Internet Engineering Task Force (IETF)
Request for Comments: 5977
Category: Experimental
ISSN: 2070-1721

A. Bader
L. Westberg
Ericsson
G. Karagiannis
University of Twente
C. Kappler
ck technology concepts
T. Phelan
Sonus

October 2010

RMD-QOSM: The NSIS Quality-of-Service Model for Resource Management in Diffserv

Abstract

This document describes a Next Steps in Signaling (NSIS) Quality-of-Service (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.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc5977.

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], and [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) [RFC5974] specifies a generic protocol for carrying 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) specified by the QSPEC template [RFC5975] 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.

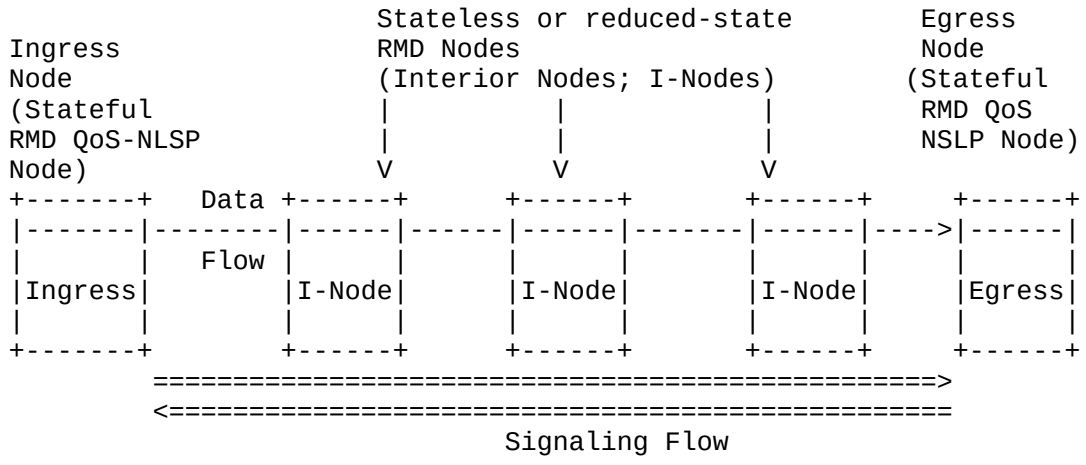

Figure 1: Actors in the RMD-QOSM

Many network scenarios, such as the "Wired Part of Wireless Network" scenario, which is described in Section 8.4 of [RFC3726], require that the impact of the used QoS signaling protocol on the network performance should be minimized. In such network scenarios, the performance of each network node that is used in a communication path

has an impact on the end-to-end performance. As such, the end-to-end performance of the communication path can be improved by optimizing the performance of the Interior nodes. One of the factors that can contribute to this optimization is the minimization of the QoS signaling protocol processing load and the minimization of the number of states on each Interior node.

Another requirement that is imposed by such network scenarios is that whenever a severe congestion situation occurs in the network, the used QoS signaling protocol should be able to solve them. In the case of a route change or link failure, a severe congestion situation may occur in the network. Typically, routing algorithms are able to adapt and change their routing decisions to reflect changes in the topology and traffic volume. In such situations, the rerouted traffic will have to follow a new path. Interior nodes located on this new path may become overloaded, since they suddenly might need to support more traffic than for which they have capacity. These severe congestion situations will severely affect the overall performance of the traffic passing through such nodes.

RMD-QOSM is an edge-to-edge (intra-domain) QoS Model that, in combination with the QoS-NSLP and QSPEC specifications, is designed to support the requirements mentioned above:

- Minimal impact on Interior node performance;
- Increase of scalability;
- Ability to deal with severe congestion

Internally to the RMD network, RMD-QOSM together with QoS-NSLP [RFC5974] defines a scalable QoS signaling model in which per-flow QoS-NSLP and NSIS Transport Layer Protocol (NTLP) states are not stored in Interior nodes but per-flow signaling is performed (see [RFC5974]) at the Edges.

In the RMD-QOSM, only routers at the Edges of a Diffserv domain (Ingress and Egress nodes) support the (QoS-NSLP) stateful operation; see Section 4.7 of [RFC5974]. Interior nodes support either the (QoS-NSLP) stateless operation or a reduced-state operation with coarser granularity than the Edge nodes.

After the terminology in Section 2, we give an overview of RMD and the RMD-QOSM in Section 3. This document specifies several RMD-QOSM/QoS-NSLP signaling schemes. In particular, Section 3.2.3 identifies which combination of sections are used for the specification of each RMD-QOSM/QoS-NSLP signaling scheme. In Section 4 we give a detailed description of the RMD-QOSM, including the role of QoS NSIS entities

(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 [RFC2119].

The terminology defined by GIST [RFC5971] and QoS-NSLP [RFC5974] applies to this document.

In addition, the following terms are used:

NSIS domain: an NSIS signaling-capable domain.

RMD domain: an NSIS domain that is capable of supporting the RMD-QOSM signaling and operations.

Edge node: a QoS-NSLP node on the boundary of some administrative domain that connects one NSIS domain to a node in either another NSIS domain or a non-NSIS domain.

NSIS-aware node: a node that is aware of NSIS signaling and RMD-QOSM operations, such as severe congestion detection and Differentiated Service Code Point (DSCP) marking.

NSIS-unaware node: a node that is unaware of NSIS signaling, but is aware of RMD-QOSM operations such as severe congestion detection and DSCP marking.

Ingress node: an Edge node in its role in handling the traffic as it enters the NSIS domain.

Egress node: an Edge node in its role in handling the traffic as it leaves the NSIS domain.

Interior node: a node in an NSIS domain that is not an Edge node.

Congestion: a temporal network state that occurs when the traffic (or when traffic associated with a particular Per-Hop Behavior (PHB)) passing through a link is slightly higher than the capacity allocated for the link (or allocated for the particular PHB). If no measures are taken, then the traffic passing through this link may temporarily slightly degrade in QoS. This type of congestion is usually solved using admission control mechanisms.

Severe congestion: the congestion situation on a particular link within the RMD domain where a significant increase in its real packet queue situation occurs, such as when due to a link failure rerouted traffic has to be supported by this particular link.

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 short active 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. Typically, 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.

The possible RMD-QOSM applicabilities are described in Section 3.2.3. Two main basic admission control modes are supported: reservation-based and measurement-based admission control that can be used in

combination with a severe congestion-handling solution. The severe congestion-handling solution is used in the situation that a link/node becomes severely congested due to the fact that the traffic supported by a failed link/node is rerouted and has to be processed by this link/node. Furthermore, RMD-QOSM supports both unidirectional and bidirectional reservations.

Another important feature of RMD-QOSM is that the intra-domain sessions supported by the Edges can be either per-flow sessions or per-aggregate sessions. In the case of the per-flow intra-domain sessions, the maintained per-flow intra-domain states have a one-to-one dependency to the per-flow end-to-end states supported by the same Edge. In the case of the per-aggregate sessions the maintained per-aggregate states have a one-to-many relationship to the per-flow end-to-end states supported by the same Edge.

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 PHB traffic classes, and signal changes in the class reservations as necessary. The reservation is quantified in terms of resource units (or bandwidth). 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. The measurement-based algorithm is used to support a predictive service where the service commitment is somewhat less reliable than the service that can be supported by the reservation-based method.

A main assumption that is made by such measurement-based admission control mechanisms is that the aggregated PHB traffic passing through an RMD Interior node is high and therefore, current measurement characteristics are considered to be an indicator of future load. Once an admission decision is made, no record of the decision need be kept at the Interior nodes. The advantage of measurement-based resource management protocols is that they do not require pre-reservation state nor explicit release of the reservations at the Interior nodes. 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. It is important to emphasize that the RMD measurement-based schemes described in this document do not use any refresh procedures, since these approaches are used in stateless nodes; see Section 4.6.1.3.

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 is 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. Note that when the predefined congestion threshold is exceeded, all packets are re-marked by a node, including NSIS messages. In this way, the Edges can admit or reject flows that are requesting resources. The frequency and duration that the congestion level is above the threshold resulting in re-marking is tracked and used to influence the admission control decisions.

* 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. With the measurement-based scheme, the requested peak bandwidth of a flow is carried by the admission control request. The admission decision is considered as 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 a value alpha. Otherwise, the admission decision is negative. It is important to emphasize that due to the fact that the RMD Interior nodes are stateless, they do not store information of previous admission control requests.

This could lead to a situation where the admission control accuracy is decreased when multiple simultaneous flows (sharing a common Interior node) are requesting admission control simultaneously. By applying measuring techniques, e.g., see [JaSh97] and [GrTs03], which use current and past information on NSIS sessions that requested resources from an NSIS-aware Interior node, the decrease in admission control accuracy can be limited. 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 the 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 QoS NSIS initiator (QNI) and QoS NSIS Receiver (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 usually QoS-NSLP-aware nodes. However, in the scenarios where the congestion notification function based on probing is used, then the Interior nodes are not NSIS aware. 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 reduced-state nodes storing per PHB aggregated QoS-NSLP states (for reservation-based 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).

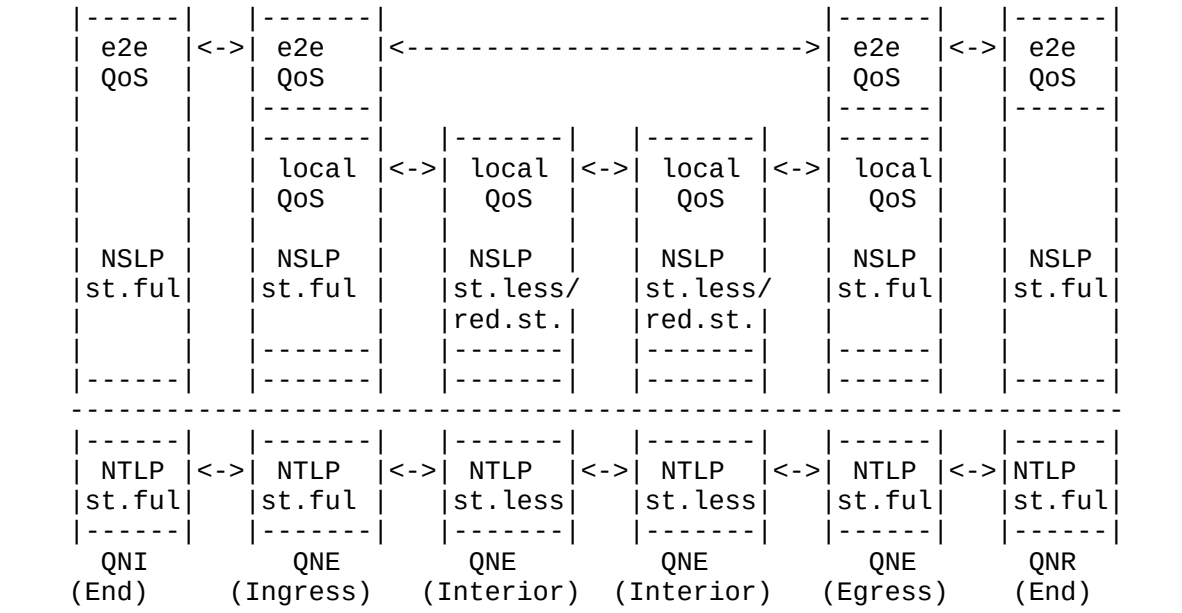

Figure 2: Protocol model of stateless/reduced-state operation

3.2.2. RMD-QOSM/QoS-NSLP Signaling

The basic RMD-QOSM/QoS-NSLP signaling is shown in Figure 3. The signaling scenarios are accomplished using the QoS-NSLP processing rules defined in [RFC5974], in combination with the Resource Management Function (RMF) triggers sent via the QoS-NSLP-RMF API described in [RFC5974].

