not and independent of the "E" flag value of the <Bandwdith> parameter, is always processed, but marked bits are not changed.

4.6.1.3.3 Operation in the Egress node

The intra-domain RESERVE(RMD-QSPEC) message is received and processed by the QNE Egress node. A new intra-domain RESPONSE (RMD-QSPEC) message is generated by the QNE Egress node and MUST include a PDR (type PDR_Refresh_Report).

The intra-domain RESPONSE (RMD-QSPEC) message MUST be sent to the QNE Ingress node, i.e., the previous stateful hop. The address of the QNE Ingress node can be found using the existing messaging association between the QNE Egress and QNE Ingress nodes. This state is associated with the end-to-end session and identified by the SESSION ID that is bound to the session of the intra-domain RESPONSE(RMD-QSPEC) message.

* the values of the <RII/RSN>, <INFO_SPEC> objects are set by the standard QoS-NSLP protocol functions.

* The value of the <PDR Control Type> parameter of the PDR container MUST be set "8" (i.e. PDR_Refresh_Report).
  In case of successful reservation the INFO_SPEC object SHOULD have the following values:
  Error Severity Class: 0x02, Success
  Error Code value: 0x01 Reservation successful

* In case of unsuccessful reservation the INFO_SPEC object SHOULD have the following values:
  Error Severity Class: 0x04, Transient Failure
  Error Code value: 0x07 Total reservation failure

The RMD-QSpec that was carried by the intra-domain RESERVE belonging to the same session as this intra-domain RESPONSE is included in the intra-domain RESPONSE message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values. If the reservation is unsuccessful then "E" flag associated with the QSPEC <QoS Reserved> object and the "E" flag associated with the <Bandwidth> parameter are set.

4.6.1.4. RMD modification of aggregated reservations

In the case when the QNE edges maintain QoS-NSLP aggregated reservation states and the aggregated reservation has to be modified (see Section 4.3.1) the following procedure is applied:

* When the modification request requires an increase of the reserved resources, the QNE Ingress node MUST include the corresponding value into the <Bandwidth> parameter of the "RMD QoS Description" field, which is sent together with a "PHR_Resource_Request" control information. If a QNE edge or QNE Interior node is not able to reserve the number of requested resources, the "PHR_Resource_Request" control information that is associated with the <Bandwidth> parameter MUST be marked. In this situation the RMD specific operation for unsuccessful reservation will be applied (see Section 4.6.1.2).

* When the modification request requires a decrease of the reserved resources, the QNE Ingress node MUST include this value into the <Bandwidth> parameter of the "RMD QoS Description" field. Subsequently an RMD release procedure SHOULD be accomplished (see Section 4.6.1.5).

4.6.1.5 RMD release procedure

If a refresh RESERVE message does not arrive at a QNE Interior node within the refresh time-out period then the resources associated with this message are removed. This soft state behavior provides certain robustness for the system ensuring that unused resources are not reserved for long time. Resources can be removed by an explicit release at any time.

When the RMD-RMF of a QNE edge or QNE Interior node processes a "PHR_Release_Request" control information it MUST identify the <PHB Class> parameter and estimate the time period that elapsed after the previous refresh, see also Section 3 of [CsTa05]. This MAY be done by indicating the time lag, say "T_lag", between the last sent "PHR_Refresh_Update" and the "PHR_Release_Request" control information container by the QNE Ingress node. The value of "T_Lag" is first normalized to the length of the refresh period, say "T_period". The ratio between the "T_Lag" and the length of the refresh period, "T_period", is calculated. This ratio is then introduced into the <Time Lag> field of the "PHR_Release_Request" control information.

When a node (QNE edge or QNE Interior) receives the "PHR_Release_Request" control information, it MUST store the arrival time. Then it MUST calculate the time difference, "Tdiff", between the arrival time and the start of the current refresh period, "T_period". Furthermore, this node MUST derive the value of the "T_Lag", from the <Time Lag> parameter. "T_Lag" can be found by multiplying the value included in the <Time Lag> parameter with the length of the refresh period, "T_period". If the derived time lag, "T_lag", is smaller than the calculated time difference, "T_diff", then this node MUST decrease the PHB reservation state with the number of resource units indicated in the <Bandwidth> parameter of the "RMD QoS Description" field that has been sent together with the "PHR_Release_Request" control information container, but not below zero.

An RMD specific release procedure can be triggered by an end-to-end RESERVE with a TEAR flag set ON (see Section 4.6.1.5.1) or it can be triggered by either an intra-domain RESPONSE, an end-to-end RESPONSE or an end-to-end NOTIFY message that includes a marked (i.e., PDR <M> and/or PDR <S> parameters are set ON) "PDR_Reservation_Report" or "PDR_Congestion_Report" and/or an INFO_SPEC object that includes one of the following error codes, see Section 4.7:

0x01 - Informational
0x03 - Protocol error
0x04 - Transient Failure
0x05 - Permanent failure
0x06 - QoS-related Error

4.6.1.5.1. Triggered by a RESERVE message

This RMD explicit release procedure can be triggered by a tear (TEAR flag set ON) end-to-end RESERVE message. When a tear (TEAR flag set ON) end-to-end RESERVE message arrives to the QNE Ingress then the QNE Ingress node SHOULD process the message in a standard QoS-NSLP way (see [QoS-NSLP]). In addition to this, the RMD RMF has to be notified.

Similar to Section 4.6.1.1, a bypassing procedure has to be initiated by the QNE Ingress node. The bypassing procedure is performed according to the description given in Section 4.4. At the QNE Ingress the end-to-end RESERVE message is marked, i.e., modifying the QoS-NSLP default NSLP-ID value to another NSLP-ID predefined value, which corresponds to a RAO value that will be used by the GIST message that carries the end-to-end RESPONSE message to bypass the QNE Interior nodes. It will generate an intra-domain RESERVE(RMD-QSPEC) message. Before generating this message, the RMD RMF is using the RMD traffic class (PHR) resources (specified in <Bandwidth>) and the PHB type (specified in <PHB Class>) for a RMD release procedure. This can be achieved by subtracting the amount of the requested resources from the total reserved amount of resources stored in the RMD traffic class state.

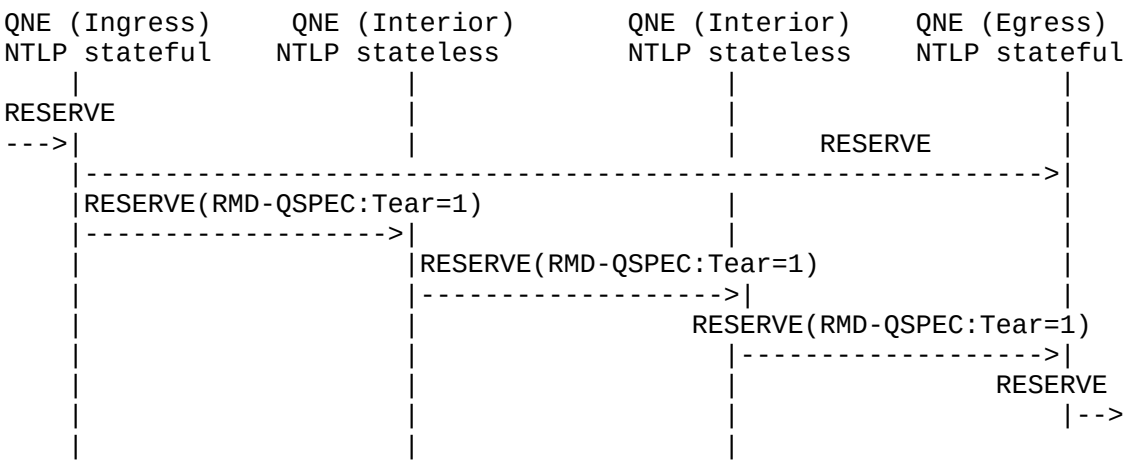

Figure 11: Explicit release triggered by RESERVE used by the RMD-QOSM

The intra-domain RESERVE (RMD-QSPEC) message MUST include a "RMD QoS Description" field and a PHR container, (i.e., "PHR_Resource_Release") and it MAY include a PDR container, (i.e., PDR_Release_Request). An example of this operation can be seen in Figure 11.

Most of the non default values of the objects contained in the tear intra-domain RESERVE message are set by the QNE Ingress node in the same way as described in Section 4.6.1.1. The following objects are set differently:

* The flag REPLACE MUST be set to FALSE = 0;

* The <RII> object MUST not included in this message. This is because the QNE Ingress node does not need to receive a response from the QNE Egress node;

* the TEAR flag MUST be set to ON;

* the PHR resource units MUST be included into the <Bandwidth> parameter of the "RMD QoS Description" field;

* the value of the <Admitted Hops> parameter MUST be set to "1";

* the value of the <Time Lag> parameter of the PHR container is calculated by the RMD-QOSM functionality (see 4.6.1.5)the value of the <Control Type> parameter of PHR container is set to "3" (i.e., PHR_Resource_Release).

The intra-domain tear RESERVE (RMD-QSPEC) message is received and processed by the QNE Interior nodes. Most of the non-default values of the objects contained in this refresh intra-domain RESERVE (RMD-QSPEC) message are set by a QNE Interior node in the same way as described in Section 4.6.1.1. The following objects are set and processed differently:

* Any QNE Interior node that receives the combination of the "RMD QoS Description" field and the "PHR_Resource_Release" control information container MUST identify the traffic class (PHB) and release the requested resources included in the <Bandwidth> parameter. This can be achieved by subtracting the amount of RMD traffic class requested resources, included in the <Bandwidth> parameter, from the total reserved amount of resources stored in the RMD traffic class state. The value of the <Time Lag> parameter of the "PHR_Resource_Release" container is used during the release procedure as explained in Section 4.6.1.5.

The intra-domain tear RESERVE (RMD-QSPEC) message is received and processed by the QNE Egress node. The "RMD QoS Description" and the "PHR RMD-QOSM control " container (and if available the "PDR RMD-QOSM control information" container) are read and processed by the RMD QoS node.

The value of the <Bandwidth>
parameter of the "RMD QoS Description" field and the value of the <Time Lag> field of the PHR container MUST be used by the RMD release procedure. This can be achieved by subtracting the amount of RMD traffic class requested resources, included in the <Bandwidth> parameter, from the total reserved amount of resources stored in the RMD traffic class state.

The end-to-end RESERVE message is forwarded by the next hop (i.e., the QNE Egress) only if the intra-domain tear RESERVE (RMD-QSPEC) message arrives at the QNE Egress node. Furthermore, the QNE Egress MUST stop the marking process that was used to bypass the QNE Interior nodes by reassigning the QoS-NSLP default NSLP-ID value to the end-to-end RESERVE message, see Section 4.4.
Note that the above described procedure applies to the situation that the QNE edges maintain a per flow QoS-NSLP reservation state. When the QNE edges maintain aggregated QoS-NSLP reservation states the RMD-QOSM functionality may start a RMD modification procedures (see Section 4.6.1.4) that uses the explicit release procedure described in this Section.

4.6.1.5.2 Triggered by a marked RESPONSE or NOTIFY message

This RMD explicit release procedure can be triggered by either an end-to-end RESPONSE message with a <M> marked PDR container (see Section 4.6.1.2) an intra-domain RESPONSE message with a <S> marked PDR container (see Section 4.6.1.6.1) or an end to end NOTIFY message (see Section 4.6.1.6) with an INFO_SPEC object with the following values:

Error Severity Class: 0x01 Informational
Error Code value: 0x05 Congestion situation

The RMD specific release procedure that is triggered by an end-to-end RESPONSE message with a <M> marked PDR container (see Section 4.6.1.2) can be terminated at any QNE edge
or any QNE Interior node using the <Max_Admitted Hops> field.

The RMD specific explicit release procedure that is terminated at a QNE Interior (or QNE edge) node is denoted as RMD partial release procedure. This explicit release procedure can be used, for example, during a RMD specific operation for unsuccessful reservation (see Section 4.6.1.2). When the RMD QoS signaling model functionality of a QNE Ingress node receives a <M> or <S> marked PDR container of type "PDR_Reservation_Report" or "PDR_Congestion_Report", it MUST start an RMD partial release procedure. The QNE Ingress node generates an intra-domain RESERVE (RMD-QSPEC) message. Before generating this message, the RMD-QOSM functionality is using the RMD traffic class (PHR) resource units for a RMD release procedure. This can be achieved by subtracting the amount of RMD traffic class requested resources from the total reserved amount of resources stored in the RMD traffic class state.

When the generation of the intra-domain RESERVE (RMD-QSPEC) message is triggered by an end-to-end NOTIFY message, which does not carry a PDR container, but it carries an INFO_SPEC object with the following values, then the intra-domain RESERVE(RMD-QSPEC) message MUST include an <RMD QoS Description> field and a PHR container, (i.e., PHR_Resource_Release) and it MAY include a PDR container, (i.e., PDR_Release_Request). Note that this procedure is accomplished during the severe congestion handling by proportional data packet marking, see Section 4.6.1.6.2. The error code values carried by this NOTIFY message are:

Error Severity Class: 0x01 Informational
Error Code value: 0x05 Congestion situation

Furthermore, note that the tear intra-domain RESERVE message is generated as it is shown in Figure 12, when it is triggered by either a NOTIFY message or RESPONSE message that do not carry a PDR container, but the INFO_SPEC object includes one of the following error codes, see Section 4.7:

- 0x01 - Informational
- 0x03 - Protocol error
- 0x04 - Transient Failure
- 0x05 - Permanent failure
- 0x06 - QoS-related Error

An example of this message exchange can be seen in Figure 12. Most of the non-default values of the objects contained in the tear intra-domain RESERVE(RMD-QSPEC) message are set by the QNE Ingress node in the same way as described in Section 4.6.1.1.