Due to the fact that within the RMD domain a QoS Model that is different than the end-to-end QoS Model applied at the Edges of the RMD domain can be supported, the RMD Interior node reduced-state reservations can be updated independently of the per-flow end-to-end reservations (see Section 4.7 of [RFC5974]). Therefore, two different RESERVE messages are used within the RMD domain. One RESERVE message that is associated with the per-flow end-to-end reservations and is used by the Edges of the RMD domain and one that is associated with the reduced-state reservations within the RMD domain.

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 initial QSPEC in the message (usually the Initiator QSPEC). The RMD-QSPEC is sent in a intra-domain, independent RESERVE message through the Interior nodes towards the QNR. This intra-domain RESERVE message uses the GIST datagram signaling mechanism. Note that the RMD-QOSM cannot directly specify that the GIST Datagram mode SHOULD be used. This can however be notified by using the GIST API Transfer-Attributes, such as unreliable, low level of security and use of local policy.

Meanwhile, the original RESERVE message is sent to the Egress node on the path to the QNR using the reliable transport mode of NTLP. Each QoS-NSLP node on the data path processes the intra-domain RESERVE message and checks the availability of resources with either the reservation-based or the measurement-based method.

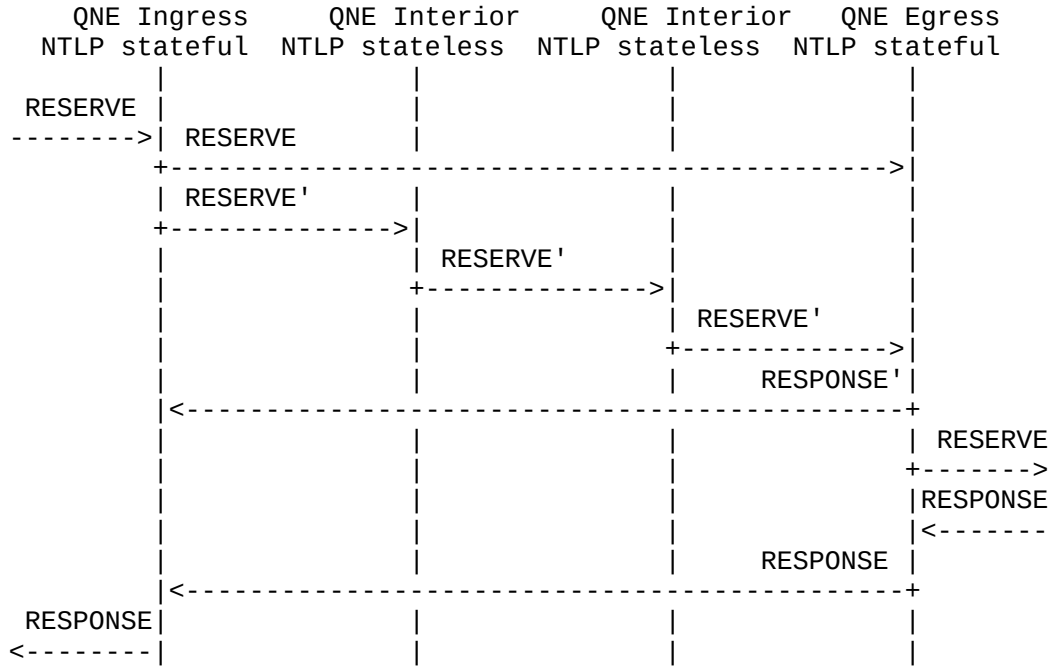

Figure 3: Sender-initiated reservation with reduced-state Interior nodes

When the message reaches the Egress node, and the reservation is successful in each Interior node, an intra-domain (local) RESPONSE' is sent towards the Ingress node and the original (end-to-end) 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, an intra-domain RESPONSE' and an original RESPONSE message are sent directly to the Ingress node.

As a consequence, in the stateless/reduced-state domain only sender-initiated reservations can be performed and functions requiring per-flow NTLP or QoS-NSLP states, like summary and reduced 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 for this is that the RMD-QOSM concept can only support the (Expedited Forwarding) EF-like functionality behavior, but is not able to support the full set of (Assured Forwarding) AF-like functionality. The bandwidth information REQUIRED by the EF-like functionality behavior can be supported by RMD-QOSM carrying the bandwidth information in the <QoS Desired> parameter (see [RFC5975]). The full set of (Assured Forwarding) AF-like functionality requires information that is specified in two token buckets. The RMD-QOSM is not supporting the use of two token buckets and therefore, it is not able to support the full set of AF-functionality. Note however, that RMD-QOSM could also support a single AF PHB, when the traffic or the upper limit of the traffic can be characterized by a single bandwidth parameter. Moreover, it is considered that in case of tunneling, the RMD-QOSM supports only the uniform tunneling mode for Diffserv (see [RFC2983]).

The RMD domain MUST be engineered in such a way that each QNE Ingress maintains information about the smallest MTU that is supported on the links within the RMD domain.

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, an RMD router can never process and use two different RMD-QOSM signaling schemes at the same time.

However, all RMD QNEs supporting this specification MUST support the combination of the "per-flow RMD reservation-based" and the "severe congestion handling by proportional data packet marking" scheme. If the RMD QNEs support more RMD-QOSM schemes, then the operator of that RMD domain MUST preconfigure all the QNE Edge nodes within one domain such that the <SCH> field included in the "PHR container" (Section

4.1.2) and the "PDR Container" (Section 4.1.3) will always use the same value, such that within one RMD domain only one of the below described RMD-QOSM schemes is used at a time.

The congestion situations (see Section 2) are solved using an admission control mechanism, e.g., "per-flow congestion notification based on probing", while the severe congestion situations (see Section 2), are solved using the severe congestion handling mechanisms, e.g., "severe congestion handling by proportional data packet marking".

The RMD domain MUST be engineered in such a way that RMD-QOSM messages could be transported using the GIST Query and DATA messages in Q-mode; see [RFC5971]. This means that the Path MTU MUST be engineered in such a way that the RMD-QOSM message are transported without fragmentation. Furthermore, the RMD domain MUST be engineered in such a way to guarantee capacity for the GIST Query and Data messages in Q-mode, within the rate control limits imposed by GIST; see [RFC5971].

The RMD domain has to be configured such that the GIST context-free flag (C-flag) MUST be set (C=1) for QUERY messages and DATA messages sent in Q-mode; see [RFC5971].

Moreover, the same deployment issues and extensibility considerations described in [RFC5971] and [RFC5978] apply to this document.

It is important to note that the concepts described in Sections 4.6.1.6.2, 4.6.2.5.2, 4.6.1.6.2, and 4.6.2.5.2 contributed to the PCN WG standardization.

The available RMD-QOSM/QoS-NSLP signaling schemes are:

* "per-flow congestion notification based on probing" (see Sections 4.3.2, 4.6.1.7, and 4.6.2.6). Note that this scheme uses, for severe congestion handling, the "severe congestion handling by proportional data packet marking" (see Sections 4.6.1.6.2 and 4.6.2.5.2). Furthermore, the Interior nodes are considered to be Diffserv aware, but NSIS-unaware nodes (see Section 4.3.2).

* "per-flow RMD NSIS measurement-based admission control" (see Sections 4.3.2, 4.6.1, and 4.6.2). Note that this scheme uses, for severe congestion handling, the "severe congestion handling by proportional data packet marking" (see Sections 4.6.1.6.2 and 4.6.2.5.2). Furthermore, the Interior nodes are considered to be NSIS-aware nodes (see Section 4.3.2).

* "per-flow RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure (see Sections 4.3.3, 4.6.1, 4.6.1.6.1, and 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 Sections 4.6.1.6.1 and 4.6.2.5.1). Furthermore, the intra-domain sessions supported by the Edge nodes are per-flow sessions (see Section 4.3.3).

* "per-flow RMD reservation-based" in combination with the "severe the congestion handling by proportional data packet marking" procedure (see Sections 4.3.3, 4.6.1, 4.6.1.6.2, and 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 Sections 4.6.1.6.2 and 4.6.2.5.2). Furthermore, the intra-domain sessions supported by the Edge nodes are per-flow sessions (see Section 4.3.3).

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure (see Sections 4.3.1, 4.6.1, 4.6.1.6.1, and 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 Sections 4.6.1.6.1 and 4.6.2.5.1). Furthermore, the intra-domain sessions supported by the Edge nodes are per-aggregate sessions (see Section 4.3.1). Moreover, this scheme can be considered to be a reservation-based scheme, since the RMD Interior nodes are reduced-state nodes, i.e., they do not store NTLP/GIST states, but they do store per PHB-aggregated QoS-NSLP reservation states.

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure (see Sections 4.3.1, 4.6.1, 4.6.1.6.2, and 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 Sections 4.6.1.6.2 and 4.6.2.5.2). Furthermore, the intra-domain sessions supported by the Edge nodes are per-aggregate sessions (see Section 4.3.1). Moreover, this scheme can be considered to be a reservation-based scheme, since the RMD Interior nodes are reduced-state nodes, i.e., they do not store NTLP/GIST states, but they do store per PHB-aggregated QoS-NSLP reservation states.

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 RMD-QOSM uses the QSPEC format specified in [RFC5975]. The Initiator/Local QSPEC bit, i.e., <I> is set to "Local" (i.e., "1") and the <QSPEC Proc> is set as follows:

* Message Sequence = 0: Sender initiated
* Object combination = 0: <QoS Desired> for RESERVE and <QoS Reserved> for RESPONSE

The <QSPEC Version> used by RMD-QOSM is the default version, i.e., "0", see [RFC5975]. The <QSPEC Type> value used by the RMD-QOSM is specified in [RFC5975] and is equal to "2". The <Traffic Handling Directives> contains the following fields:

<Traffic Handling Directives> = <PHR container> <PDR container>

The Per-Hop Reservation container (PHR container) and the Per-Domain Reservation container (PDR container) are specified in Sections 4.1.2 and 4.1.3, respectively. The <PHR container> contains the traffic handling directives for intra-domain communication and reservation. The <PDR container> contains additional traffic handling directives that are needed for edge-to-edge communication. The parameter IDs used by the <PHR container> and <PDR container> are assigned by IANA; see Section 6.

The RMD-QOSM <QoS Desired> and <QoS Reserved>, are specified in Section 4.1.1. The RMD-QOSM <QoS Desired> and <QoS Reserved> and 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 Desired> and <QoS Reserved>

The RESERVE message contains only the <QoS Desired> object [RFC5975]. The <QoS Reserved> object is carried by the RESPONSE message.

In RMD-QOSM, the <QoS Desired> and <QoS Reserved> objects contain the following parameters:

<QoS Desired> = <TMOD-1> <PHB Class> <Admission Priority>
<QoS Reserved> = <TMOD-1> <PHB Class> <Admission Priority>

The bit format of the <PHB Class> (see [RFC5975] and Figures 4 and 5) and <Admission Priority> complies with the bit format specified in [RFC5975].

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.

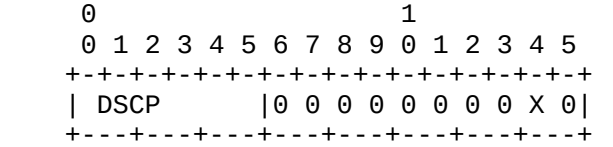

Figure 4: DSCP parameter

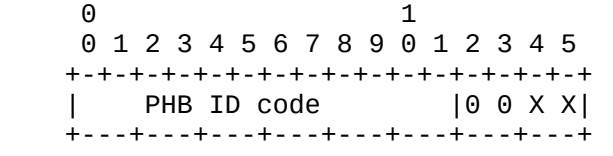

Figure 5: PHB ID Code parameter

4.1.2. PHR Container

This section describes the parameters used by the PHR container, which are used by the RMD-QOSM functionality available at the Interior nodes.