The following objects MUST be used and/or set differently:

* The flag REPLACE MUST be set to FALSE;

* The value of the <M> parameter of the PHR container MUST be set to "1".

* the value of the <S> parameter of the PHR container MUST be set to "1".

* The RESERVE message MAY include a PDR container.

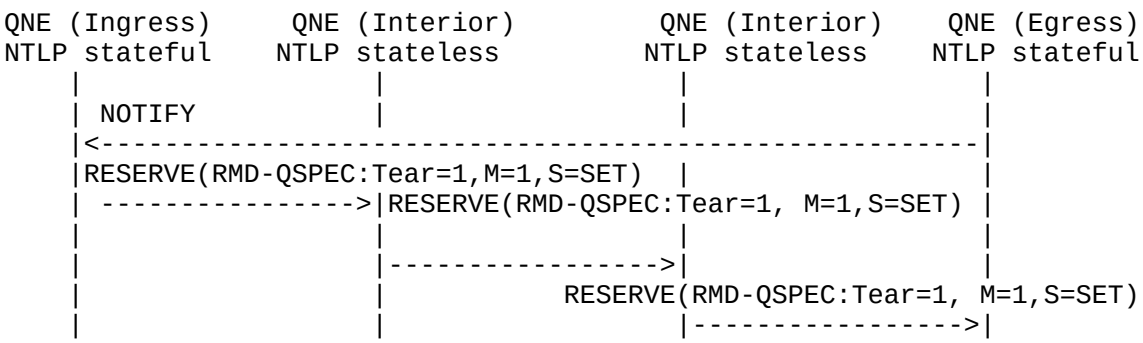

Figure 12: Basic operation during RMD explicit release procedure triggered by NOTIFY used by the RMD-QOSM

When the generation of the intra-domain RESERVE(RMD-QSPEC) message is triggered by an end-to-end RESPONSE(PDR) message then this generated intra-domain RESERVE(RMD-QSPEC) message MUST include a <RMD QoS Description> field and a PDR container, (i.e., PHR_Resource_Release) and it MAY include a PDR container, (i.e., PDR_Release_Request). An example of this operation can be seen in Figure 13.

Most of the non-default values of the objects contained in the tear intra-domain RESERVE(RMD-QSPEC) message are set by the QNE Ingress node in the same way as described in Section 4.6.1.1.

The following objects MUST be used and/or set differently:

* The flag REPLACE MUST be set to FALSE;

* The value of the <M> parameter of the PHR container MUST be set to "1".

- The RESERVE message MAY include a PDR container.

* When the tear intra-domain RESERVE message is triggered by an intra-domain RESPONSE(RMD-QSPEC) message, then the value of the <Max Admitted Hops> parameter of the PDR container included in the received <M> or <S> marked intra-domain RESPONSE(PDR) message MUST be included in the <Max Admitted Hops> parameter of the PDR container of the RESERVE message. Note that this procedure is applied for the severe congestion handling by the RMD-QOSM refresh procedure (see Section 4.6.1.6.1). The tear intra-domain RESERVE message propagates in this case until the QNE egress (similar to Figure 12).

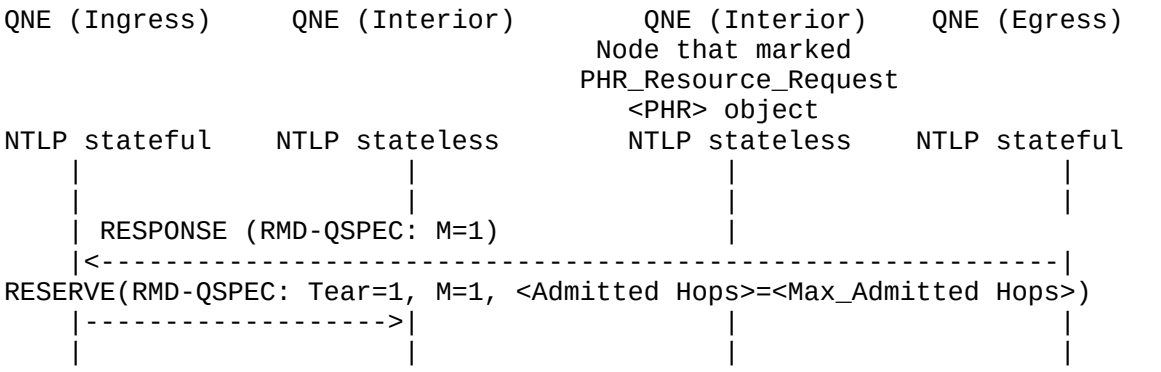

Figure 13: Basic operation during RMD explicit release procedure Triggered by RESPONSE used by the RMD-QOSM

Any QNE edge or QNE Interior node that receives the "RMD QoS Description" field and the PHR container MUST identify the traffic class state (PHB), using the <PHB Class> parameter, and release the requested resources included in the <Bandwidth> field. This can be achieved by subtracting the amount of RMD traffic class requested resources, included in the <Bandwidth> field, from the total reserved amount of resources stored in the RMD traffic class state. The value of the <Time Lag> parameter of the PHR field is used during the release procedure as explained in Section 4.6.1.5.

The <Admitted Hops> value included in the PHR container is increased by one. If the value of <M> parameter of the "PHR_Resource_Release" control information container is "1" and if the value of the <S> parameter is set to "0" then the <Max_Admitted Hops> value included in the PDR container MUST be compared with the calculated <Admitted Hops> value. When these two values are equal then the intra-domain RESERVE(RMD-QSPEC) has to be terminated and it will not be forwarded downstream. The reason of this is that the QNE node that is currently processing this message was the last QNE node that successfully processed the "RMD QoS Description" field and PHR container of its associated initial reservation request (i.e., initial intra-domain RESERVE(RMD-QSPEC) message). Its next QNE downstream node was unable to successfully process the initial

reservation request, therefore, this QNE node marked the <M> parameter of the "PHR_Resource_Request" control information. When the values of the <M> and <S> parameters are set to "0", then this message will not be terminated by a QNE Interior node, but it will be forwarded in the downstream direction. The QNE Egress node will receive and process the PHR_Resource_Release control information. Afterwards, the QNE Egress node MUST terminate the intra-domain RESERVE(RMD-QSPEC) message.

Note that the above described procedure applies to the situation that the QNE edges maintain a per flow QoS-NSLP reservation state. When the QNE edges maintain aggregated QoS-NSLP reservation states the RMD-QOSM functionality MAY start a RMD modification procedures (see Section 4.6.1.4) that uses the explicit release procedure described in this section.

4.6.1.6. Severe congestion handling

This section describes the operation of the RMD-QOSM when a severe congestion occurs within the Diffserv domain.

When a failure in a communication path, e.g., a router or a link failure occurs, the routing algorithms will adapt to failures by changing the routing decisions to reflect changes in the topology and traffic volume. As a result, the re-routed traffic will follow a new path, which may result in overloaded nodes as they need to support more traffic. This may cause severe congestion in the communication path. In this situation the available resources, are not enough to meet the required QoS for all the flows along the new path. Therefore, one or more flows SHOULD be terminated, or forwarded in a lower priority queue.

Interior nodes notify edge nodes by data marking or marking the refresh messages.

4.6.1.6.1 Severe congestion handling by the RMD-QOSM refresh procedure

The QoS-NSLP and RMD are able to cope with congested situations using the refresh procedure, see Section 4.6.1.3. If the refresh is not successful in an QNE Interior node, edge nodes are notified by "S" marking the refresh messages and by including the percentage of overload into the <Overload %> field in the "PHR_Refresh_Update" container, carried by the intra-domain RESERVE message.
The intra-domain RESPONSE message that is sent by the QNE Egress towards QNE Ingress will contain a PDR container with a Parameter/Container ID = 10, i.e., "PDR_Congestion_Report". The values of the <S> and <Overload %> fields of this container should be set equal to the values of the <S> and <Overload %> fields, respectively, carried by the "PHR_Refresh_Update" message. Part of

the flows, corresponding to the <Overload %>, are terminated, or forwarded in a lower priority queue. The flows can be terminated by the RMD release procedure described in Section 4.6.1.5. Note that the above described functionality applies to the RMD reservation-based and to the NSIS measurement-based admission control schemes. Furthermore, note that the above functionalities apply also for the scenario where the QNE Edge nodes maintain either per flow QoS-NSLP reservation states or aggregated QoS-NSLP reservation states.

In general, relying on the soft state refresh mechanism solves the congestion within the time frame of the refresh period. If this mechanism is not fast enough additional functions should be used, which are described in Section 4.6.1.6.2.

4.6.1.6.2 Severe congestion handling by proportional data packet marking

This severe congestion handling method requires the following functionalities.

4.6.1.6.2.1 Operation in the Interior nodes

The Interior node detecting severe congestion remarks data packets passing the node. For this remarking, two additional DSCPs can be allocated for each traffic class. One DSCP MAY be used to indicate that the packet passed a congested node. This type of DSCP is denoted in this document as "affected DSCP" and is used to indicate that a packet passed through a severe congested node. The use of this DSCP type eliminates the possibility that, due to e.g. ECMP (Equal Cost Multiple Paths) enabled routing, the egress node either does not detect packets passed a severe congested node or erroneously detects packets that actually did not pass the severe congested node. Note that this type of DSCP MUST only be used if all the nodes within the RMD domain are configured to use it. Otherwise, this type of DSCP MUST not be applied. The other DSCP MUST be used to indicate the degree of congestion by marking the bytes proportionally to the degree of congestion. This type of DSCP is denoted in this document as "encoded DSCP".

Note that in this document the terms marked packets or marked bytes refer to the "encoded DSCP". The terms unmarked packets or unmarked bytes are representing the packets or the bytes belonging to these packets that their DSCP is either the "affected DSCP" or the original DSCP. Furthermore, in the algorithm described below it is considered that the router may drop received packets. The counting/measuring of marked or unmarked bytes described in this section is accomplished within measurement periods. All nodes within a RMD domain use the same, fixed measurement interval, say T seconds, which MUST be pre-configured.

It is RECOMMENDED that the total number of additional DSCPs needed for severe congestion handling within an RMD domain should be as low as possible and it should not exceed the limit of 16. One possibility to reduce the number of used DSCPs is to use only the "encoded DSCP" and not to use "affected DSCP" marking. Another possible solution is for example, to allocate one DSCP for severe congestion indication for each of the AF classes, independently from their dropping precedence. Assuming 4 AF classes and 1 EF class, and using one DSCP per traffic class then the number of DSCPs used in this situation for severe congestion is 5. If two additional DSCP's are used then the total number in this case is 10.

An example of a remarking procedure can be found in Appendix A.1.1.

4.6.1.6.2.2 Operation in the Egress nodes

The QNE Egress node applies a predefined policy to solve the severe congestion situation, by selecting a number of inter-domain (end-to-end) flows that SHOULD be terminated, or forwarded in a lower priority queue.

When the RMD domain does not use the "affected DSCP" marking then the egress MUST generate an ingress/egress pair aggregated state, for each ingress and for each supported PHB. This is because the edges must be able to detect in which ingress/egress pair a severe congestion occurs. When the RMD domain supports the "affected DSCP" marking then the egress is able to detect all flows that are affected by the severe congestion situation. Therefore, when the RMD domain supports the "affected DSCP" marking, then the Egress MAY not generate and maintain the ingress/egress pair aggregated states.

The ingress/egress pair aggregated state can be derived by detecting, which flows are using the same PHB and are sent by the same Ingress (via the per flow end-to-end QoS-NSLP states).

Some flows, belonging to the same PHB traffic class might get other priority than other flows belonging to the same PHB traffic class. This difference in priority can be notified to the egress and ingress nodes either by the RESERVE message that carries the QSPEC associated with the end-to-end QoS model, i.e., <Preemption Priority> & <Defending Priority> parameter, or by using a local defined policy.

The terminated flows are selected from the flows having the same PHB traffic class as the PHB of the marked (as "encoded DSCP") and "affected DSCP" (when applied in the complete RMD domain) packets and (when the ingress/egress pair aggregated states are available).that are belonging to the same ingress/egress pair aggregate.

For flows associated with the same PHB traffic class the priority of the flow plays a significant role. An example of calculating the number of flows associated with each priority class that have to be terminated is explained in Appendix A.1.2.

For the flows (sessions) that have to be terminated, the QNE Egress node generates and sends a NOTIFY message to the QNE Ingress node (its upstream stateful QoS-NSLP peer) to indicate the severe congestion in the communication path.

The non-default values of the objects contained in the NOTIFY message MUST be set by the QNE Egress node as follows:

* the values of the <INFO_SPEC> object is set by the standard QoS-NSLP protocol functions.

* the INFO_SPEC object SHOULD include information that notifies that the end-to-end flow SHOULD be terminated. This information is as follows:

  Error Severity Class: 0x01 Informational
  Error Code value: 0x05 Congestion situation

The selection and notification process of the end-to-end is identical for the scenarios where the QNE Edges maintain per-flow or aggregated QoS-NSLP reservation states.

Furthermore, note that QNE egress SHOULD restore the original DSCP values of the remarked packets, otherwise multiple actions for the same event might occur. However, this value MAY not be restored if there is an SLA agreement between domains that a downstream domain handles the remarking problem.

4.6.1.6.2.3 Operation in the Ingress nodes

Upon receiving the (end-to-end) NOTIFY message, the QNE Ingress node resolves the severe congestion by a predefined policy, e.g., by refusing new incoming flows (sessions), terminating the affected and notified flows (sessions), or shifting them to an alternative RMD traffic class (PHB). This operation is depicted in Figure 14, where the QNE Ingress, for each flow (session) to be terminated, receives a NOTIFY message. The NOTIFY message SHOULD include an INFO-SPEC object with the following information:

Error Severity Class: 0x1 Informational
Error Code value: 0x05 Congestion situation

When the QNE Ingress node receives the end-to-end NOTIFY message, it associates this NOTIFY message with its bound intra-domain session, via the BOUND_SESSION_ID information included in the end-to-end per-flow QoS-NSLP state. The QNE Ingress uses the operation described in Section 4.6.1.5.2 to terminate the intra-domain session.

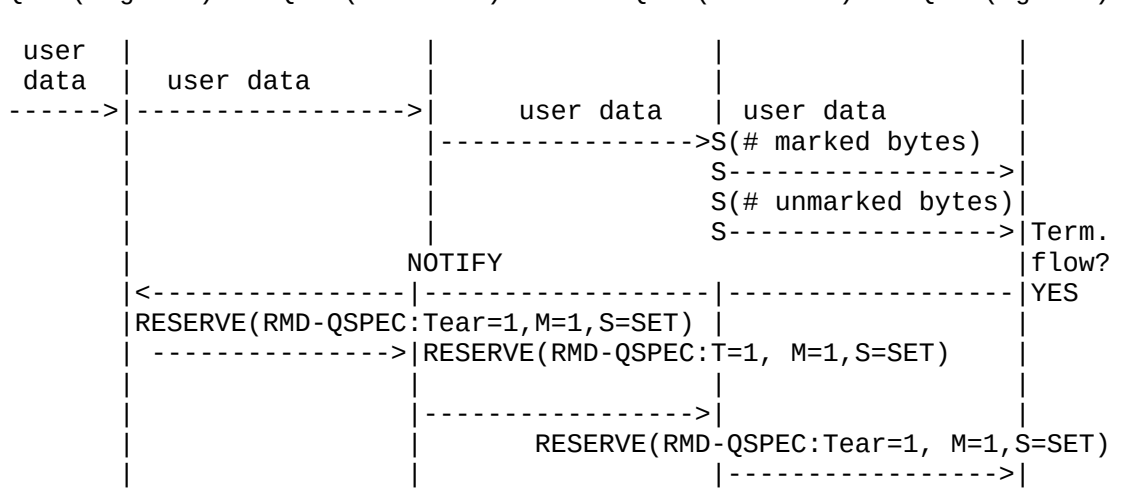

Figure: 14 RMD severe congestion handling

Note that the above functionality applies to the RMD reservation-based and to both measurement-based admission control methods (i.e., congestion notification based on probing and the NSIS measurement-based admission control). The above functionality applies also for the scenario where the QNE Edge nodes maintain either per flow QoS-NSLP reservation states or aggregated QoS-NSLP reservation states.

In the case that the edges support aggregated QoS-NSLP reservation states the following actions take place. When the QNE Ingress node receives the end-to-end NOTIFY message, it associates the NOTIFY message with the intra-domain aggregated QoS-NSLP state via the BOUND_SESSION_ID information included in the end-to-end per-flow QoS-NSLP state. The QNE Ingress node should reduce the bandwidth associated with the end-to-end flow from the aggregated bandwidth associated with its bound aggregated QoS-NSLP reservation state. This is accomplished by triggering the RMD modification for aggregated reservations procedure described in Section 4.6.1.4.

4.6.1.7 Admission control using congestion notification based on probing

The congestion notification function based on probing can be used to implement a simple measurement-based admission control within a Diffserv domain. At interior nodes along the data path congestion notification thresholds are set in the measurement based admission control function for the traffic belonging to different PHBs. These interior nodes are not NSIS aware nodes.

4.6.1.7.1 Operation in Ingress nodes

When an end-to-end reservation request (RESERVE) arrives at the Ingress node (QNE), see Figure 15, it is processed based on the procedures defined by the end-to-end QoS model.

If the ingress is configured to neither process this type of admission control nor any other admission control scheme specified in the previous sections, then the <NON QOSM Hop> parameter that is carried by the end-to-end QSpec SHOULD be set.

The DSCP field of the GIST datagram message that is used to transport this probe RESERVE message, SHOULD be marked with the same value of DSCP as the data path packets associated with the same session.

When (end-to-end) RESPONSE message is received by the Ingress node,it will be processed based on the procedures defined by the end-to-end QoS model.

4.6.1.7.2 Operation in Interior nodes

These Interior nodes are not needed to be NSIS aware nodes and they do not need to process NSIS functionality of NSIS messages. Using standard functionalties congestion notification thresholds are set for the traffic belonging to different PHBs, see Section 4.3.2.

The end-to-end RESERVE message, see Figure 15, is used as a probe packet.

The DSCP field of the GIST message carrying the RESERVE message will be re-marked when the corresponding "congestion notification" threshold is exceeded, see Section 4.3.2. Note that when the data rate is higher than the congestion notification threshold then also the data packets are remarked. An example of the detailed operation of this procedure is given in Appendix A.2.1.

4.6.1.7.3 Operation in Egress nodes

As emphasised in Section 4.6.1.6.2.2, the egress node, by using the per flow end-to-end QoS-NSLP states, can derive which flows are using the same PHB and are sent by the same ingress.

For each ingress, the egress SHOULD generate an ingress/egress pair aggregated state for each supported PHB.

In Appendix A.2.2 an example is described how and when a (probe) RESERVE message that arrives at the egress, is admitted or rejected.

If the request is rejected then the Egress node SHOULD generate an (end-to-end) RESPONSE message to notify that the reservation is unsuccesfull. In particular it will generate an INFO_SPEC object of:

Error Severity Class: 0x04, Transient failure
Error Code value: 0x07 Total reservation failure

The QSpec that was carried by the end to end RESERVE belonging to the same session as this end to end RESPONSE is included in this message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values. The "E" flag associated with the <QoS Reserved> object and the "E" flag associated with <Bandwidth> parameter are also set. This RESPONSE message will be sent to the Ingress node and it will be processed based on the end-to-end QoS model.

Note that QNE egress SHOULD restore the original DSCP values of the remarked packets, otherwise multiple actions for the same event might occur. However, this value MAY not be restored if there is an SLA agreement between domains that a downstream domain handles the remarking problem.

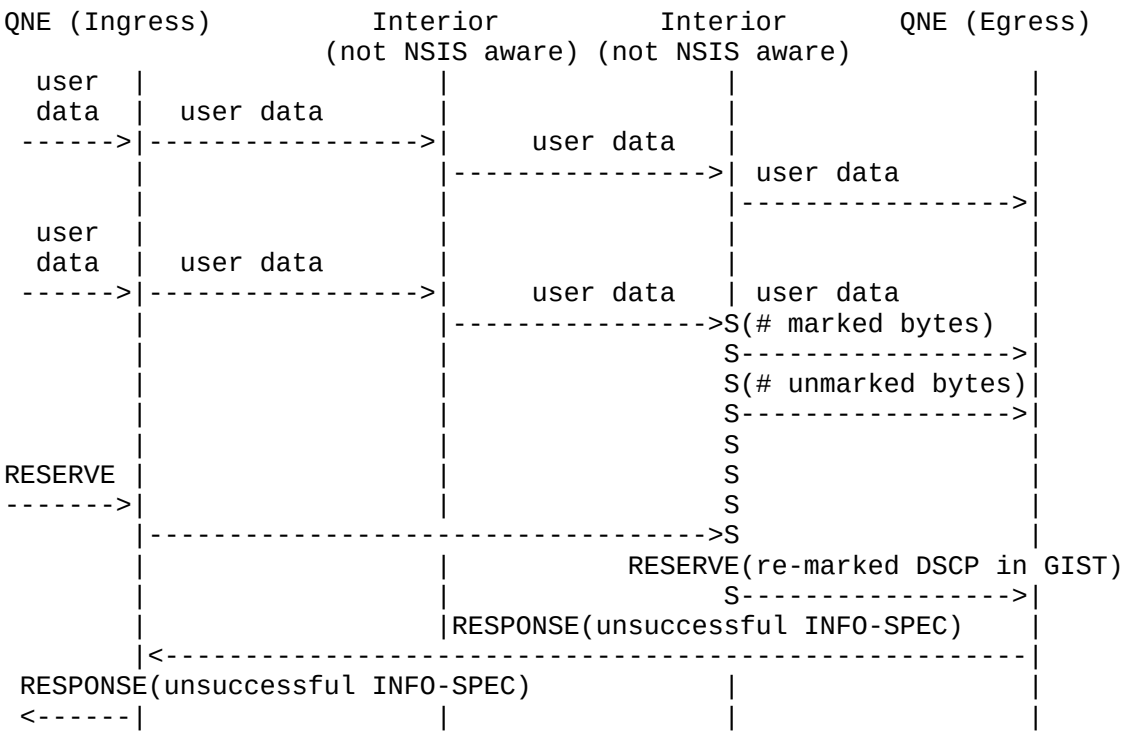

Figure: 15 Using RMD congestion notification function for admission control based on probing

4.6.2 Bi-directional operation

RMD assumes asymmetric routing by default. Combined sender-receiver initiated reservation cannot be efficiently done in the RMD domain because upstream NTLP states are not stored in Interior routers. Therefore, the bi-directional operation SHOULD be performed by two sender-initiated reservations (sender&sender). We assume that the QNE edge nodes are common for both upstream and downstream directions, therefore, the two reservations/sessions can be bound at the QNE edge nodes.

This bi-directional sender&sender procedure can then be applied between the QNE edges (QNE Ingress and QNE Egress) nodes of the RMD QoS signaling model. In the situation a security association exists between the QNE Ingress and QNE Egress nodes (see Figure 15), and the QNE Ingress node has the required <Bandwidth> parameters for both directions, i.e., QNE Ingress towards QNE Egress and QNE Egress towards QNE Ingress, then the QNE Ingress MAY include both <Bandwidth> parameters (needed for both directions) into the RMD-QSPEC within a RESERVE message. In this way the QNE Egress node is able to use the QoS parameters needed for the "Egress towards Ingress" direction (QoS-2). The QNE Egress is then able to create a RESERVE with the right QoS parameters included in the QSPEC, i.e., RESERVE (QoS-2). Both directions of the flows are bound by inserting the <BOUND_SESSION_ID> object at the QNE Ingress and QNE Egress.

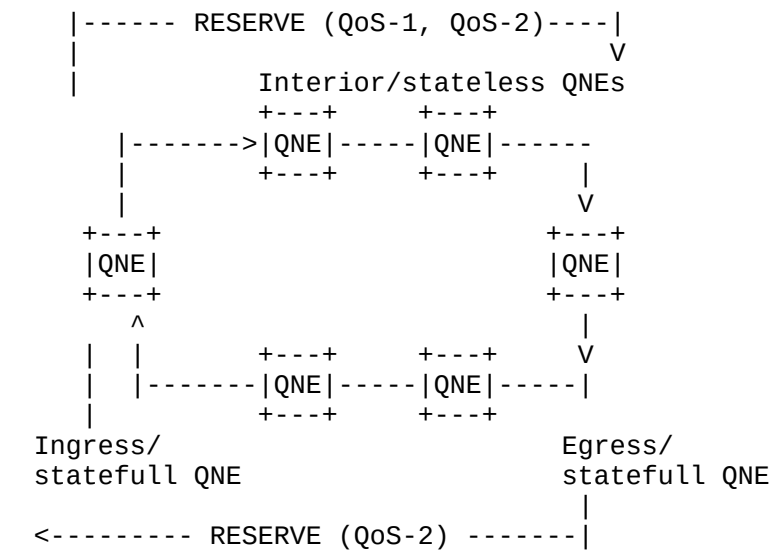

Figure 16: The bi-directional reservation scenario in the RMD domain

A bidirectional reservation, within the RMD domain, is indicated by the PHR <B> and PDR <B> flags, which are set in all messages.

In this case two BOUND_SESSION_ID objects SHOULD be used.
The first BOUND_SESSION_ID object is applied in the following way. The end-to-end RESERVE SHOULD contain in the BOUND_SESSION_ID the SESSION_ID of its bound intra-domain session. Furthermore, if the QNE Edge nodes maintain intra-domain per flow QoS-NSLP reservation states then the value of Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions). If the QNE Edge nodes maintain intra-domain aggregated QoS-NSLP reservation states then the value of Binding_Code MUST be set to 0x03 (Aggregate sessions).

The intra-domain RESERVE SHOULD contain in the BOUND_SESSION_ID the SESSION_ID of its bound end to end session. The value of the Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions).