<PHR container> = <O> <K> <S> <M>, <Admitted Hops>, <B> <Hop_U> <Time Lag> <SCH> <Max Admitted Hops>

The bit format of the PHR container can be seen in Figure 6. Note that in Figure 6 <Hop_U> is represented as <U>. Furthermore, in Figure 6, <Max Admitted Hops> is represented as <Max Adm Hops>.

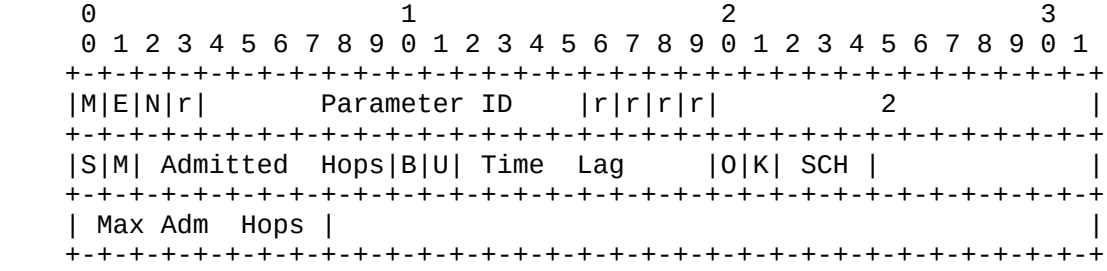

Figure 6: PHR container

Parameter ID: 12-bit field, indicating the PHR type: PHR_Resource_Request, PHR_Release_Request, PHR_Refresh_Update.

"PHR_Resource_Request" (Parameter ID = 17): 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_Release_Request" (Parameter ID = 18): explicitly release, by subtraction, the reserved resources for a particular flow from a traffic class reservation state.

"PHR_Refresh_Update" (Parameter ID = 19): 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.

<S> (Severe Congestion): 1 bit. In the 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, severe congestion occurs. Severe congested Interior nodes SHOULD notify Edge QNEs about the congestion by setting the <S> bit.

<O> (Overload): 1 bit. This field is used during the severe congestion handling scheme that is using the RMD-QOSM refresh procedure. This bit is set when an overload on a QNE Interior node is detected and when this field is carried by the "PHR_Refresh_Update" container. <O> SHOULD be set to"1" if the <S> bit is set. For more details, see Section 4.6.1.6.1.

<M>: 1 bit. In the 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 MUST be incremented by each Interior QNE, provided that the <Hop_U> bit is not set. 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, by setting the <Hop_U> bit to "1". Note that the <Admitted Hops> parameter in combination with the <Max Admitted Hops> and <K> parameters are used during the RMD partial release procedures (see Section 4.6.1.5.2).

<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 to "1", the <Admitted

Hops> SHOULD NOT be changed. Note that this flag, in combination with the <Admitted Hops> flag, are used to locate the last node that successfully processed a reservation request (see Section 4.6.1.2).

<B>: 1 bit. When set to "1", it indicates a bidirectional reservation.

<Time Lag>: It represents the ratio between the "T_Lag" parameter, which is the time difference between the departure time of the last sent "PHR_Refresh_Update" control information container and the departure time of the "PHR_Release_Request" control information container, and the length of the refresh period, "T_period", see Section 4.6.1.5.

<K>: 1 bit. When set to "1", it indicates that the resources/bandwidth carried by a tearing RESERVE MUST NOT be released, and the resources/bandwidth carried by a non-tearing RESERVE MUST NOT be reserved/refreshed. For more details, see Section 4.6.1.5.2.

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

<SCH>: 3 bits. The <SCH> value that is used to specify which of the 6 RMD-QOSM scenarios (see Section 3.2.3) MUST be used within the RMD domain. The operator of an RMD domain MUST preconfigure all the QNE Edge nodes within one domain such that the <SCH> field included in the "PHR container", will always use the same value, such that within one RMD domain only one of the below described RMD-QOSM schemes can be used at a time. All the QNE Interior nodes MUST interpret this field before processing any other PHR container payload fields. The currently defined <SCH> values are:

- 0: RMD-QOSM scheme MUST be "per-flow congestion notification based on probing";
- 1: RMD-QOSM scheme MUST be "per-flow RMD NSIS measurement-based admission control",
- 2: RMD-QOSM scheme MUST be "per-flow RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;
- 3 : RMD-QOSM scheme MUST be "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;

- o 4: RMD-QOSM scheme MUST be "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

- o 5: RMD-QOSM scheme MUST be "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;

- o 6 - 7: reserved.

The default value of the <SCH> field MUST be set to the value equal to 3.

4.1.3. PDR Container

This section describes the parameters of the PDR container, which are used by the RMD-QOSM functionality available at the Edge nodes.

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

<PDR container> = <O> <S> <M>
<Max Admitted Hops> <B> <SCH> [<PDR Bandwidth>]

In Figure 7, note that <Max Admitted Hops> is represented as <Max Adm Hops>.

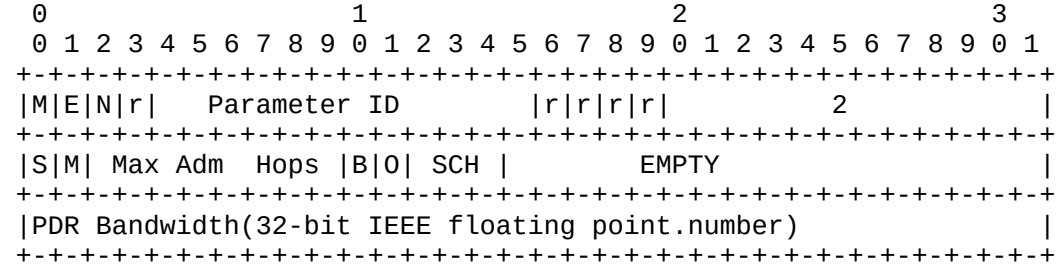

Figure 7: PDR container

Parameter ID: 12-bit field identifying the type of <PDR container> field.

"PDR_Reservation_Request" (Parameter ID = 20): 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 ID = 21): 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 ID = 22): 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 ID = 23): 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" traffic handling directive field have been received and that the request has been admitted or rejected.

"PDR_Refresh_Report" (Parameter ID = 24) generated and sent by the QNE(Egress) node in case needed, to the QNE(Ingress) node to report that a "PHR_Refresh_Update" traffic handling directive field has been received and has been processed.

"PDR_Release_Report" (Parameter ID = 25) 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" traffic handling directive field have been received and have been processed.

"PDR_Congestion_Report" (Parameter ID = 26): 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.

<O> (Overload): 1 bit. This field is used during the severe congestion handling scheme that is using the RMD-QOSM refresh procedure. This bit is set when an overload on a QNE Interior node is detected and when this field is carried by the "PDR_Congestion_Report" container. <O> SHOULD be set to "1" if the <S> bit is set. For more details, see Section 4.6.1.6.1.

<M> (PDR Marked): 1 bit. Carries the <M> value of the "PHR_Resource_Request" or "PHR_Refresh_Update" traffic handling directive field.

<B>: 1 bit. Indicates bidirectional reservation.

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

<PDR Bandwidth>: 32 bits. This field specifies the bandwidth that either applies when the <B> flag is set to "1" and when this parameter is carried by a RESPONSE message or when a severe congestion occurs and the QNE Edges maintain an aggregated intra-domain QoS-NSLP operational state and it is carried by a NOTIFY message. In the situation that the <B> flag is set to "1", this parameter specifies the requested bandwidth that has to be reserved by a node in the reverse direction and when the intra-domain signaling procedures require a bidirectional reservation procedure. In the severe congestion situation, this parameter specifies the bandwidth that has to be released.

<SCH>: 3 bits. The <SCH> value that is used to specify which of the 6 RMD scenarios (see Section 3.2.3) MUST be used within the RMD domain. The operator of an RMD domain MUST preconfigure all the QNE Edge nodes within one domain such that the <SCH> field included in the "PDR container", will always use the same value, such that within one RMD domain only one of the below described RMD-QOSM schemes can be used at a time. All the QNE Interior nodes MUST interpret this field before processing any other <PDR container> payload fields. The currently defined <SCH> values are:

- 0: RMD-QOSM scheme MUST be "per-flow congestion notification based on probing";
- 1: RMD-QOSM scheme MUST be "per-flow RMD NSIS measurement-based admission control";
- 2: RMD-QOSM scheme MUST be "per-flow RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;
- 3 : RMD-QOSM scheme MUST be "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;
- 4: RMD-QOSM scheme MUST be "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;
- 5: RMD-QOSM scheme MUST be "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;

o 6 - 7: reserved.

The default value of the <SCH> field MUST be set to the value equal to 3.

4.2. Message Format

The format of the messages used by the RMD-QOSM complies with the QoS-NSLP and QSPEC template specifications. The QSPEC used by RMD-QOSM is denoted in this document as RMD-QSPEC and is described in Section 4.1.

4.3. RMD Node State Management

The QoS-NSLP state creation and management is specified in [RFC5974]. This section describes the state creation and management functions of the Resource Management Function (RMF) in the RMD nodes.

4.3.1. Aggregated Operational and Reservation States at the QNE Edges

The QNE Edges maintain both the intra-domain QoS-NSLP operational and reservation states, while the QNE Interior nodes maintain only reservation states. The structure of the intra-domain QoS-NSLP operational state used by the QNE Edges is specified in [RFC5974].

In this case, the intra-domain sessions supported by the Edges are per-aggregate sessions that have a one-to-many relationship to the per-flow end-to-end states supported by the same Edge.

Note that the method of selecting the end-to-end sessions that form an aggregate is not specified in this document. An example of how this can be accomplished is by monitoring the GIST routing states used by the end-to-end sessions and grouping the ones that use the same <PHB Class>, QNE Ingress and QNE Egress addresses, and the value of the priority level. Note that this priority level should be deduced from the priority parameters carried by the initial QSPEC object.

The operational state of this aggregated intra-domain session MUST contain a list with BOUND-SESSION-IDs.

The structure of the list depends on whether a unidirectional reservation or a bidirectional reservation is supported.

When the operational state (at QNE Ingress and QNE Egress) supports unidirectional reservations, then this state MUST contain a list with BOUND-SESSION-IDs maintaining the <SESSION-ID> values of its bound end-to-end sessions. The Binding_Code associated with this BOUND-

SESSION-ID is set to code (Aggregated sessions). Thus, the operational state maintains a list of BOUND-SESSION-ID entries. Each entry is created when an end-to-end session joins the aggregated intra-domain session and is removed when an end-to-end session leaves the aggregate.

It is important to emphasize that, in this case, the operational state (at QNE Ingress and QNE Egress) that is maintained by each end-to-end session bound to the aggregated intra-domain session MUST contain in the BOUND-SESSION-ID, the <SESSION-ID> value of the bound tunneled intra-domain (aggregate) session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Aggregated sessions).

When the operational state (at QNE Ingress and QNE Egress) supports bidirectional reservations, the operational state MUST contain a list of BOUND-SESSION-ID sets. Each set contains two BOUND-SESSION-IDs. One of the BOUND-SESSION-IDs maintains the <SESSION-ID> value of one of bound end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Aggregated sessions). Another BOUND-SESSION-ID, within the same set entry, maintains the SESSION-ID of the bidirectional bound end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Bidirectional sessions).