The SESSION_ID field of the second BOUND_SESSION_ID object depends on the direction of the message. An upstream RMD QoS-NSLP message SHOULD contain the SESSION_ID of the bound downstream end-to-end flow. A downstream RMD QoS-NSLP message SHOULD contain the SESSION_ID of the bound upstream end-to-end flow. In both cases the value of the Binding_Code associated with this BOUND_SESSION_ID object SHOULD be equal to 0x02.

If no security association exists between the QNE Ingress and QNE Egress nodes the bi-directional reservation for the sender&sender scenario in the RMD domain SHOULD use the scenario specified in [QoS-NSLP] as "Bi-directional reservation for sender&sender scenario".

In the following sections it is considered that the QNE edge nodes are common for both upstream and downstream directions and therefore, the two reservations/sessions can be bound at the QNE edge nodes. Furthermore, it is considered that a security association exists between the QNE Ingress and QNE Egress nodes, and the QNE Ingress node has the required <Bandwidth> parameters for both directions, i.e., QNE Ingress towards QNE Egress and QNE Egress towards QNE Ingress.

4.6.2.1 Successful and unsuccessful reservations

This section describes the operation of the RMD-QOSM where a RMD bi-directional reservation operation is either successfully or unsuccessfully accomplished.

The bi-directional successful reservation is similar to a combination of two unidirectional successful reservations that are accomplished in opposite directions, see Figure 17. The main differences of the bi-directional successful reservation procedure with the combination of two unidirectional successful reservations accomplished in opposite directions are as follows. The intra-domain RESERVE message sent by the QNE Ingress node towards the QNE Egress node, is denoted in Figure 17 as RESERVE (RMD-QSPEC): "forward". The main differences between the RESERVE (RMD-QSPEC): "forward" message used for the bi-directional successful reservation procedure and a RESERVE (RMD-QSPEC) message used for the unidirectional successful reservation are as follows:

* Two BOUND_SESSION_ID objects MUST be used. The first BOUND_SESSION_ID object contains the SESSION_ID of its bound End-to-end session. The value of the Binding_Code MUST be set to 0x01 (Tunnel and end-to-end sessions). The SESSION_ID field of the second BOUND_SESSION_ID object SHOULD contain the SESSION_ID of the bound "reverse" end-to-end flow. The value of the Binding_Code associated with this BOUND_SESSION_ID object SHOULD be equal to 0x02.

* the RII object MUST NOT included in the message. This is because no RESPONSE message is expected to arrive.

* the <B> bit of the PHR container indicates a bi-directional reservation and it MUST be set to "1".

* the PDR container is also included into the RESERVE(RMD-QSPEC): "forward" message. The value of the Parameter/Container ID is "4", i.e., "PDR_Reservation_Request". Note that the response PDR container sent by a QNE Egress to a QNE Ingress node is not carried by an end-to-end RESPONSE message, but it is carried by an intra-domain RESERVE message that is sent by the QNE Egress node towards the QNE Ingress node (denoted in Figure 16 as RESERVE(RMD-QSPEC):"reverse").

* the <B> PDR bit indicates a bi-directional reservation and is set to "1".

* the <PDR Reverse Requested Resources> field specifies the requested bandwidth that has to be used by the QNE Egress node to initiate another intra-domain RESERVE message in the reverse direction.

The RESERVE(RMD-QSPEC):"reverse" message is initiated by the QNE Egress node at the moment that the RESERVE(RMD-QSPEC):"forward" message is successfully processed by the QNE Egress node.
The main differences between the RESERVE(RMD-QSPEC):"reverse" message used for the bi-directional successful reservation procedure and a RESERVE(RMD-QSPEC) message used for the unidirectional successful reservation are as follows:

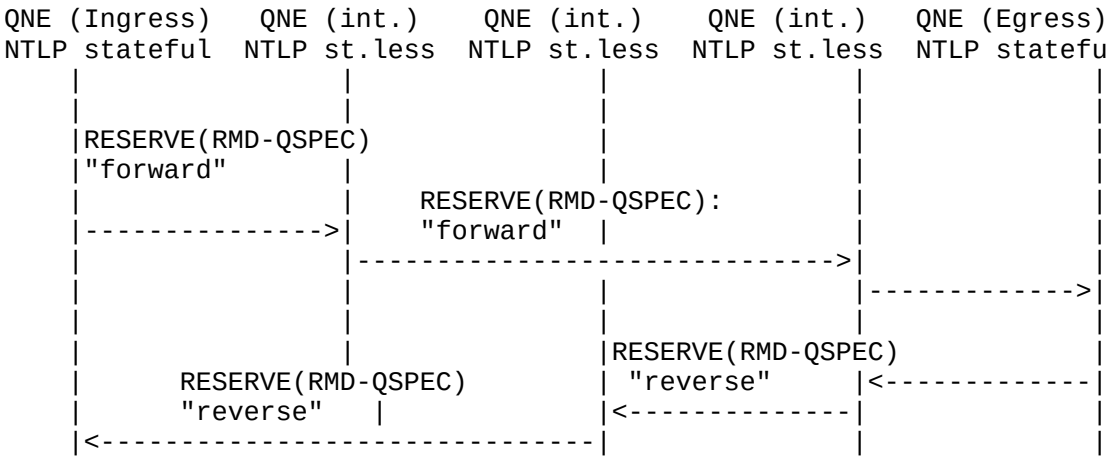

Figure 17: Intra-domain signaling operation for successful bi-directional reservation

* two BOUND_SESSION_ID objects SHOULD be used.
  The first BOUND_SESSION_ID object contains the SESSION ID of its bound end to end session. The value of the Binding_Code = 0x01 (Tunnel and end-to-end sessions). The SESSION_ID field of the second BOUND_SESSION_ID object SHOULD contain the SESSION_ID of the bound "forward" end-to-end flow. The value of the Binding_Code associated with this BOUND_SESSION_ID object SHOULD be equal to 0x02.

* the RII object is not included in the message. This is because no RESPONSE message is expected to arrive;

* the value of the <Bandwidth> parameter is set equal to the value of the <PDR Reverse Requested Resources> field included in the RESERVE(RMD-QSPEC):"forward" message that triggered the generation of this RESERVE(RMD-QSPEC): "reverse" message;

* the <B> bit of the PHR container indicates a bi-directional reservation and is set to "1";

* the PDR container is included into the RESERVE(RMD-QSPEC):"reverse" message. The value of the Parameter/Container ID is "7", i.e., "PDR_Reservation_Report";

* the <B> PDR bit indicates a bi-directional reservation and is set to "1".

Figure 18 and Figure 19 show the flow diagrams used in case of a unsuccessful bi-directional reservation. In Figure 18 it is considered that the QNE that is not able to support the requested <Bandwidth> is located in the direction QNE Ingress towards QNE Egress. In Figure 19 it is considered that the QNE that is not able to support the requested <Bandwidth> is located in the direction QNE Egress towards QNE Ingress.

The main differences between the bi-directional unsuccessful procedure shown in Figure 18 and the bi-directional successful procedure are as follows:

* the QNE node that is not able to reserve resources for a certain request is located in the "forward" path, i.e., path from QNE Ingress towards the QNE Egress.

* the QNE node that is not able to support the requested <Bandwidth> it MUST mark the <M> bit, i.e., set to value "1", of the RESERVE(RMD-QSPEC): "forward".

The operation for this type of unsuccessful bi-directional reservation is similar to the operation for unsuccessful uni-directional reservation shown in Figure 9. The main difference is that the QNE Egress generates an intra-domain (local) RESPONSE(PDR) message that is sent towards QNE Ingress node.

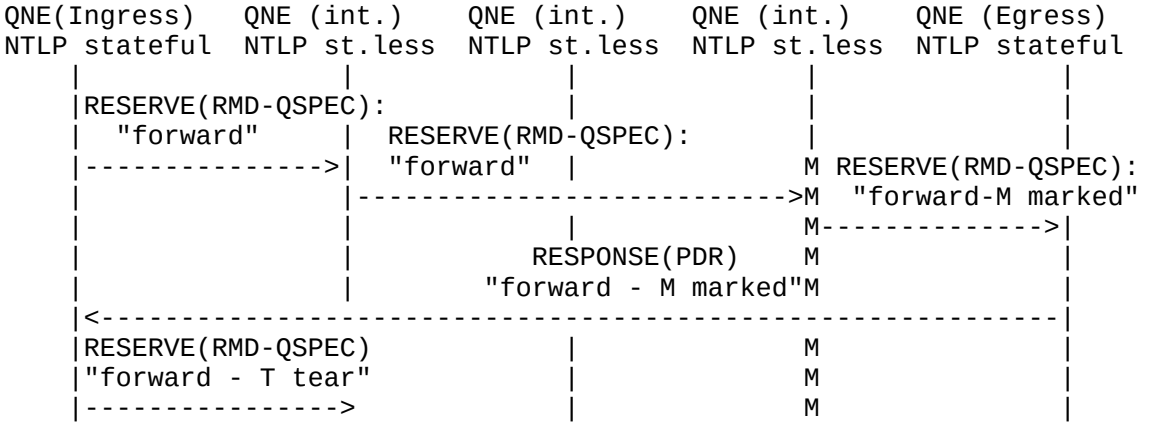

Figure 18: Intra-domain signaling operation for unsuccessful bi-directional reservation (rejection on path QNE(Ingress) towards QNE(Egress))

The main differences between the bi-directional unsuccessful procedure shown in Figure 19 and the in bi-directional successful procedure are as follows:

* the QNE node that is not able to reserve resources for a certain request is located in the "reverse" path, i.e., path from QNE Egress towards the QNE Ingress.

* the QNE node that is not able to support the requested <Bandwidth> it MUST mark the <M> bit, i.e., set to value "1", the RESERVE(RMD-QSPEC):"reverse".

* the QNE Ingress uses the information contained in the received PHR and PDR containers of the RESERVE(RMD-QSPEC): "reverse" and generates a tear intra-domain (local) RESERVE(RMD-QSPEC): "forward - T tear" message. This message carriers a "PHR_Release_Request" and a "PDR_Release_Request" control information. This message is sent to QNE Egress node. The QNE Egress node by using the information contained in the "PHR_Release_Request" and the "PDR_Release_Request" control info containers it generates a RESERVE(RMD-QSPEC):"reverse - T tear" message that is sent towards the QNE Ingress node.

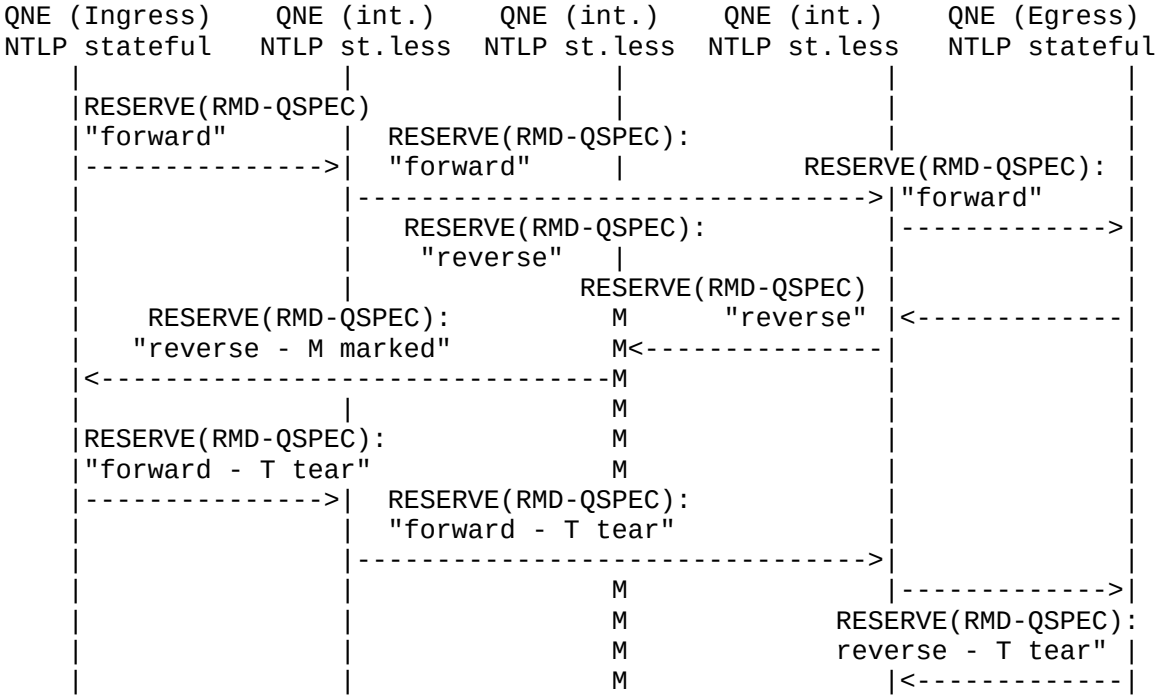

Figure 19: Intra-domain signaling normal operation for unsuccessful bi-directional reservation (rejection on path QNE(Egress) towards QNE(Ingress)

4.6.2.2 Refresh reservations

This section describes the operation of the RMD-QOSM where a RMD bi-directional refresh reservation operation is accomplished.

The refresh procedure in case of RMD reservation-based method follows a similar scheme as the successful reservation procedure, described in Section 4.6.2.1, and depicted in Figure 17 and the way of how the refresh process of the reserved resources is maintained, is similar to the refresh process used for the intra-domain uni-directional reservations (see Section 4.6.1.3).