Note that, in each set, a one-to-one relation exists between each BOUND-SESSION-ID with Binding_Code set to (Aggregate sessions) and each BOUND-SESSION-ID with Binding_Code set to (bidirectional sessions). Each set is created when an end-to-end session joins the aggregated operational state and is removed when an end-to-end session leaves the aggregated operational state.

It is important to emphasize that, in this case, the operational state (at QNE Ingress and QNE Egress) that is maintained by each end-to-end session bound to the aggregated intra-domain session it MUST contain two types of BOUND-SESSION-IDs. One is the BOUND-SESSION-ID that MUST contain the <SESSION-ID> value of the bound tunneled aggregated intra-domain session that is using the Binding_Code set to (Aggregated sessions). The other BOUND-SESSION-ID maintains the SESSION-ID of the bound bidirectional end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Bidirectional sessions).

When the QNE Edges use aggregated QoS-NSLP reservation states, then the <PHB Class> value and the size of the aggregated reservation, e.g., reserved bandwidth, have to be maintained. Note that this type of aggregation is an edge-to-edge aggregation and is similar to the aggregation type specified in [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/sessions it aggregates (e.g., see Section 1.4.4 of [RFC3175]).

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).

The QNE Interior nodes are reduced-state nodes, i.e., they do not store NTLP/GIST states, but they do store per PHB-aggregated QoS-NSLP reservation states. These reservation states are maintained and refreshed in the same way as described in Section 4.3.3.

4.3.2. Measurement-Based Method

The QNE Edges maintain per-flow intra-domain QoS-NSLP operational and reservation states that contain similar data structures as those described in Section 4.3.1. The main difference is associated with the different types of the used Message-Routing-Information (MRI) and the bound end-to-end sessions. The structure of the maintained BOUND-SESSION-IDs depends on whether a unidirectional reservation or a bidirectional reservation is supported.

When unidirectional reservations are supported, the operational state associated with this per-flow intra-domain session MUST contain in the BOUND-SESSION-ID the <SESSION-ID> value of its bound end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Tunneled and end-to-end sessions).

When bidirectional reservations are supported, the operational state (at QNE Ingress and QNE Egress) MUST contain two types of BOUND-SESSION-IDs. One is the BOUND-SESSION-ID that maintains the <SESSION-ID> value of the bound tunneled per-flow intra-domain session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Tunneled and end-to-end sessions).

The other BOUND-SESSION-ID maintains the SESSION-ID of the bound bidirectional end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Bidirectional sessions).

Furthermore, the QoS-NSLP reservation state maintains the <PHB Class> value, the value of the bandwidth requested by the end-to-end session bound to the intra-domain session, and the value of the priority level.

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 PHB traffic class, a predefined congestion notification threshold is set. The predefined congestion notification threshold is set according to an engineered bandwidth limitation based, e.g., on a 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, an 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. It is required that the re-marking happens to all packets that belong to the congested PHB traffic class so that the probe can't pass the congested router without being re-marked. In this way, it is ensured that the end-to-end NSIS message passed through the node that is congested. This feature is very useful when flow-based ECMP (Equal Cost Multiple Path) routing is used to detect only flows that are passing through the congested node.

* NSIS measurement-based admission control:

The measurement-based admission control is implemented in NSIS-aware stateless routers. Thus, the main difference between this type of the measurement-based admission control and the congestion notification-based admission control is the fact that the Interior nodes are NSIS-aware nodes. 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 intra-domain 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

The QNE Edges maintain intra-domain QoS-NSLP operational and reservation states that contain similar data structures as described in Section 4.3.1.

In this case, the intra-domain sessions supported by the Edges are per-flow sessions that have a one-to-one relationship to the per-flow end-to-end states supported by the same Edge.

The 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. This state is identified by the <PHB Class> value and it maintains the number of currently reserved resource units (or bandwidth). 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 in the Interior nodes, is based on the following equation:

$$last + p <= T,$$

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

The PHB group reservation states maintained in the Interior nodes are soft states, which are refreshed by sending periodic refresh intra-domain RESERVE messages, which are initiated by the Ingress QNEs. If a refresh message corresponding to a number of reserved resource units (i.e., bandwidth) 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 intra-domain RESERVE release/tear message, which is generated by the Ingress QNEs.

The use 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 the case of measurement- and (per-flow and aggregated) RMD reservation-based methods, the way in which 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 models 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 with the same PHB, are denoted in this document as PHB_low_priority, PHB_normal_priority, and PHB_high_priority, and are identified by the <PHB Class> value and the priority value, which is carried in the <Admission Priority> RMD-QSPEC parameter.

4.4. Transport of RMD-QOSM Messages

As mentioned in Section 1, the RMD-QOSM aims to support a number of additional requirements, e.g., Minimal impact on Interior node performance. Therefore, RMD-QOSM is designed to be very lightweight signaling with regard to the number of signaling message round trips and the amount of state established at involved signaling nodes with and without reduced state on QNEs. The actions allowed by a QNE Interior node are 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. Moreover, RMD signaling is targeted towards intra-domain signaling only. Therefore, RMD-QOSM relies on the security and reliability support that is provided by the bound end-to-end session, which is running between the boundaries of the RMD domain (i.e., the RMD-QOSM QNE Edges), and the security provided by the D-mode. This implies the use of the Datagram Mode.

Therefore, the intra-domain messages used by the RMD-QOSM are intended to operate in the NTLP/GIST Datagram mode (see [RFC5971]). The NSLP functionality available in all RMD-QOSM-aware QoS-NSLP nodes requires the intra-domain GIST, via the QoS-NSLP RMF API see [RFC5974], 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.

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

* not create any NTLP routing state by the Interior nodes. This can be satisfied by passing this requirement from the QoS-NSLP layer to the GIST layer via the API. However, between the QNE Egress and QNE Ingress routing states SHOULD be created that are associated with intra-domain sessions and that can be used for the communication of GIST Data messages sent by a QNE Egress directly to a QNE Ingress. This type of routing state associated with an intra-domain session can be generated and used in the following way:

* When the QNE Ingress has to send an initial intra-domain RESERVE message, the QoS-NSLP sends this message by including, in the GIST API SendMessage primitive, the Unreliable and No security attributes. In order to optimize this procedure, the RMD domain MUST be engineered in such a way that GIST will piggyback this NSLP message on a GIST Query message. Furthermore, GIST sets the C-flag (C=1), see [RFC5971] and uses the Q-mode. The GIST functionality in each QNE Interior node will receive the GIST Query message and by using the RecvMessage GIST API primitive it will pass the intra-domain RESERVE message to the QoS-NSLP functionality. At the same time, the GIST functionality uses the Routing-State-Check boolean to find out if the QoS-NSLP needs to create a routing state. The QoS-NSLP sets this boolean to inform GIST to not create a routing state and to forward the GIST Query further downstream with the

modified QoS-NSLP payload, which will include the modified intra-domain RESERVE message. The intra-domain RESERVE is sent in the same way up to the QNE Egress. The QNE Egress needs to create a routing state.

Therefore, at the same moment that the GIST functionality passes the intra-domain RESERVE message, via the GIST RecvMessage primitive, to the QoS-NSLP, the QoS-NSLP sets the Routing-State-Check boolean such that a routing state is created. The GIST creates the routing state using normal GIST procedures. After this phase, the QNE Ingress and QNE Egress have, for the particular session, routing states that can route traffic directly from QNE Ingress to QNE Egress and from QNE Egress to QNE Ingress. The routing state at the QNE Egress can be used by the QoS-NSLP and GIST to send an intra-domain RESPONSE or intra-domain NOTIFY directly to the QNE Ingress using GIST Data messages. Note that this routing state is refreshed using normal GIST procedures. Note that in the above description, it is considered that the QNE Ingress can piggyback the initial RESERVE (NSLP) message on the GIST Query message. If the piggybacking of this NSLP (initial RESERVE) message would not be possible on the GIST Query message, then the GIST Query message sent by the QNE Ingress node would not contain any NSLP data. This GIST Query message would only be processed by the QNE Egress to generate a routing state.

After the QNE Ingress is informed that the routing state at the QNE Egress is initiated, it would have to send the initial RESERVE message using similar procedures as for the situation that it would send an intra-domain RESERVE message that is not an initial RESERVE, see next bullet. This procedure is not efficient and therefore it is RECOMMENDED that the RMD domain MUST be engineered in such a way that the GIST protocol layer, which is processed on a QNE Ingress, will piggyback an initial RESERVE (NSLP) message on a GIST Query message that uses the Q-mode.

* When the QNE Ingress needs to send an intra-domain RESERVE message that is not an initial RESERVE, then the QoS-NSLP sends this message by including in the GIST API SendMessage primitive such attributes that the use of the Datagram Mode is implied, e.g., the Unreliable attribute. Furthermore, the Local policy attribute is set such that GIST sends the intra-domain RESERVE message in a Q-mode even if there is a routing state at the QNE Ingress. In this way, the GIST functionality uses its local policy to send the intra-domain RESERVE message by piggybacking it on a GIST Data message and sending it in Q-mode even if there is a routing state for this session. The intra-domain RESERVE message is piggybacked on the GIST Data message that is forwarded and processed by the QNE Interior nodes up to the QNE Egress.

The transport of the original (end-to-end) RESERVE message is accomplished in the following way:

At the QNE 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 NSLPID values (see [RFC5974]). This is accomplished by marking the end-to-end RESERVE message, i.e., modifying the QoS-NSLP default NSLPID value to another NSLPID predefined value.

The marking MUST be accomplished by the Ingress by modifying the QoS_NSLP default NSLPID value to a NSLPID predefined value. In this way, the Egress MUST stop this marking process by reassigning the QoS-NSLP default NSLPID value to the original (end-to-end) RESERVE message. Note that the assignment of these NSLPID values is a QoS-NSLP issue, which SHOULD be accomplished via IANA [RFC5974].

4.5. Edge Discovery and Message Addressing

Mainly, the Egress node discovery can be performed by using either the GIST discovery mechanism [RFC5971], manual configuration, or any other discovery technique. The addressing of signaling messages depends on which GIST transport mode is used. The RMD-QOSM/QoS-NSLP signaling messages that are processed only by the Edge nodes use the peer-peer addressing of the GIST Connection (C) mode.

RMD-QOSM/QoS-NSLP signaling messages that are processed by all nodes of the Diffserv domain, i.e., Edges and Interior nodes, use the end-to-end addressing of the GIST Datagram (D) mode. Note that the RMD-QOSM cannot directly specify that the GIST Connection or the GIST Datagram mode SHOULD be used. This can only be specified by using, via the QoS-NSLP-RMF API, the GIST API Transfer-Attributes, such as Reliable or Unreliable, high or low level of security, and by the use of local policies. RMD QoS signaling messages that are addressed to the data path end nodes are intercepted by the Egress nodes. In particular, at the ingress and for downstream intra-domain messages, the RMD-QOSM instructs the GIST functionality, via the GIST API to do the following:

* use unreliable and low level security Transfer-Attributes,

* do not create a GIST routing state, and

* use the D-mode MRI.

The intra-domain RESERVE messages can then be transported by using the Query D-mode; see Section 4.4.

At the QNE Egress, and for upstream intra-domain messages, the RMD-QOSM instructs the GIST functionality, via the GIST API, to use among others:

* unreliable and low level security Transfer-Attributes

* the routing state associated with the intra-domain session to send an upstream intra-domain message directly to the QNE Ingress; see Section 4.4.

4.6. Operation and Sequence of Events

4.6.1. Basic Unidirectional Operation

This section describes the basic unidirectional operation and sequence of events/triggers 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 (Section 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 Edges can also operate as either QNI or as QNR and at the same time they can operate as either sender or receiver of the data path.