Note that the RMD traffic class refresh periods used by the bound bi-directional sessions MUST be equal in all QNE edge and QNE Interior nodes.

The main differences between the RESERVE(RMD-QSPEC):"forward" message used for the bi-directional refresh procedure and a RESERVE(RMD-QSPEC):"forward" message used for the bi-directional successful reservation procedure are as follows:

* the value of the Parameter/Container ID of the PHR container is "2", i.e., "PHR_Refresh_Update".

* the value of the Parameter/Container ID of the PDR container is "5", i.e., "PDR_Refresh_Request".

The main differences between the RESERVE(RMD-QSPEC):"reverse" message used for the bi-directional refresh procedure and the RESERVE (RMD-QSPEC): "reverse" message used for the bi-directional successful reservation procedure are as follows:

* the value of the Parameter/Container ID of the PHR container is "2", i.e., "PHR_Refresh_Update".

* the value of the Parameter/Container ID of the PDR container is "8", i.e., "PDR_Refresh_Report".

4.6.2.3 Modification of aggregated reservations

This section describes the operation of the RMD-QOSM where a RMD

In the case when the QNE edges maintain, for the RMD QoS model, QoS-NSLP aggregated reservation states and if such an aggregated reservation has to be modified (see Section 4.3.1) then similar procedures to Section 4.6.1.4 are applied. In particular:

* When the modification request requires an increase of the reserved resources, the QNE Ingress node MUST include the corresponding value into the <Bandwidth> parameter of the "RMD QoS Description" field, which is sent together with a "PHR_Resource_Request" control information. If a QNE edge or QNE Interior node is not able to reserve the number of requested resources, then the "PHR_Resource_Request" control information associated with the <Bandwidth> parameter MUST be marked. In this situation the RMD specific operation for unsuccessful reservation will be applied (see Section 4.6.2.1).

* When the modification request requires a decrease of the reserved resources, the QNE Ingress node MUST include this value into the <Bandwidth> parameter of the "RMD QoS Description" field. Subsequently an RMD release procedure SHOULD be accomplished (see Section 4.6.2.4).

4.6.2.4 Release procedure

This section describes the operation of the RMD-QOSM where a RMD bi-directional reservation release operation is accomplished. The message sequence diagram used in this procedure is similar to the one used by the successful reservation procedures, described in Section 4.6.2.1, and depicted in Figure 17. However, the way of how the release of the reservation is accomplished, is similar to the RMD release procedure used for the intra-domain uni-directional reservations (see Section 4.6.1.5 and Figure 18 and Figure 19).

The main differences between the RESERVE (RMD-QSPEC): "forward" message used for the bi-directional release procedure and a RESERVE (RMD-QSPEC): "forward" message used for the bi-directional successful reservation procedure are as follows:

* the value of the Parameter/Container ID of the PHR container is "3", i.e."PHR_Release_Request";

* the value of the Parameter/Container ID of the PDR container is "6", i.e., "PDR_Release_Request";

The main differences between the RESERVE (RMD-QSPEC): "reverse" message used for the bi-directional release procedure and the RESERVE (RMD-QSPEC): "reverse" message used for the bi-directional successful reservation procedure are as follows:

* the value of the Parameter/Container ID of the PHR container is "3", i.e., "PHR_Release_Request";

* the PDR container is not included in the RESERVE (RMD-QSPEC): "reverse" message.

4.6.2.5 Severe congestion handling

This section describes the severe congestion handling operation used in combination with bi-directional reservation procedures. This severe congestion handling operation is similar to the one described in Section 4.6.1.6.

4.6.2.5.1 Severe congestion handling by the RMD-QOSM bi-directional refresh procedure

This procedure is similar to the severe congestion handling procedure described in Section 4.6.1.6.1. The difference is related to how the refresh procedure is accomplished, see Section 4.6.2.2 and to how the flows are terminated, see Section 4.6.2.4.

4.6.2.5.2 Severe congestion handling by proportional data packet marking

This section describes the severe congestion handling by proportional data packet marking when this is combined with a bi-directional reservation procedure.

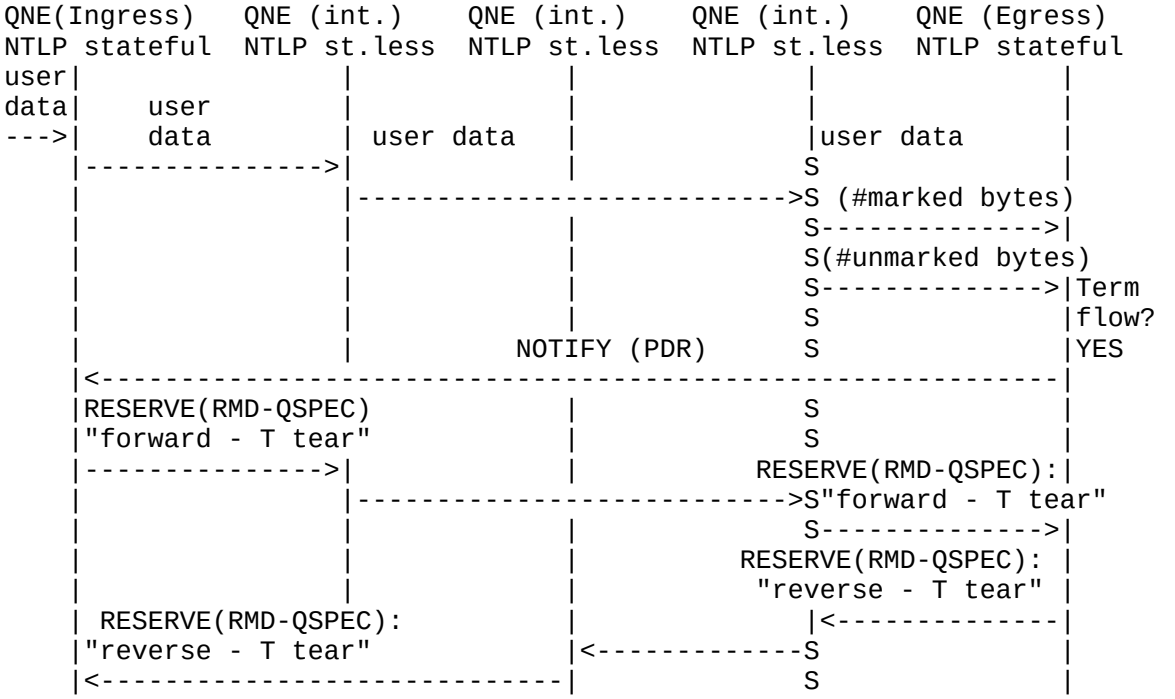

Figure 20: Intra-domain RMD severe congestion handling for bi-directional reservation (congestion on path QNE(Ingress) towards QNE(Egress))

This procedure is similar to the severe congestion handling procedure described in Section 4.6.1.6.2. The main difference is related to the location of the severe congested node, i.e., "forward" path (i.e., path between QNE Ingress towards QNE Egress) or "reverse" path (i.e., path between QNE Egress towards QNE Ingress). Another difference is associated with the way of how the egress node selects the flows that have to be terminated. Note that when a severe congestion situation occurs on e.g.a forward path, and flows are terminated to solve the severe congestion in forward path, then the reserved bandwidth associated with the terminated bidirectional flows will also be released. Therefore, a careful selection of the flows that have to be terminated should take place. An example of such a selection is given in Appendix A.3.1.

Furthermore, a special case of this operation is associated to the severe congestion situation occuring simultaneously on the forward and reverse paths. An example of this operation is given in Appendix A.3.2.

Figure 20 shows the scenario where the severe congested node is located in the "forward" path. This scenario is very similar to the severe congestion handling scenario described in Section 4.6.1.6.2 and shown in Figure 14. The difference is related to the release procedure, which is accomplished in the same way as described in Section 4.6.2.4.

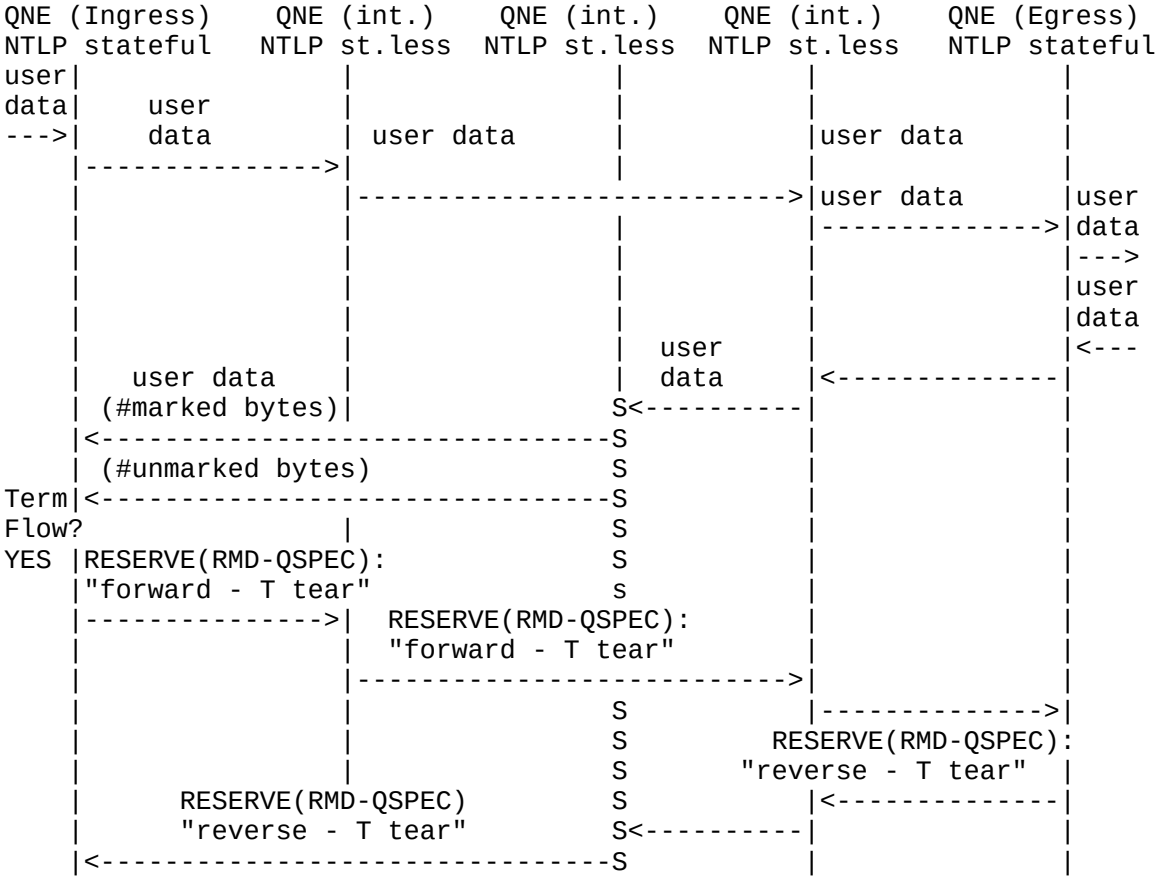

Figure 21: Intra-domain RMD severe congestion handling for bi-directional reservation (congestion on path QNE(Egress) towards QNE(Ingress))

Figure 21 shows the scenario where the severe congested node is located in the "reverse" path. The main difference between this scenario and the scenario shown in Figure 20 is that no intra-domain NOTIFY(PDR) message has to be generated by the QNE Egress node. This is because the (#marked and #unmarked) user data is arriving at the QNE Ingress. The QNE Ingress node will be able to calculate the number of flows that have to be terminated or forwarded in a lower priority queue.

For the flows that have to be terminated a release procedure, see Section 4.6.2.4, is initiated to release the reserved resources on the "forward" and "reverse" paths.

4.6.2.6 Admission control using congestion notification based on probing

This section describes the admission control scheme that uses the congestion notification function based on probing when bi-directional reservations are supported.

This procedure is similar to the congestion notification for admission control procedure described in Section 4.6.1.7. The main difference is related to the location of the severe congested node, i.e., "forward" path (i.e., path between QNE Ingress towards QNE Egress) or "reverse" path (i.e., path between QNE Egress towards QNE Ingress).

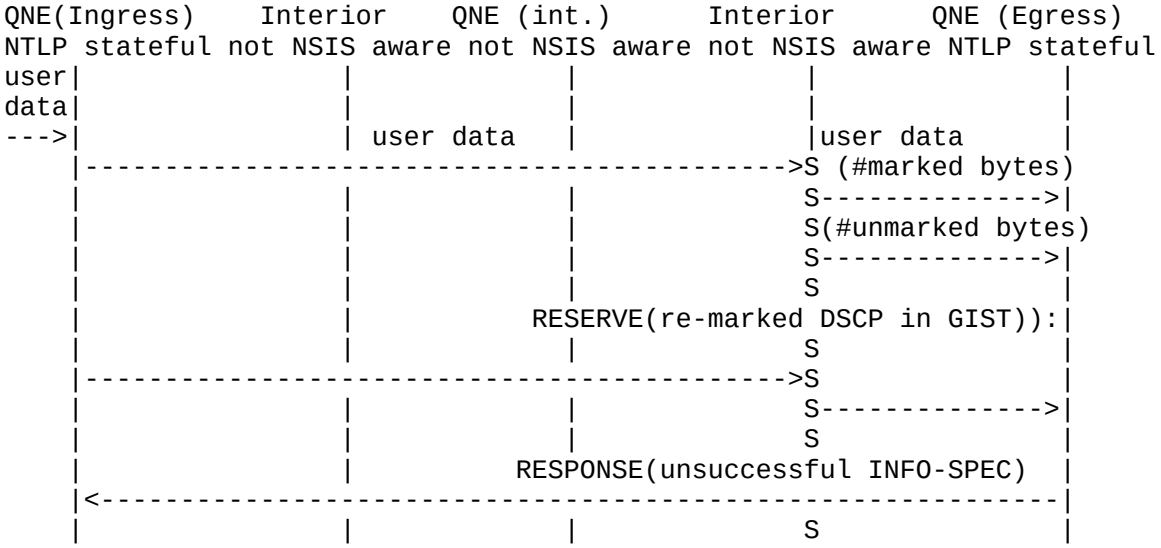

Figure 22: Intra-domain RMD congestion notification based on probing for bi-directional admission control (congestion on path from QNE(Ingress) towards QNE(Egress))

Figure 22 shows the scenario where the severe congested node is located in the "forward" path. The functionality of providing admission control is very similar to the one described in Section 4.6.1.7, Figure 15.