It is important to emphasize that the content of this section is used for the specification of the following RMD-QOSM/QoS-NSLP signaling schemes, when basic unidirectional operation is assumed:

* "per-flow congestion notification based on probing";

* "per-flow RMD NSIS measurement-based admission control";

* "per-flow RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

* "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure.

For more details, please see Section 3.2.3.

In particular, the functionality described in Sections 4.6.1.1, 4.6.1.2, 4.6.1.3, 4.6.1.5, 4.6.1.4, and 4.6.1.6 applies to the RMD reservation-based and to the NSIS measurement-based admission control methods. The described functionality in Section 4.6.1.7 applies to the admission control procedure that uses the congestion notification based on probing. The QNE Edge nodes maintain either per-flow QoS-NSLP operational and reservation states or aggregated QoS-NSLP operational and reservation states.

When the QNE Edges maintain aggregated QoS-NSLP operational and reservation states, the RMD-QOSM functionality MAY accomplish an RMD modification procedure (see Section 4.6.1.4), instead of the reservation initiation procedure that is described in this subsection. Note that it is RECOMMENDED that the QNE implementations of RMD-QOSM process the QoS-NSLP signaling messages with a higher priority than data packets. This can be accomplished as described in Section 3.3.4 of [RFC5974] and it can be requested via the QoS-NSLP-RMF API described in [RFC5974]. The signaling scenarios described in this section are accomplished using the QoS-NSLP processing rules defined in [RFC5974], in combination with the RMF triggers sent via the QoS-NSLP-RMF API described in [RFC5974].

According to Section 3.2.3, it is specified that only the "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" scheme MUST be implemented within one RMD domain. However, all RMD QNEs supporting this specification MUST support the combination the "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" scheme. If the RMD QNEs support more RMD-QOSM schemes, then the operator of that RMD domain MUST preconfigure all the QNE Edge nodes within one domain such that the <SCH> field included in the "PHR container" (Section

4.1.2) and the "PDR Container" (Section 4.1.3) will always use the same value, such that within one RMD domain only one of the below described RMD-QOSM schemes is used at a time.

All QNE nodes located within the RMD domain MUST read and interpret the <SCH> field included in the "PHR container" before processing all the other "PHR container" payload fields. Moreover, all QNE Edge nodes located at the boarder of the RMD domain, MUST read and interpret the <SCH> field included in the "PDR container" before processing all the other <PDR container> payload fields.

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 [RFC5971].

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. Subsequently, the combination of <TMOD-1>, <PHB Class>, and <Admission Priority> is derived from the <QoS Desired> object of the initial QSPEC.

The QNE Ingress MUST maintain information about the smallest MTU that is supported on the links within the RMD domain.

The <Maximum Packet Size-1 (MPS)> value included in the end-to-end QoS Model <TMOD-1> parameter is compared with the smallest MTU value that is supported by the links within the RMD domain. If the "Maximum Packet Size-1 (MPS)" is larger than this smallest MTU value within the RMD domain, then the end-to-end reservation request is rejected (see Section 4.6.1.1.2). Otherwise, the admission process continues.

The <TMOD-1> parameter contained in the original initiator QSPEC is mapped into the equivalent RMD-Qspec <TMOD-1> parameter representing only the peak bandwidth in the local RMD-QSPEC. This can be accomplished by setting the RMD-QSPEC <TMOD-1> fields as follows: token rate (r) = peak traffic rate (p), the bucket depth (b) = large, and the minimum policed unit (m) = large.

Note that the bucket size, (b), is measured in bytes. Values of this parameter may range from 1 byte to 250 gigabytes; see [RFC2215]. Thus, the maximum value that (b) could be is in the order of 250

gigabytes. The minimum policed unit, [m], is an integer measured in bytes and must be less than or equal to the Maximum Packet Size (MPS). Thus, the maximum value that (m) can be is (MPS). [Part94] and [TaCh99] describe a method of calculating the values of some Token Bucket parameters, e.g., calculation of large values of (m) and (b), when the token rate (r), peak rate (p), and MPS are known.

The <Peak Data Rate-1 (p)> value of the end-to-end QoS Model <TMOD-1> parameter is copied into the <Peak Data Rate-1 (p)> value of the <Peak Data Rate-1 (p)> value of the local RMD-Qspec <TMOD-1>.

The MPS value of the end-to-end QoS Model <TMOD-1> parameter is copied into the MPS value of the local RMD-Qspec <TMOD-1>.

If the initial QSPEC does not contain the <PHB Class> parameter, then the selection of the <PHB Class> that is carried by the intra-domain RMD-QSPEC is defined by a local policy similar to the procedures discussed in [RFC2998] and [RFC3175].

For example, in the situation that the initial QSPEC is used by the IntServ Controlled Load QOSM, then the Expedited Forwarding (EF) PHB is appropriate to set the <PHB Class> parameter carried by the intra-domain RMD-QSPEC (see [RFC3175]).

If the initial QSPEC does not carry the <Admission Priority> parameter, then the <Admission Priority> parameter in the RMD-QSPEC will not be populated. If the initial QSPEC does not carry the <Admission Priority> parameter, but it carries other priority parameters, then it is considered that Edges, as being stateful nodes, are able to control the priority of the sessions that are entering or leaving the RMD domain in accordance with the priority parameters.

Note that the RMF reservation states (see Section 4.3) in the QNE Edges store the value of the <Admission Priority> parameter that is used within the RMD domain in case of preemption and severe congestion situations (see Section 4.6.1.6).

If the RMD domain supports preemption during the admission control process, then the QNE Ingress node can support the building blocks specified in [RFC5974] and during the admission control process use the example preemption handling algorithm described in Appendix A.7.

Note that in the above described case, the QNE Egress uses, if available, the tunneled initial priority parameters, which can be interpreted by the QNE Egress.

If the initial QSPEC carries the <Excess Treatment> parameter, then the QNE Ingress and QNE Egress nodes MUST control the excess traffic that is entering or leaving the RMD domain in accordance with the <Excess Treatment> parameter. Note that the RMD-QSPEC does not carry the <Excess Treatment> parameter.

If the requested <TMOD-1> parameter carried by the initial QSPEC, cannot be satisfied, then an end-to-end RESPONSE message has to be generated. However, in order to decide whether the end-to-end reservation request was locally (at the QNE Ingress) satisfied, a local (at the QNE_Ingress) RMD-QOSM admission control procedure also has to be performed. In other words, the RMD-QOSM functionality has to verify whether the value included in the <Peak Data Rate-1 (p)> field of RMD-QOSM <TMOD-1> can be reserved and stored in the RMD-QOSM reservation states (see Sections 4.6.1.1.2 and 4.3).

An initial QSPEC object MUST be included in the end-to-end 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 local RMD-QSPEC <TMOD-1> 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: Transient Failure Error code value: Reservation failure

Furthermore, all of the other RESPONSE parameters are set according to the end-to-end QoS Model or according to [RFC5974] and [RFC5975].

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 message. Note however, that when the aggregated QoS-NSLP operational and reservation states are used by the QNE Ingress, then the generation of the intra-domain RESERVE message depends on the availability of the aggregated QoS-NSLP operational state. If this aggregated QoS-NSLP operational state is available, then the RMD modification of aggregated reservations described in Section 4.6.1.4 is used.

It is important to note that when the "per-flow RMD reservation-based" scenario is used within the RMD domain, the retransmission within the RMD domain SHOULD be disallowed. The reason for this is related to the fact that the QNI Interior nodes are not able to differentiate between a retransmitted RESERVE message associated with a certain session and an initial RESERVE message belonging to another session. However, the QNE Ingress have to report a failure situation

upstream. When the QNE Ingress transmits the (intra-domain or end-to-end) RESERVE with the <RII> object set, it waits for a RESPONSE from the QNE Egress for a QOSNSLP_REQUEST_RETRY period.

If the QNE Ingress transmitted an intra-domain or end-to-end RESERVE message with the <RII> object set and it fails to receive the associated intra-domain or end-to-end RESPONSE, respectively, after the QOSNSLP_REQUEST_RETRY period expires, it considers that the reservation failed. In this case, the QNE Ingress SHOULD generate an end-to-end RESPONSE message that will include, among others, an <INFO-SPEC> object. The error code used by this <INFO-SPEC> object is:

Error severity class: Transient Failure
Error code value: Reservation failure

Furthermore, all of the other RESPONSE parameters are set according to the end-to-end QoS Model or according to [RFC5974] and [RFC5975].

Note however, that if the retransmission within the RMD domain is not disallowed, then the procedure described in Appendix A.8 SHOULD be used on QNE Interior nodes; see also [Chan07]. In this case, the stateful QNE Ingress uses the retransmission procedure described in [RFC5974].

If a rerouting takes place, then the stateful QNE Ingress is following the procedures specified in [RFC5974].

At this point, the intra-domain and end-to-end operational states MUST be initiated or modified according to the REQUIRED binding procedures. The way of how the BOUND-SESSION-IDs are initiated and maintained in the intra-domain and end-to-end QoS-NSLP operational states is described in Sections 4.3.1 and 4.3.2.

These two messages 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 code "Tunnel and end-to-end sessions" (see Section 4.3.2).

In addition to this, the intra-domain and end-to-end RESERVE messages are bound using the Message binding procedure described in [RFC5974].

In particular the <MSG-ID> object is included in the intra-domain RESERVE message and its bound <BOUND-MSG-ID> object is carried by the end-to-end RESERVE message. Furthermore, the <Message_Binding_Type> flag is SET (value is 1), such that the message dependency is bidirectional.

If the QoS-NSLP Edges maintain aggregated intra-domain QoS-NSLP operational states, then the value of Binding_Code MUST be set to code "Aggregated sessions".

Furthermore, in this case, the retransmission within the RMD domain is allowed and the procedures described in Appendix A.8 SHOULD be used on QNE Interior nodes. This is necessary due to the fact that when retransmissions are disallowed, then the associated with (micro) flows belonging to the aggregate will loose their reservations. Note that, in this case, the stateful QNE Ingress uses the retransmission procedure described in [RFC5974].

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 [RFC5971]). 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 supported, 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 is sent using the NTLP datagram mode (see Sections 4.4 and 4.5). Note that the GIST Datagram mode can be selected using the unreliable GIST API Transfer-Attributes. In addition, the intra-domain RESERVE (RMD-QSPEC) message MUST include a PHR container (PHR_Resource_Request) and the RMD QOSM <QoS Desired> object.

The end-to-end RESERVE message includes the initial QSPEC and it is sent towards the Egress QNE.

Note that after completing the initial discovery phase, the GIST Connection mode can be used between the QNE Ingress and QNE Egress. Note that the GIST Connection mode can be selected using the reliable GIST API Transfer-Attributes.

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 NSLPID value to another NSLPID predefined 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 [RFC5971]) are configured to handle only certain NSLP-IDs (see [RFC5974]).

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. This can be accomplished only if the GIST implementation is configured to perform that, e.g., via a local policy. However, the selection of the discovery procedure cannot be selected by the RMD-QOSM.

The (initial) intra-domain RESERVE message MUST be sent by the QNE Ingress and it MUST contain the following values (see the QoS-NSLP-RMF API described in [RFC5974]):

- the <RSN> object, whose value is generated and processed as described in [RFC5974];
- 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 RMD boundary nodes and the QNE Interior nodes MUST be configured as Interior (intermediary) nodes;
- the <RII> MUST be included in this message, see [RFC5974];
- the <REPLACE> flag MUST be set to FALSE = 0;

* The value of the <Message ID> value carried by the <MSG-ID> object is set according to [RFC5974]. The value of the <Message_Binding_Type> is set to "1".