Figure 23 shows the scenario where the congested node is located in the "reverse" path. The probe RESERVE message sent in the "forward" direction will not be affected by the severe congested node, while the DSCP value in the IP header of the GIST message that carries the probe RESERVE message sent in the "reverse" direction will be remarked by the congested node. The QNE ingress is in this way notified that a congestion occurred in the network and therefore it is able to refuse the new initiation of the reservation.

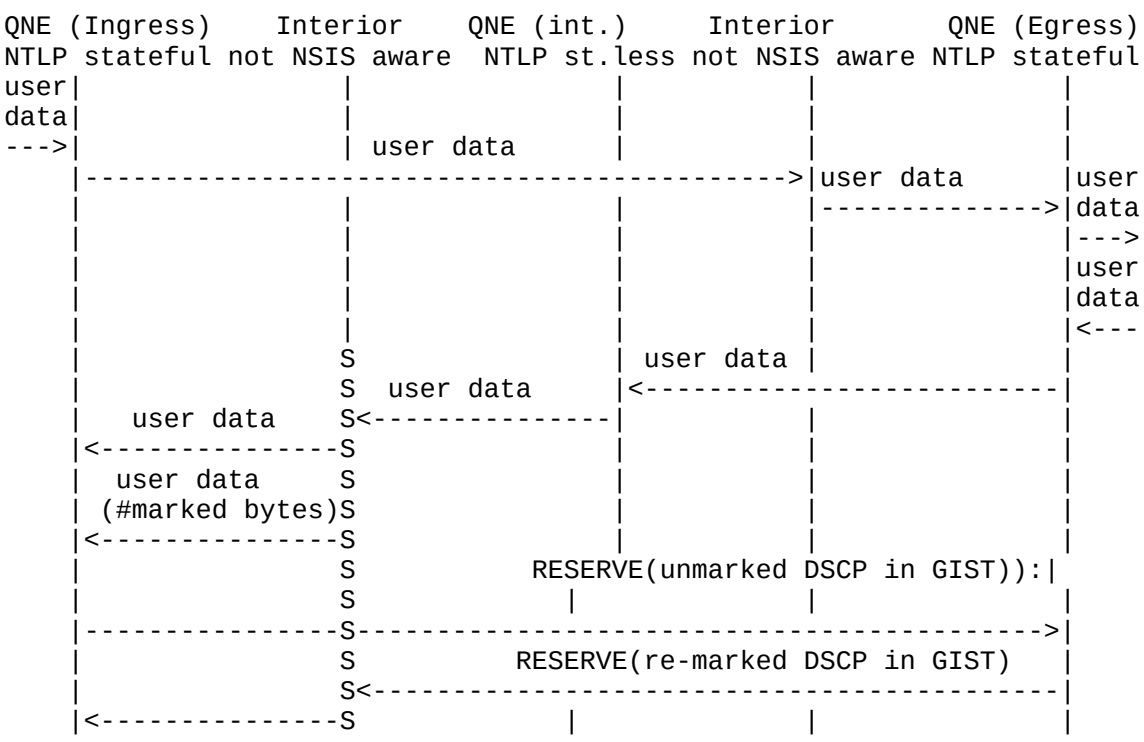

Figure 23: Intra-domain RMD congestion notification for bi-directional admission control (congestion on path QNE(Egress) towards QNE(Ingress))

4.7 Handling of additional errors

During the QSpec processing, additional errors may occur. The way of how these additional errors are handled and notified is specified in [QSP-T] and [QoS-NSLP].

5. Security Considerations

A router implementing a QoS signaling protocol can, similar to a router without QoS signaling, do a lot of harm to a system. If taken over by an adversary, a router can delay, drop, inject, duplicate or modify packets. Additional threats are, however, introduced with new protocols and they are subject for a discussion below.

The RMD-QOSM aims to be very lightweight signaling with regard to the number of signaling message roundtrips and the amount of state established at involved signaling nodes with and without reduced state on QNEs. This implies the usage of the Datagram Mode which does not allow channel security to be used. As such, RMD signaling is targeted towards intra-domain signaling only.

In the context of RMD-QOSM signaling a classification between on-path adversaries and off-path adversaries needs to be made. Furthermore, it might be necessary to differentiate between off-path nodes that never participate in the RMD signaling exchange and nodes that are only off-path with regard to a specific signaling session whereby routing asymmetry might even mean that the downstream and the upstream signaling direction matters for this classification.

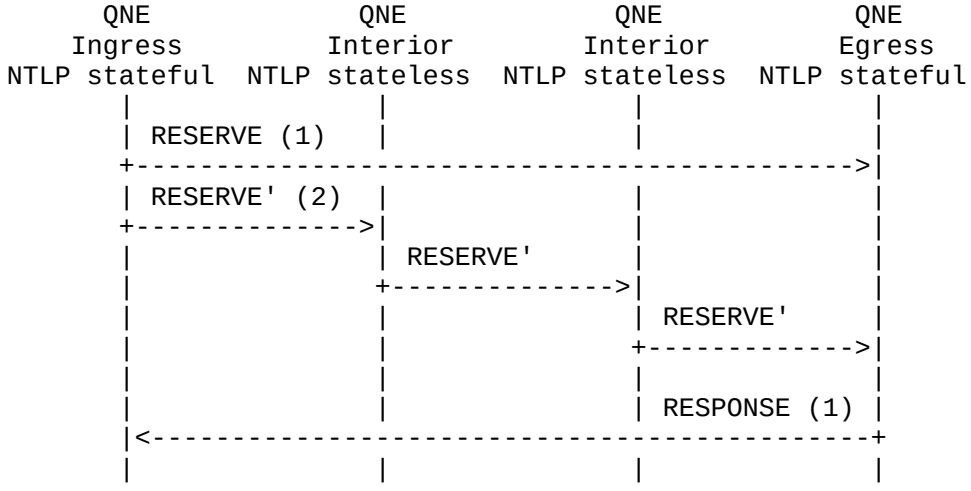

Figure 24: RMD message exchange

Note that RMD always uses the message exchange shown in Figure 24 even if there is no end-to-end signaling session. If the RMD-QOSM is triggered based on an E2E signaling exchange then the RESERVE message is created by a node outside the RMD domain and will subsequently travel further on (e.g., to the data receiver). Such an exchange is shown in Figure 3. As such, an evaluation of RMD's security must always been seen as a combination of the two signaling sessions, (1) and (2) of Figure 24.

The following security requirements are set as goals for the intra-domain communication, namely:

* Nodes, which are never supposed to participate in the NSIS signaling exchange, SHOULD NOT interfere with QNE Interior nodes. Off-path nodes (off-path with regard to the path taken by a particular signaling message exchange) SHOULD NOT be able to interfere with other on-path signaling nodes.
* The actions allowed by a QNE Interior node SHOULD be minimal (i.e., only those specified by the RMD-QOSM). For example, only the QNE Ingress and the QNE Egress nodes are allowed to initiate certain signaling messages. QNE Interior nodes are, for example, allowed to modify certain signaling message payloads.

Note that the term 'interfere' refers to all sorts of security threats, such as denial of service, spoofing, replay, signaling message injection, etc.

If we assume that the RESERVE/RESPONSE is sent in C-Mode and protected between the QNE Ingress and the QNE Egress node then we can be sure that the payloads of these messages MUST be authenticated, integrity, replay protected and encrypted. Encryption is necessary to prevent an adversary that is located along the path of the RESERVE message to learn information about the session that can later be used to inject a valid RESERVE'. The following messages need to relate to each other to make sure that the occurrence of one message is not without the other one:

a) the RESERVE and the RESERVE' relate to each other at the QNE Egress and

b) the RESPONSE and the RESERVE relate to each other at the QNE Ingress and

c) the RESERVE' and the RESPONSE' (carried in the RESPONSE) relate to each other

The RESERVE and the RESERVE' message are tied together using the BOUND_SESSION_ID. Hence, there cannot be a RESERVE' without a corresponding RESERVE. The SESSION_ID can fulfill this purpose quite well if the aim is to provide protection against off-path adversaries that do not see the SESSION_ID carried in the RESERVE and the RESERVE' messages. If, however, the path changes (due to re-routing or due to mobility) then an adversary could inject RESERVE' messages (with a previously seen SESSION_ID) and could potentially cause harm.

An off-path adversary can, of course, create RESERVE' messages that cause intermediate nodes to create some state (and cause other actions) but the message would finally hit the QNE Egress node. The QNE Egress node would then be able to determine that there is something going wrong.

The severe congestion handling can be triggered by intermediate nodes (unlike other messages). In many cases, however, intermediate nodes experiencing congestion use refresh messages modify the <S> and <Overload %> parameters of the message. These messages are still initiated by the QNE Ingress node and carry the SESSION_ID. The QNE Egress node will use the SESSION_ID and subsequently the BOUND_SESSION_ID to refer to a flow that might be terminated. The aspect of intermediate nodes initiating messages for severe congestion handling is for further study.

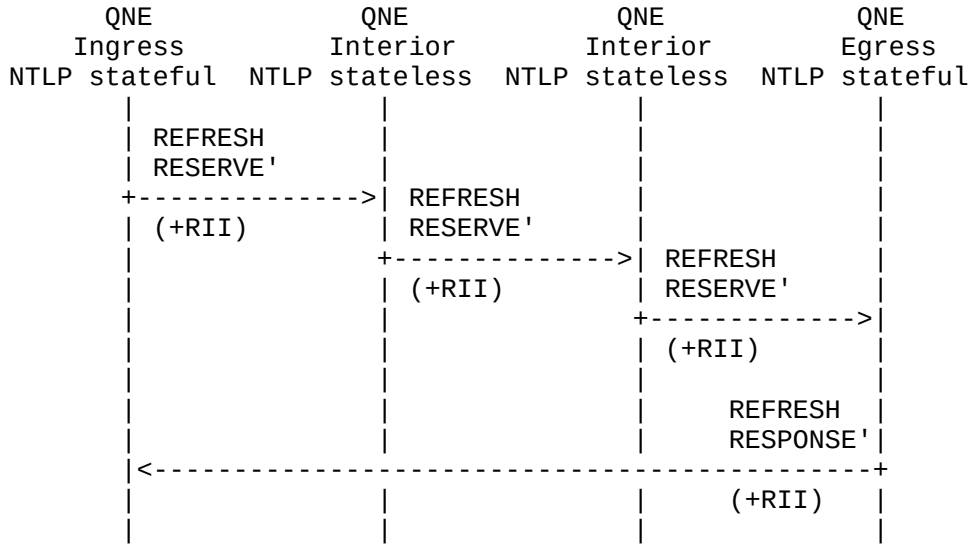

Figure 25: RMD REFRESH message exchange

During the refresh procedure a RESERVE' creates a RESPONSE', see Figure 25. The RII is carried in the RESERVE' message and the RESPONSE' message that is generated by the QNE Egress node contains the same RII as the RESERVE'.

The RII can be used by the QNE Ingress to match the RESERVE' with the RESPONSE'. The QNE Egress is able to determine whether the RESERVE' (as a refresh) was created by the QNE Ingress node since the BOUND_SESSION_ID is included in the RESERVE' message.

With the initial RESERVE'/RESERVE exchange there is a one-to-one mapping between the RESERVE and the RESERVE' message based on the SESSION_ID that is used in the two messages and the BOUND_SESSION_ID. With the REFRESH' message this is not the case since they relate to one RESERVE message exchange.

A further aspect is marking of data traffic. Data packets can be modified by an intermediary without any relationship to a signaling session (and a SESSION_ID). The problem appears if an off-path adversary injects spoofed data packets. The adversary thereby needs to spoof data packets that relate to the flow identifier of an existing end-to-end reservation that should be terminated. Therefore the question arises how an off-path adversary should create a data packet that matches an existing flow identifier (if a 5-tuple is used). Hence, this might not turn out to be simple for an adversary unless we assume the previously mentioned mobility/re-routing case where the path through the network changes and the set of nodes that are along a path changes over time.

6. IANA Considerations

RMD-QOSM requires a new IANA registry for RMD QoS Model Identifiers. It is a 32-bit value carried in a QSPEC object [QSP-T].

RMD-QOSM defines 2 new objects for the QSPEC Template: PHR container and PDR container, see 4.1.2 and 4.1.3. For these new containers, new IDs in the QSPEC Template Object Type registry should be assigned.

7. Acknowledgments

The authors express their acknowledgement to people who have worked on the RMD concept: Z. Turanyi, R. Szabo, G. Pongracz, A. Marquetant, O. Pop, V. Rexhepi, G. Heijenk, D. Partain, M. Jacobsson, S. Oosthoek, P. Wallentin, P. Goering, A. Stienstra, M. de Kogel, M. Zoumaro-Djayoon, M. Swanink, R. Klaver G. Stokkink, J. W. van Houwelingen, D. Dimitrova

8. Authors' Addresses

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

Lars Westberg
Ericsson Research
Torshamnsgatan 23
SE-164 80 Stockholm, Sweden
EMail: Lars.Westberg@ericsson.com

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

Cornelia Kappler
Siemens AG
Siemensdamm 62
Berlin 13627, Germany
Email: cornelia.kappler@siemens.com

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

Tom Phelan
Sonus Networks
250 Apollo Dr.
Chelmsford, MA USA 01824
EMail: tphelan@sonusnet.com

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

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

9. Normative References

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

[QoS-NSLP] Manner, J., Karagiannis, G.,McDonald, A., Van de Bosch, S., "NSLP for Quality-of-Service signaling", draft-ietf-nsis-qos-nslp (work in progress).

[QSP-T] Ash, J., Bader, A., Kappler C., "QoS-NSLP QSpec Template" draft-ietf-nsis-QSpec (work in progress).

10. Informative References

[CsTa05] Csaszar, A., Takacs, A., Szabo, R., Henk, T., "Resilient Reduced-State Resource Reservation", Journal of Communication and Networks, Vol. 7, Nr. 4, December 2005.

[JaSh97] Jamin, S., Shenker, S., Danzig, P., "Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service", Proceedings IEEE Infocom '97, Kobe, Japan, April 1997

[GrTs03] Grossglauser, M., Tse, D.N.C, "A Time-Scale Decomposition Approach to Measurement-Based Admission Control", IEEE/ACM Transactions on Networking, Vol. 11, No. 4, August 2003

[RFC2961] Berger, L., Gan, D., Swallow, G., Pan, P., Tommasi, F. and S. Molendini, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, April 2001.

[RFC3175] Baker, F., Iturralde, C. Le Faucher, F., Davie, B., "Aggregation of RSVP for IPv4 and IPv6 Reservations", IETF RFC 3175, 2001.

[RFC4125] Le Faucheur & Lai, "Maximum Allocation Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering", RFC 4125, June 2005.

[RFC4127] Le Faucheur et al, Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering, RFC 4127, June 2005

[GIST] Schulzrinne, H., Hancock, R., "GIST: General Internet Messaging Protocol for Signaling", draft-ietf-nsis-ntlp (work in progress).

[RFC1633] Braden R., Clark D., Shenker S., "Integrated Services in the Internet Architecture: an Overview", RFC 1633

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

[RFC2638] Nichols K., Jacobson V., Zhang L. "A Two-bit Differentiated Services Architecture for the Internet", RFC 2638, July 1999

[RMD1] Westberg, L., et al., "Resource Management in Diffserv (RMD): A Functionality and Performance Behavior Overview", IFIP PFHSN'02

[RMD2] G. Karagiannis, et al., "RMD - a lightweight application of NSIS" Networks 2004, Vienna, Austria.

[RMD3] Marquetant A., Pop O., Szabo R., Dinnyes G., Turanyi Z., "Novel Enhancements to Load Control - A Soft-State, Lightweight Admission Control Protocol", Proc. of the 2nd Int. Workshop on Quality of Future Internet Services, Coimbra, Portugal, Sept 24-26, 2001, pp. 82-96.

[RMD4] A. Csaszar et al., "Severe congestion handling with resource management in diffserv on demand", Networking 2002

Appendix A.1.1 Example of a remarking operation during severe congestion in the Interior nodes

Per supported PHB, the interior node can support the operation states depicted in Figure A.1, when the per-flow congestion notification based on probing signaling scheme is used in combination with this severe congestion type. Figure A.2 depicts the same functionality when the per-flow congestion notification based on probing scheme is not used in combination with the severe congestion scheme.

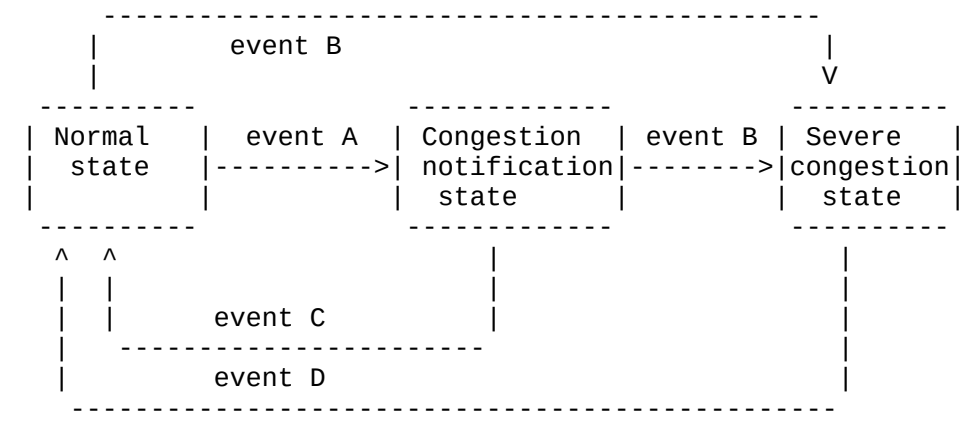

Figure A.1: States of operation, severe congestion combined with congestion notification based on probing

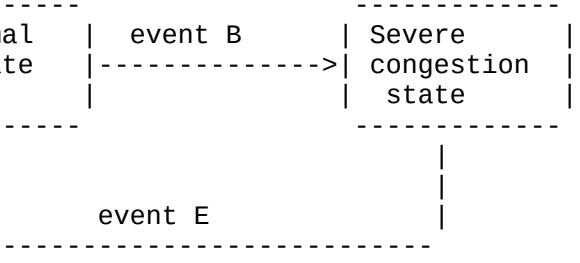

Figure A.2: States of operation, severe congestion without congestion notification based on probing

The terms used in Figure A.1 and Figure A.2 are:

Normal state: represents the normal operation conditions of the node, i.e. no congestion

Severe congestion state: it represents the state when state the interior node is severely congested related to a certain PHB

Congestion notification: state where the load is relatively high, close to the level when congestion can occur

event A: this event occurs when the incoming PHB rate is higher than the "congestion notification detection" threshold. This threshold is used by the congestion notification based on probing scheme, see Section 4.6.1.7, 4.6.2.6.

event B: this event occurs when the incoming PHB rate is higher than the "severe congestion detection" threshold.

event C: this event occurs when the incoming PHB rate is lower than the "congestion notification detection" threshold.

event D: this event occurs when the incoming PHB rate is lower than the "severe_congestion_restoration" threshold.

event E: this event occurs when the incoming PHB rate is lower than the "severe congestion restoration" threshold.

Note that the "severe congestion detection", "severe congestion restoration" and admission thresholds should be higher than the "congestion notification detection" threshold, i.e.,:
"severe congestion detection" > "congestion notification detection" and "severe congestion restoration" > "congestion notification detection"

Furthermore, the "severe congestion detection" threshold should be higher than or equal to the admission threshold that is used by the reservation based and NSIS measurement based signaling schemes.
"severe congestion detection" >= admission threshold

Moreover, the "severe congestion restoration" threshold should be lower than or equal to the "severe congestion detection" threshold that is used by the reservation based and NSIS measurement based signaling schemes, i.e.,:

"severe congestion restoration" <= "severe congestion detection"

During severe congestion the interior node calculates, per traffic class (PHB), the incoming rate that is above the "severe congestion restoration" threshold, denoted as signaled_overload_rate, in the following way:

* A severe congested interior node should take into account that packets might be dropped. Therefore, before queuing and eventually dropping packets, the interior node should count the total number of unmarked and remarked bytes received by the severe congested node, denote this number as total_received_bytes. Note that there are situations when more than one interior nodes in the same path become severe congested. Therefore, any interior node located behind a severe congested node may receive marked bytes.

* before queuing and eventually dropping the packets, at the end of each measurement interval of T seconds, calculate the current estimated overloaded rate, say measured_overload_rate, by using the following equation:

measured_overload_rate =
=((total_received_bytes)/T) - severe_congestion_restoration)

Note that since marking is done in interior nodes, the decisions are made at egress nodes, and termination of flows are performed by ingress nodes, there is a significant delay until the overload information is learned by the ingress nodes, see Section 6 of [CsTa05]). The delay consists of the trip time of data packets from the severe congested interior node to the egress, the measurement interval, i.e., T, and the trip time of the notification signaling messages from egress to ingress. Moreover, until the overload decreases at the severe congested interior node, an additional trip time from the ingress node to the severe congested interior node must expire. This is because immediately before receiving the congestion notification, the ingress may have sent out packets in the flows that where selected for termination. That is, a terminated flow may contribute to congestion for a time longer that is taken from the ingress to the interior node. Without considering the above, interior nodes would continue marking the packets until the measured utilization falls below the severe congestion restoration threshold. In this way, in the end more flows will be terminated than necessary, i.e., an over-reaction takes place. [CsTa05] provides a solution to this problem, where the interior nodes use a sliding window memory to keep track of the signaling overload in a couple of previous measurement intervals. At the end of a measurement intervals, T, before encoding and signaling the overloaded rate as "encoded DSCP" packets, the actual overload is decreased with the sum of already signaled overload stored in the sliding window memory, since that overload is already being handled in the severe congestion handling control loop. The sliding window memory consists of an integer number of cells, i.e, n = maximum number of cells. Guidelines for configuring the sliding window parameters are given in [CsTa05].

At the end of each measurement interval, the newest calculated overload is pushed into the memory, and the oldest cell is dropped.

If Mi is the overload_rate stored in ith memory cell (i = [1..n]), then at the end of every measurement interval, the overload rate that is signaled to the egress node, i.e., signaled_overload_rate is calculated as follows:

```
Sum_Mi =0
For i =1 to n
{
Sum_Mi = Sum_Mi + Mi
}
```

signaled_overload_rate = measured_overload_rate - Sum_Mi,

where Sum_Mi is calculated as above.

Next, the sliding memory is updated as follows:

```
    for i = 1..(n-1): Mi <- Mi+1
    Mn <- signaled_overload_rate
```

The bytes that have to be remarked to satisfy the signaled overload rate: signaled_remarked_bytes, are calculated as follows:

signaled_remarked_bytes = signaled_overload_rate*T/N

The signal_remarked_bytes represents also the number of the outgoing packets (after the dropping stage) that must be remarked, during each measurement interval T, by a node when operates in severe congestion mode.

Note that in order to process an overload situation higher than 100% of the maintained severe congestion threshold all the nodes within the domain MUST be configured and maintain a scaling parameter, e.g., N used in the above equation, which in combination with the marked bytes, e.g., signaled_remarked_bytes, such a high overload situation ca be calculated and represented.

Note that when incoming remarked bytes are dropped, the operation of the severe congestion algorithm may be affected, e.g., the algorithm may become in certain situations slower. An implementation of the algorithm may assure as much as possible that the incoming marked bytes are not dropped. This could for example be accomplished by using different dropping rate thresholds for marked and unmarked bytes.

Note that when the "affected DSCP" marking is applied by a severe congested node then all the outgoing packets that are not marked (i.e., by using the "encoded DSCP") have to be remarked using the "affected DSCP" code. Furthermore, note that when the congestion notification based on probing is used in combination with severe congestion, then in addition to the possible "encoded DSCP" and "affected DSCP" another DSCP for the remarking of the same PHB might be used, see Section 4.6.1.7. This additional DSCP might be denoted in this document as "notified DSCP". When an interior node operates in the severe congested state, see Figure A.2, and receives "notified DSCP" packets, these packets are considered to be unmarked packets (but not "affected DSCP" packets).

Appendix A.1.2 Example of a detailed severe congestion operation in the Egress nodes

The states of operation in Egress nodes are similar to the ones described in A.1.1. The definition of the events, see below, is how ever different than the definition of the events given in Figure A.1 and Figure A.2:

* event A: the egress node measures the rate of the incoming "notified_DSCP" marked packets and compare it with a predefined congestion notification detection threshold at the egress. When the measured rate of "notified DSCP" bytes is higher than this threshold then event_A is activated, see Section 4.6.1.7 and A.2.2. This is applied when the whole RMD domain uses "notified DSCP" for this purpose. If the "notified DSCP" marking is not used in the whole RMD domain, the "encoded_DSCP" marking is used to notify the congestion notification state. In this case the egress should measure the rate of the incoming "encoded_DSCP" marked packets and compare it with a predefined congestion notification detection threshold and to a severe congestion detection threshold in the egress. Note that the detection thresholds used in the egress for congestion notification and severe congestion may be different than the ones used in interior nodes. When the measured rate of "encoded DSCP" bytes is higher than the congestion notification threshold but lower than the severe congestion threshold then event_A is activated.