* the value of the <REFRESH-PERIOD> object MUST be calculated and set by the QNE Ingress node as described in Section 4.6.1.3;

* the value of the <PACKET-CLASSIFIER> object is associated with the path-coupled routing Message Routing Message (MRM), since RMD-QOSM is used with the path-coupled MRM. The flag that has to be set is the <T> flag (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 used in

the MRI can be derived by the value of <PHB Class> parameter, which MUST be carried by the intra-domain RESERVE message. Note that the QNE Ingress being a QNI for the intra-domain session it can pass this value to GIST, via the GIST API.

* the PHR resource units MUST be included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter of the <QoS Desired> object.

  When the QNE Edges use per-flow intra-domain QoS-NSLP states, then the <Peak Data Rate-1 (p)> value included in the initial QSPEC <TMOD-1> parameter is copied into the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter.

  When the QNE Edges use aggregated intra-domain QoS-NSLP operational states, then the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter can be obtained by using the bandwidth aggregation method described in Section 4.3.1;

* the value of the <PHB Class> parameter can be defined by using the method of copying the <PHB Class> parameter carried by the initial QSPEC into the <PHB Class> carried by the RMD-QSPEC, which is described above in this subsection.

* the value of the <Parameter ID> field of the PHR container MUST be set to "17", (i.e., PHR_Resource_Request).

* the value of the <Admitted Hops> parameter in the PHR container MUST be set to "1". Note that during a successful reservation, each time an RMD-QOSM-aware node processes the RMD-QSPEC, the <Admitted Hops> parameter is increased by one.

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

* the value of the <Max Admitted Hops> is set to "0".

* If the initial QSPEC carried an <Admission Priority> parameter, then this parameter SHOULD be copied into the RMD-QSPEC 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 of 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, and PHB_high_priority) MAY be associated with the same PHB (see Section 4.3.3).

* In a single RMD domain case, the PDR container MAY not be included in the message.

Note that the intra-domain RESERVE message does not carry the <BOUND-SESSION-ID> object. The reason for this is that the end-to-end RESERVE carries, in the <BOUND-SESSION-ID> object, the <SESSION-ID> value of the intra-domain session.

When an end-to-end RESPONSE message is received by the QNE Ingress node, which was sent by a QNE Egress node (see Section 4.6.1.1.3), then it is processed according to [RFC5974] and end-to-end QoS Model rules.

When an intra-domain RESPONSE message is received by the QNE Ingress node, which was sent by a QNE Egress (see Section 4.6.1.1.3), it uses the QoS-NSLP procedures to match it to the earlier sent intra-domain RESERVE message. After this phase, the RMD-QSPEC has to be identified and processed.

The RMD QOSM reservation has been successful if the <M> bit carried by the "PDR Container" is equal to "0" (i.e., not set).

Furthermore, the <INFO-SPEC> object is processed as defined in the QoS-NSLP specification. In the case of successful reservation, the <INFO-SPEC> object MUST have the following values:

* Error severity class: Success
* Error code value: Reservation successful

If the end-to-end RESPONSE message has to be forwarded to a node outside the RMD-QOSM-aware domain, then the 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. If an end-to-end QUERY is received by the QNE Ingress, then the same bypassing procedure has to be used as the one applied for an end-to-end RESERVE message. In particular, it is forwarded using the GIST forwarding procedure to bypass the Interior stateless or reduced-state QNE nodes.

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 (see QoS-NSLP-RMF API described in [RFC5974]):

* the values of the <RSN>, <RII>, <PACKET-CLASSIFIER>, <REFRESH-PERIOD>, 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 flag that has to be set in this case is the <T> flag (traffic class). The value of the <DSCP> value MUST be obtained via the MRI parameters that the QoS-NSLP receives from GIST. A QNE Interior MUST be able to associate the value carried by the RMD-QSPEC <PHB Class> parameter and the <DSCP> value obtained via GIST. This is REQUIRED, because there are situations in which the <PHB Class> parameter is not carrying a <DSCP> value but a PHB ID code, see Section 4.1.1.

* the flag <REPLACE> MUST be set to FALSE = 0;

* when the RMD reservation-based methods, described in Section 4.3.1 and 4.3.3, are used, the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter is used by the QNE Interior node for admission control. Furthermore, if the <Admission Priority> parameter is carried by the RMD-QOSM <QoS Desired> object, then this parameter is processed as described in the following bullets.

* in the case of the RMD reservation-based procedure, and if these resources are admitted (see Sections 4.3.1 and 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 on allocating bandwidth can be used, which are beyond the scope of this document.

* If the RMD domain supports preemption during the admission control process, then the QNE Interior node can support the building blocks specified in the [RFC5974] and during the admission control process use the preemption handling algorithm specified in Appendix A.7.

* in the case of the RMD measurement-based method (see Section 4.3.2), and if the requested into the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter is admitted, using a measurement-based admission control (MBAC) algorithm, then the number of this resource will be used to update the MBAC algorithm according to the operation described in Section 4.3.2.

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 NSLPID 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 MUST be removed after processing. Note that the received end-to-end RESERVE was tunneled within the RMD domain. Therefore, the tunneled initial QSPEC carried by the end-to-end RESERVE message has to be processed/set according to the [RFC5975] specification.

If a rerouting takes place, then the stateful QNE Egress is following the procedures specified in [RFC5974].

At this point, the intra-domain and end-to-end operational states MUST be initiated or modified according to the REQUIRED binding procedures.

The way in which the BOUND-SESSION-IDs are initiated and maintained in the intra-domain and end-to-end QoS-NSLP operational states is described in Sections 4.3.1 and 4.3.2.

If 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 to have failed.

Furthermore, if the initial QSPEC object used an object combination of type 1 or 2 where the <QoS Available> is populated, and the intra-domain RESERVE(RMD-QSPEC) was not successful at all nodes in the RMD domain MUST be considered that the <QoS Available> is not satisfied and that the inter-domain (end-to-end) reservation is considered to have failed.

Furthermore, note that when the QNE Egress uses per-flow intra-domain QoS-NSLP operational states (see Sections 4.3.2 and 4.3.3), the QNE Egress SHOULD support the message binding procedure described in [RFC5974], which can be used to synchronize the arrival of the end-

to-end RESERVE and the intra-domain RESERVE (RMD-QSPEC) messages, see Section 5.7, and QoS-NSLP-RMF API described in [RFC5974]. Note that the intra-domain RESERVE message carries the <MSG-ID> object and its bound end-to-end RESERVE message carries the <BOUND-MSG-ID> object. Both these objects carry the <Message_Binding_Type> flag set to the value of "1". If these two messages do not arrive during the time defined by the MsgIDWait timer, then the reservation is considered to have failed. Note that the timer has to be preconfigured and it has to have the same value in the RMD domain. In this case, an end-to-end RESPONSE message, see QoS-NSLP-RMF API described in [RFC5974], is sent towards the QNE Ingress with the following <INFO-SPEC> values:

Error class: Transient Failure
Error code: 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 according to the specification given in [RFC5974]. The SESSION-ID included in the BOUND-SESSION-ID parameter stored in the intra-domain QoS-NSLP operational state object is the SESSION-ID of the session associated with the end-to-end RESERVE message(s). Note that if the QNE Edge nodes maintain per-flow intra-domain QoS-NSLP operational states, then the value of Binding_Code = (Tunnel and end-to-end sessions) is used. If the QNE Edge nodes maintain per-aggregated QoS-NSLP intra-domain reservation states, then the value of Binding_Code = (Aggregated sessions), see Sections 4.3.1 and 4.3.2.

If the RMD domain supports preemption during the admission control process, then the QNE Egress node can support the building blocks specified in the [RFC5974] and during the admission control process use the example preemption handling algorithm described in Appendix A.7.

The end-to-end RESERVE message is generated/forwarded further upstream according to the [RFC5974] and [RFC5975] specifications. Furthermore, the <B> (BREAK) QoS-NSLP flag in the end-to-end RESERVE message MUST NOT be set, see the QoS-NSLP-RMF API described in QoS-NSLP.

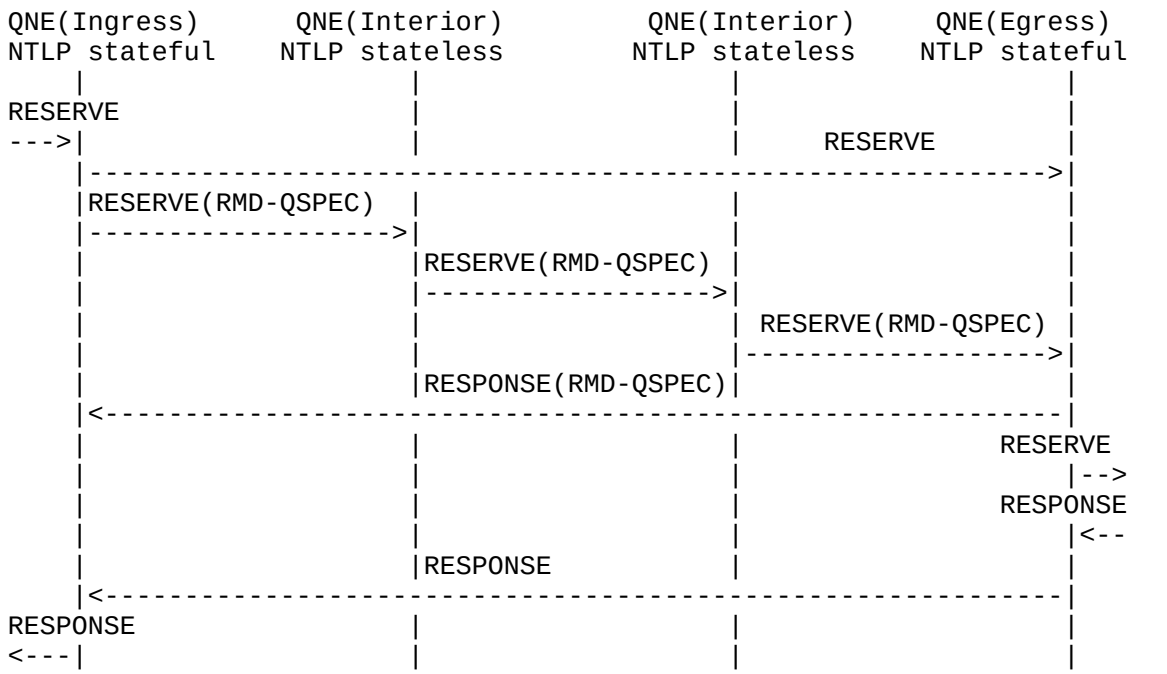

Figure 8: Basic operation of successful reservation procedure used by the RMD-QOSM

The QNE Egress MUST generate an intra-domain RESPONSE (RMD-Qspec) message. The intra-domain RESPONSE (RMD-QSPEC) message MUST be sent to the QNE Ingress node, i.e., the previous stateful hop by using the procedures described in Sections 4.4 and 4.5.

The values of the RMD-QSPEC that are carried by the intra-domain RESPONSE message MUST be used and/or set in the following way (see the QoS-NSLP-RMF API described in [RFC5974]):

* the <RII> object carried by the intra-domain RESERVE message, see Section 4.6.1.1.1, has to be copied and carried by the intra-domain RESPONSE message.

* the value of the <Parameter ID> field of the PDR container MUST be set to "23" (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. This is REQUIRED since the value of the <M> parameter is used to indicate the status if the RMD reservation request to the Ingress Edge.

If the binding between the intra-domain session and the end-to-end session uses a Binding_Code that is (Aggregated sessions), and there is no aggregated QoS-NSLP operational state associated with the intra-domain session available, then the RMD modification of aggregated reservation procedure described in Section 4.6.1.4 can be used.