* event B: this event occurs when the egress receives packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain). However, when the "encoded_DSCP" marking is also used for congestion notification detection purposes, see description of event_A, then event_B is only activated if either "affected DSCP" packets are received or if the rate of the incoming "encoded_DSCP" marked packets is higher than the preconfigured severe congestion detection egress threshold.

* event C: this event occurs when the rate of incoming "notified DSCP" packets decreases below the congestion notification detection threshold. This is applied when whole RMD domain uses "notified DSCP" for this purpose. When the "encoded_DSCP" marking is also used for congestion notification detection, see description of event_A, then event_C is activated when the rate of incoming "encoded DSCP" packets decreases below the congestion notification threshold.

* event D: this event occurs when the egress does not receive packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain). When the "encoded_DSCP" marking is also used for congestion notification detection, see description of event_A, event_B, event_C, then the event_D is only activated if either "affected DSCP" packets are not anymore received or if the rate of the incoming "encoded_DSCP" marked packets is slower than the preconfigured severe congestion restoration threshold in egress.

* event E: this event occurs when the egress does not receive packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain)

An example of the algorithm for calculation of the number of flows associated with each priority class that have to be terminated is explained by the pseudocode below.

First, when the egress operates in the severe congestion state then the total amount of remarked bandwidth associated with the PHB traffic class, say total_congested_bandwidth, is calculated. Note that when the node maintains information about each ingress/egress pair aggregate, then the total_congested_bandwidth must be calculated per ingress/egress pair aggregate. This bandwidth represents the severe congested bandwidth that should be terminated. The total_congested_bandwidth can be calculated as follows:

total_congested_bandwidth = N*input_remarked_bytes/T

Where, input_remarked_bytes represents the number of marked bytes that arrive at the ingress, during one measurement interval T, N is defined as in Section 4.6.1.6.2.1. The term denoted as terminated_bandwidth is a temporal variable representing the total bandwidth that have to be terminated, belonging to the same PHB traffic class. The terminate_flow_bandwidth(priority_class) is the total of bandwidth associated with flows of priority class equal to priority_class. The parameter priority_class is an integer fulfilling

0 < priority_class =< Maximum_priority.

The calculate_terminate_flows(priority_class) function determines the Flows for a given priority class and per PHB that has to be Terminated. This function also calculates the term sum_bandwidth_terminate(priority_class), which is the sum of the bandwith associated with the flows that will be terminated. The constraint of finding the total number of flows that have to be terminated is that sum_bandwidth_terminate(priority_class), should be smaller or approximatelly equal to the variable terminate_bandwidth(priority_class).

```
terminated_bandwidth = 0;
priority_class = 0;
while terminated_bandwidth < total_congested_bandwidth
 {
  terminate_bandwidth(priority_class) =
  = total_congested_bandwidth - terminated_bandwidth
  calculate_terminate_flows(priority_class);
  terminated_bandwidth =
  = sum_bandwidth_terminate(priority_class) + terminated_bandwidth;
  priority_class = priority_class + 1;
 }
```

If the egress node maintains ingress/egress pair aggregates, then the above algorithm is performed for each ingress/egress pair aggregate.

Appendix A.2.1 Example of a detailed remarking admission control (congestion notification) operation in Interior nodes

In particular, the predefined congestion notification threshold is set according to, and usually less than, an engineered bandwidth limitation, i.e., admission threshold, based on e.g. agreed Service Level Agreement or a capacity limitation of specific links.

The difference between the congestion notification threshold and the engineered bandwidth limitation, i.e., admission threshold, provides an interval where the signaling information on resource limitation is already sent by a node but the actual resource limitation is not reached. This is due to the fact that data packets associated with an admitted session have not yet arrived, while allows the admission control process available at the egress to interpret the signaling information and reject new calls before reaching congestion. Note that in the situation when the data rate is higher than the preconfigured congestion notification rate, also data packets are re-marked, see section 4.6.1.6.2.1. To distinguish between congestion notification and severe congestion, two methods may be used (see Appendix 1.1.1):

* using different DSCP values (re-marked DSCP values). The remarked DSCP that is used for this purpose is denoted as "notified DSCP" in this document. When this method is used and when the interior node is in "congestion notification" state, see A.1.1, then the node should remark the unmarked bytes using the "notified DSCP". Note that this method can only be applied if all nodes in RMD domain use the "notified" DSCP marking.

* Using the "encoded DSCP" marking for congestion notification and severe congestion. This situation is applied when the "notified DSCP" marking is not applied in the RMD domain. When this method is used and when the interior node is in "congestion notification" state, see A.1.1, then the node should remark the unmarked bytes using the "encoded DSCP".

Note that if a node starts dropping packets belonging to a PHB that suports both "severe congestion" and "congestion notification" states, see section 4.6.1.6.2.1, then it is considered that the packet rate associated to this PHB is higher than the severe congestion detection threshold and that the operation state of this node has moved to the severe congestion state, see Appendix A.1.1.

Appendix A.2.2 Example of a detailed admission control (congestion notification) operation in Egress nodes

The admission control congestion notification procedure can be applied only if the egress maintais the ingress/egress pair aggregate. When the operation state of the ingress/egress pair aggregate is the "congestion notification", see Appendix A.1.2, then the implementation of the algorithm depends on how the congestion notification situation is notified to the egress. As mentioned in Section A.2.1, two methods are used:

* using the "notified DSCP". During a measurement interval T, the egress counts the number of "notified DSCP" marked bytes that belong to the same PHB and are associated with the same ingress/egress pair aggregate, say input_notified_bytes. We denote the rate as incoming_notified_rate.

* using the "encoded DSCP". In this case, during a measurement interval T, the egress measures the input_notified_bytes by counting instead of the "notified DSCP", the "encoded DSCP" bytes.

The incoming congestion_rate can be then calculated as follows:

incoming_congestion_rate = N*input_notified_bytes/T

If the incoming_congestion_rate is higher than a preconfigured congestion notification threshold, then the communication path between ingress and egress is considered to be congested. In this situation if the end-to-end RESERVE (probe) arrives at the egress, then this request SHOULD be rejected. Note that this choice is independent of the DSCP marking status of the packet that carries the RESERVE message.
If such an ingress/egress pair aggregated state is not available when the (probe) RESERVE message arrives at the egress, then this request is accepted if the DSCP of the packet carrying the RESERVE messsage is unmarked. Otherwise (if the packet is either "notified DSCP" or "encoded DSCP" marked), it is rejected.

Appendix A.3.1 Example of selecting bi-directional flows for termination during severe congestion

When a severe congestion occurs on e.g., in the forward path, and when the algorithm terminates flows to solve the severe congestion in forward path, then the reserved bandwidth associated with the terminated bidirectional flows is also released. Therefore, a careful selection of the flows that have to be terminated should take place. A possible method of selecting the flows belonging to the same priority type passing through the severe congestion point on a unidirectional path can be the following:

* the egress node should select, if possible, first unidirectional flows instead of bidirectional flows
* the egress node should select, if possible, bidirectional flows that reserved a relatively small amount of resources on the path reversed to the path of congestion.

Appendix A.3.2 Example of a severe congestion solution for bi-directional flows congested simultaneously on forward and reverse path

This scenario describes a solution using the combination of the severe congestion solutions described in Section 4.6.2.5.2.
It is considered that the severe congestion occurs simultaneously on forward and reverse directions, which may affect the same bi-directional flows. This situation is depicted in Figure A.3. Consider that the egress node selects a number of bi-directional flows to be terminated. In this case the egress will send for each bi-directional flows a NOTIFY message to ingress. If the Ingress receives these NOTIFY messages and its operational state (associated with reverse path) is in the severe congestion state (see Figure A.1 and A.2), then the ingress operates in the following way:

* For each NOTIFY message, the Ingress should identify the bidirectional flows have to be terminated.

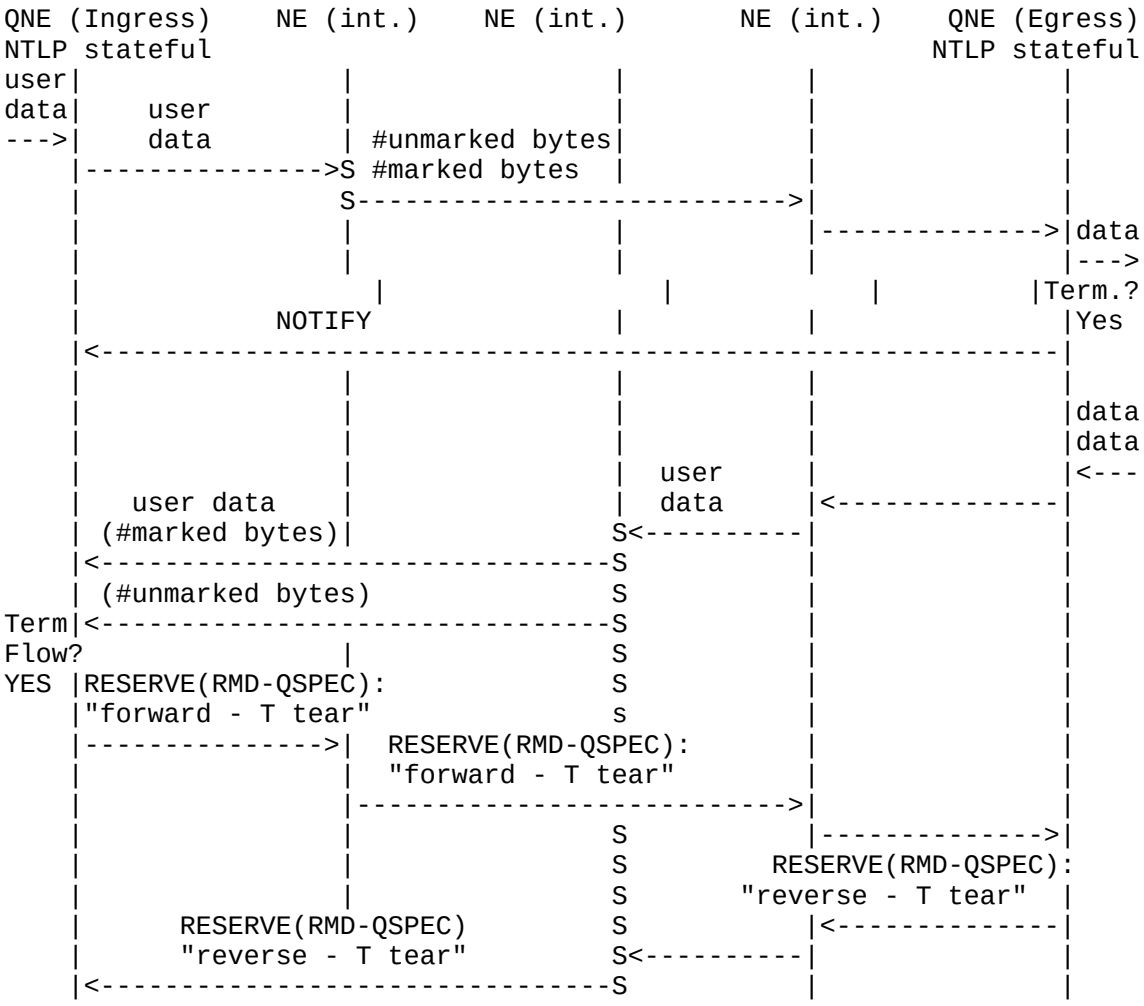

Figure A.3: Intra-domain RMD severe congestion handling for bi-directional reservation (congestion on both forward and reverse direction)

* The ingress then calculates the total bandwidth that should be released in the reverse direction (thus not in forward direction) if the bidirectional flows will be terminated (preempted), say "notify_reverse_bandwidth".

* Furthermore, using the received marked packets (from the reverse path) the ingress will calculate, using the algorithm used by an egress and described in A.1.2, the total bandwidth that has to be terminated in order to solve the congestion in the reverse path direction, say "marked_reverse_bandwidth".

* The ingress then calculates the bandwidth of the additional flows that have to be terminated, say "additional_reverse_bandwidth", in order to solve the severe congestion in reverse direction, by taking into account:

** the bandwidth in the reverse direction of the bidirectional flows that were appointed by the egress (the ones that received a NOTIFY message) to be preempted, i.e., "notify_reverse_bandwidth"

** the total amount of bandwidth in the reverse direction that has been calculated by using the received marked packets, i.e., "marked_reverse_bandwidth".
This additional bandwidth can be calculated using the following algorithm:

```
 IF ("marked_reverse_bandwidth" > "notify_reverse_bandwidth") THEN
 "additional_reverse_bandwidth" =
  = "marked_reverse_bandwidth"- "notify_reverse_bandwidth";
 ELSE
 "additional_reverse_bandwidth" = 0
```

* Ingress terminates the flows that received a (preemption) NOTIFY message

* If possible the ingress SHOULD terminate unidirectional flows that are using the same egress-ingress reverse direction communication path to satisfy the release of a total bandiwtdh up equal to the: "additional_reverse_bandwidth", see Appendix 3.1.

* If the number of required uni-directional flows (to satisfy the above issue) is not available, then a number of bi-directional flows that are using the same egress-ingress reverse direction communication path MAY be selected for preemption in order to satisfy the release of a total bandiwtdh up equal to the: "additional_reverse_bandwidth". Note that using the guidelines given in Appendix A.3.1, first the bidirectional flows that reserved a relatively small amount of resources on the path reversed to the path of congestion should be selected for termination.

Intellectual Property Statement

Disclaimer of Validity