If the QNE Egress receives an end-to-end RESPONSE message, it is processed and forwarded towards the QNE Ingress. In particular, the non-default values of the objects contained in the end-to-end RESPONSE message MUST be used and/or set by the QNE Egress as follows (see the QoS-NSLP-RMF API described in [RFC5974]):

* the values of the <RII>, <RSN>, <INFO-SPEC>, [<QSPEC>] objects are set according to [RFC5974] and/or [RFC5975]. The <INFO-SPEC> object SHOULD be set by the QoS-NSLP functionality. In the case of successful reservation, the <INFO-SPEC> object SHOULD have the following values:

  Error severity class: Success Error code value: Reservation successful

* furthermore, an initial QSPEC object MUST be included in the end-to-end RESPONSE message. The parameters included in the QSPEC <QoS Reserved> object are copied from the original <QoS Desired> values.

The end-to-end RESPONSE message is delivered as normal, i.e., is addressed and sent to its upstream QoS-NSLP neighbor, i.e., the QNE Ingress node.

Note that if a QNE Egress receives an end-to-end QUERY that was bypassed through the RMD domain, it MUST stop the marking process that was used to bypass the QNE Interior nodes. This can be done by reassigning the QoS-NSLP default NSLPID value to the end-to-end QUERY message; see Section 4.4.

4.6.1.2. Unsuccessful Reservation

This subsection 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 a similar 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, e.g., 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 methods (see Sections 4.3.1 and 4.3.2) and to the NSIS measurement-based admission control method (see Section 4.3.2).

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 an 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 (1) the "Maximum Packet Size-1 (MPS)" included in the end-to-end QoS Model <TMOD-1> is larger than this smallest MTU value within the RMD domain or (2) there are no resources available, the QNE Ingress MUST reject this end-to-end RESERVE message and send an end-to-end RESPONSE message back to the sender, as described in the QoS-NSLP specification, see [RFC5974] and [RFC5975].

When an end-to-end RESPONSE message is received by an Ingress node (see Section 4.6.1.2.3), the values of the <RII>, <RSN>, <INFO-SPEC>, and [<QSPEC>] objects are processed according to the QoS-NSLP procedures.

If the end-to-end RESPONSE message has to be forwarded upstream to a node outside the RMD-QOSM-aware domain, then the 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.

When an intra-domain RESPONSE message is received by the QNE Ingress node, which was sent by a QNE Egress (see Section 4.6.1.2.3), it uses the QoS-NSLP procedures to match it to the intra-domain RESERVE message that was previously sent. After this phase, the RMD-QSPEC has to be identified and processed. Note that, in this case, the RMD Resource Management Function (RMF) is notified that the reservation has been unsuccessful, by reading the <M> parameter of the PDR container. 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 an RMD modification procedure (see Section 4.6.1.4).

4.6.1.2.2. Operation in the Interior Nodes

In the 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. Moreover, the value of the <Max Admitted Hops> counter MUST be set equal to the <Admitted Hops> value.

Furthermore, the <E> flag associated with the QSPEC <QoS Desired> object and the <E> flag associated with the local RMD-QSPEC <TMOD-1> parameter SHOULD be set. In the 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 local RMD-QSPEC <TMOD-1> parameter SHOULD be set. Note that the <M> flag seems to be set in a similar way to the <E> flag used by the local RMD-QSPEC <TMOD-1> parameter. However, the ways in which the two flags are processed by a QNE are different.

In general, if a QNE Interior node receives an RMD-QSPEC <TMOD-1> 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-QOSM <QoS Desired> object) MUST NOT be processed. Furthermore, when the <K> parameter that is included in the "PHR Container" and carried by a RESERVE message is set to "1", then this "PHR Container" and the RMD-QOSM <QoS Desired> object) MUST NOT be processed.

4.6.1.2.3. Operation in the Egress Nodes

In the RMD reservation-based (Section 4.3.3) and RMD NSIS measurement-based scenarios (Section 4.3.2), 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 end-to-end session MUST be bound.

Moreover, if the initial QSPEC object (used by the end-to-end QoS Model) used an object combination of type 1 or 2 where the <QoS Available> is populated, and the intra-domain RESERVE(RMD-QSPEC) was not successful at all nodes in the RMD domain, i.e., the intra-domain RESERVE(RMD-QSPEC) message is marked, it MUST be considered that the <QoS Available> is not satisfied and that the inter-domain (end-to-end) reservation is considered as to have failed.

When the QNE Egress uses per-flow intra-domain QoS-NSLP operational states (see Sections 4.3.2 and 4.3.3), then the QNE Egress node MUST generate an end-to-end RESPONSE message that has to be sent to its previous stateful QoS-NSLP hop (see the QoS-NSLP-RMF API described in [RFC5974]).

* the values of the <RII>, <RSN> and <INFO-SPEC> objects are set by the standard QoS-NSLP protocol functions. In the case of an unsuccessful reservation, the <INFO-SPEC> object SHOULD have the following values:

  Error severity class: Transient Failure
  Error code value: Reservation failure

The QSPEC that was carried by the end-to-end RESERVE message that belongs to the same session as this end-to-end RESPONSE message is included in this message.

In particular, the parameters included in the QSPEC <QoS Reserved> object of the end-to-end RESPONSE message are copied from the initial <QoS Desired> values included in its associated end-to-end RESERVE message. The <E> flag associated with the QSPEC <QoS Reserved> object and the <E> flag associated with the <TMOD-1> parameter included in the end-to-end RESPONSE are set.

In addition to the above, similar to the successful operation, see Section 4.6.1.1.3, the QNE Egress MUST generate an intra-domain RESPONSE message that has to be sent to its previous stateful QoS-NSLP hop.

The values of the <RII>, <RSN> and <INFO-SPEC> objects are set by the standard QoS-NSLP protocol functions. In the case of an unsuccessful reservation, the <INFO-SPEC> object SHOULD have the following values (see the QoS-NSLP-RMF API described in [RFC5974]):

Error severity class: Transient Failure
Error code value: Reservation failure

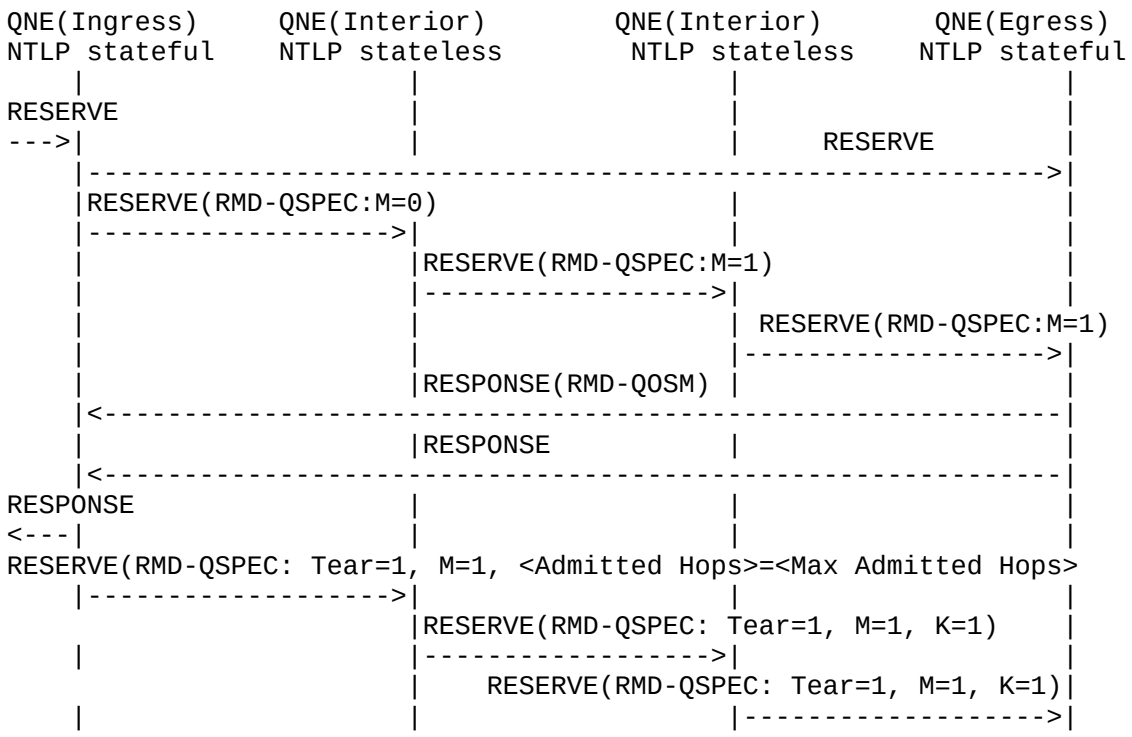

Figure 9: Basic operation during unsuccessful reservation initiation used by the RMD-QOSM

The values of the RMD-QSPEC MUST be used and/or set in the following way (see the QoS-NSLP-RMF API described in [RFC5974]):

* the value of the <PDR Control Type> of the PDR container MUST be set to "23" (PDR_Reservation_Report);

* the value of the <Max Admitted Hops> parameter of the PHR container included in the received <M> marked intra-domain RESERVE (RMD-QSPEC) 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 "1".

4.6.1.3. RMD Refresh Reservation

In the case of the RMD measurement-based method, see Section 4.3.2, QoS-NSLP reservation states in the RMD domain are not typically maintained, therefore, this method typically does not use an intra-domain refresh procedure.

However, there are measurement-based optimization schemes, see [GrTs03], that MAY use the refresh procedures described in Sections 4.6.1.3.1 and 4.6.1.3.3. However, this measurement-based optimization scheme can only be applied in the RMD domain if the QNE Edges are configured to perform intra-domain refresh procedures and if all the QNE Interior nodes are configured to perform the measurement-based optimization schemes.

In the description given in this subsection, it is assumed that the RMD measurement-based scheme does not use the refresh procedures.

When the QNE Edges maintain aggregated or per-flow QoS-NSLP operational and reservation states (see Sections 4.3.1 and 4.3.3), then the refresh procedures are very similar. If the RESERVE messages arrive within the soft state timeout 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 [RFC5974]).

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 defined by the refresh period/timer. 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 an RMD-QOSM <QoS Desired> and a PHR container (i.e., PHR_Refresh_Update).

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

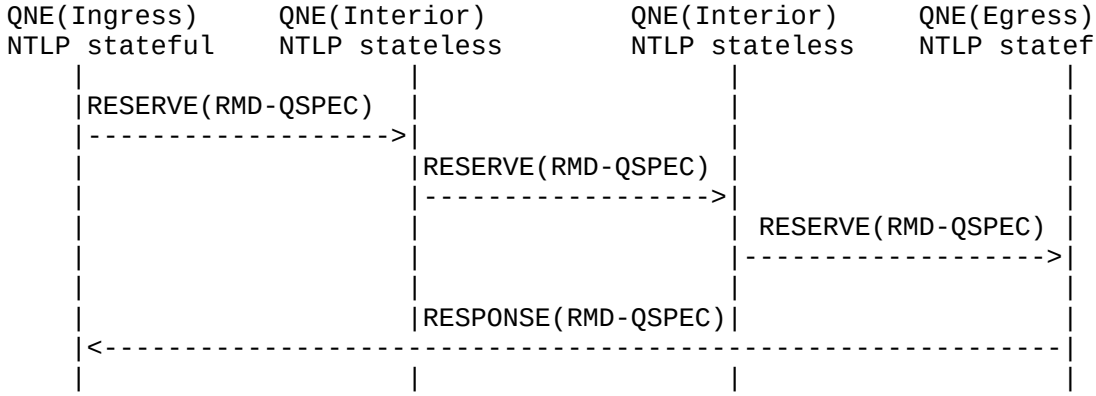

Figure 10: Basic operation of RMD-specific refresh procedure

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 PHR resource units MUST be included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter. The <Peak Data Rate-1 (p)> field value of the local RMD-QSPEC <TMOD-1> 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); see Section 4.3.1. If no QoS-NSLP aggregation is accomplished by the QNE Ingress node, the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter SHOULD be equal to the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of its associated new (initial) intra-domain RESERVE (RMD-QSPEC) message; see Section 4.3.3.

* the value of the Container field of the <PHR Container> MUST be set to "19", i.e., "PHR_Refresh_Update".

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>, and [RFC5975] objects are processed by the standard QoS-NSLP protocol functions (see Section 4.6.1.1);

* the "PDR Container" has to be processed 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

sessions associated with this RMD-specific refresh session MUST be informed about the success or failure of the refresh procedure. (When aggregated QoS-NSLP operational and reservation states are used (see Section 4.3.1), there will be more than one session.) In the 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 the 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> 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 <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired> is used by the QNE Interior node for refreshing the RMD traffic class state. These resources (included in the <Peak Data Rate-1 (p)> value of local RMD-QSPEC <TMOD-1>), if reserved, are added to the currently reserved resources per PHB and therefore they will become a part of the per-traffic class (PHB) reservation state (see Sections 4.3.1 and 4.3.3). If the refresh procedure cannot be fulfilled then the <M> and <S> fields carried by the PHR container MUST be set to "1".

* furthermore, the <E> flag associated with <QoS Desired> object and the <E> flag associated with the local RMD-QSPEC <TMOD-1> parameter SHOULD be set.

Any PHR container of type "PHR_Refresh_Update", and its associated local RMD-QSPEC <TMOD-1>, whether or not it is marked and independent of the <E> flag value of the local RMD-QSPEC <TMOD-1> 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 (refresh) intra-domain RESPONSE (RMD-QSPEC) message MUST be sent to the QNE Ingress node, i.e., the previous stateful hop. The (refresh) intra-domain RESPONSE (RMD-QSPEC) message MUST be explicitly routed to the QNE Ingress node, i.e., the previous stateful hop, using the procedures described in Section 4.5.

* the values of the <RII>, <RSN>, and <INFO-SPEC> objects are set by the standard QoS-NSLP protocol functions, see [RFC5974].

* the value of the <PDR Control Type> parameter of the PDR container MUST be set "24" (i.e., PDR_Refresh_Report). In case of successful reservation, the <INFO-SPEC> object SHOULD have the following values:

  Error severity Class: Success
  Error code value: Reservation successful

* In the case of unsuccessful reservation the <INFO-SPEC> object SHOULD have the following values:

  Error severity class: Transient Failure
  Error code value: 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 the <E> flag associated with the QSPEC <QoS Reserved> object and the <E> flag associated with the local RMD-QSPEC <TMOD-1> parameter are set. Furthermore, the <M> and <S> PDR container bits are set to "1".

4.6.1.4. RMD Modification of Aggregated Reservations

In the case when the QNE Edges maintain QoS-NSLP-aggregated operational and 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 <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired>, 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" that is associated with the local RMD-QSPEC <TMOD-1> parameter MUST be <M> marked,

i.e., the <M> bit is set to the value of "1". 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 <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired>. Subsequently, an RMD release procedure SHOULD be accomplished (see Section 4.6.1.5). Note that if the complete bandwidth associated with the aggregated reservation maintained at the QNE Ingress does not have to be released, then the <TEAR> flag MUST be set to OFF. This is because the NSLP operational states associated with the aggregated reservation states at the Edge QNEs MUST NOT be turned off. However, if the complete bandwidth associated with the aggregated reservation maintained at the QNE Ingress has to be released, then the <TEAR> flag MUST be set to ON.

It is important to emphasize that this RMD modification scheme only applies to the following two RMD-QOSM schemes:

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure.

4.6.1.5. RMD Release Procedure

This procedure is applied to all RMD mechanisms that maintain reservation states. If a refresh RESERVE message does not arrive at a QNE Interior node within the refresh timeout period, then the bandwidth requested by this refresh RESERVE message is not updated. This means that the reserved bandwidth associated with the reduced state is decreased in the next refresh period by the amount of the corresponding bandwidth that has not been refreshed, see Section 4.3.3.

This soft state behavior provides certain robustness for the system ensuring that unused resources are not reserved for a long time. Resources can be removed by an explicit release at any time. However, in the situation that an end-to-end (tear) RESERVE is retransmitted (see Section 5.2.4 in [RFC5974]), then this message MUST NOT initiate an intra-domain (tear) RESERVE message. This is because the amount of bandwidth within the RMD domain associated with

the (tear) end-to-end RESERVE has already been released, and therefore, this amount of bandwidth within the RMD domain MUST NOT once again be released.

When the RMD-RMF of a QNE Edge or QNE Interior node processes a "PHR_Release_Request" PHR container, 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, see [RMD1] and [CsTa05] for more details. 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". When the above mentioned procedure of indicating the "T_Lag" is used and when a node (QNE Egress or QNE Interior) receives the "PHR_Release_Request" PHR container, it MUST store the arrival time. Then, it MUST calculate the time difference, "T_diff", 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 <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired> that has been sent together with the "PHR_Release_Request" "PHR Container", but not below zero.

An RMD-specific release procedure can be triggered by an end-to-end RESERVE with a <TEAR> flag set to 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 to ON) "PDR_Reservation_Report" or "PDR_Congestion_Report" and/or an <INFO-SPEC> object.

4.6.1.5.1. Triggered by a RESERVE Message

This RMD-explicit release procedure can be triggered by a tear (<TEAR> flag set to ON) end-to-end RESERVE message. When a tear (<TEAR> flag set ON) end-to-end RESERVE message arrives to the QNE Ingress, the QNE Ingress node SHOULD process the message in a standard QoS-NSLP way (see [RFC5974]). In addition to this, the RMD RMF is notified, as specified in [RFC5974].

Like the scenario described in 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 NSLPID value to another NSLPID predefined value that will be used by the GIST message that carries the end-to-end RESERVE message to bypass the QNE Interior nodes.

Before generating an intra-domain tear RESERVE, the RMD-QOSM has to release the requested RMD-QOSM bandwidth from the RMD traffic class state maintained at the QNE Ingress.

This can be achieved by identifying the traffic class (PHB) and then subtracting the amount of RMD traffic class requested resources, included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter, from the total reserved amount of resources stored in the RMD traffic class state. The <Time Lag> is used as explained in the introductory part of Section 4.6.1.5.

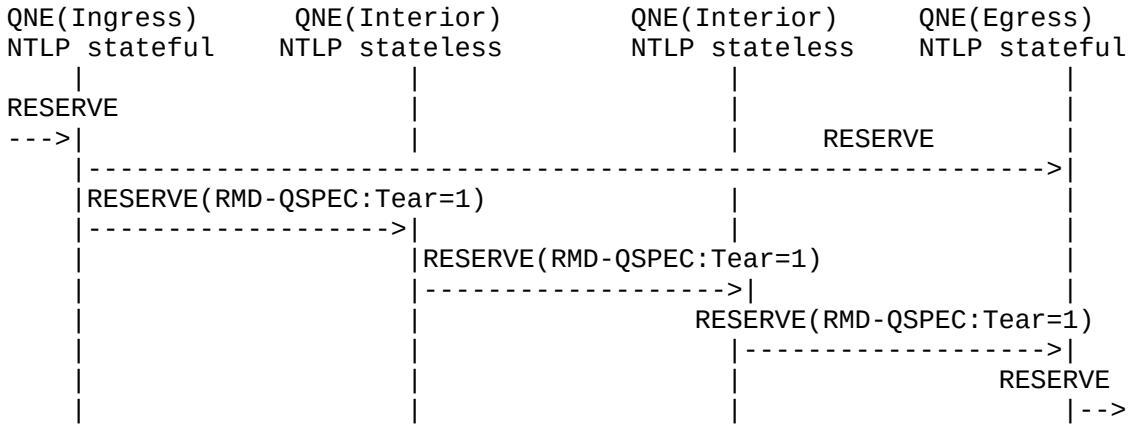

Figure 11: Explicit release triggered by RESERVE used by the RMD-QOSM

After that, the REQUIRED bandwidth is released from the RMD-QOSM traffic class state at the QNE Ingress, an intra-domain RESERVE (RMD-QOSM) message has to be generated. The intra-domain RESERVE (RMD-QSPEC) message MUST include an <RMD QoS object combination> field and a PHR container, (i.e., "PHR_Release_Request") 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 (see the QoS-NSLP-RMF API described in [RFC5974]):

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

* if the release procedure is not applied for the RMD modification of aggregated reservation procedure (see Section 4.6.1.4), then the <TEAR> flag MUST be set to ON;

* the PHR resource units MUST be included into the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired>;

* 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 Section 4.6.1.5) the value of the <Control Type> parameter of the PHR container is set to "18" (i.e., PHR_Release_Request).

Any QNE Interior node that receives the combination of the RMD-QOSM <QoS Desired> object and the "PHR_Release_Request" control information container MUST identify the traffic class (PHB) and release the requested resources included in the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter. This can be achieved by subtracting the amount of RMD traffic class requested resources, included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> 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_Release_Request" container is used during the release procedure as explained in the introductory part of Section 4.6.1.5.

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

The value of the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired> 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 <Peak Data Rate-1 (p)> field value of the local RMD-QSPEC <TMOD-1> 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 NSLPID value to the end-to-end RESERVE message (see Section 4.4).

Note that when the QNE Edges maintain aggregated QoS-NSLP reservation states, the RMD-QOSM functionality MAY start an RMD modification procedure (see Section 4.6.1.4) that uses the explicit release procedure, described above in this subsection. Note that if the complete bandwidth associated with the aggregated reservation maintained at the QNE Ingress has to be released, then the <TEAR> flag MUST be set to ON. Otherwise, the <TEAR> flag MUST be set to OFF, see Section 4.6.1.4.

4.6.1.5.2. Triggered by a Marked RESPONSE or NOTIFY Message

This RMD explicit release procedure can be triggered by either an intra-domain RESPONSE message with a PDR container carrying among others the <M> and <S> parameters with values <M>=1 and <S>=0 (see Section 4.6.1.2), an intra-domain (refresh) RESPONSE message carrying a PDR container with <M>=1 and <S>=1 (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: Informational
Error code value: Congestion situation

When the aggregated intra-domain QoS-NSLP operational states are used, an end-to-end NOTIFY message used to trigger an RMD release procedure MAY contain a PDR container that carries an <M> and an <S> with values <M>=1 and <S>=1, and a bandwidth value in the <PDR Bandwidth> parameter included in a "PDR_Refresh_Report" or "PDR_Congestion_Report" container.

Note that in all explicit release procedures, before generating an intra-domain tear RESERVE, the RMD-QOSM has to release the requested RMD-QOSM bandwidth from the RMD traffic class state maintained at the QNE Ingress. This can be achieved by identifying the traffic class (PHB) and then subtracting the amount of RMD traffic class requested

resources, included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter, from the total reserved amount of resources stored in the RMD traffic class state.

Figure 12 shows the situation that the intra-domain tear RESERVE is generated after being triggered by either an intra-domain (refresh) RESPONSE message that carries a PDR container with <M>=1 and <S>=1 or by an end-to-end NOTIFY message that does not carry a PDR container, but an <INFO-SPEC> object. The error code values carried by this NOTIFY message are:

Error severity class: Informational
Error code value: Congestion situation

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 (see the QoS-NSLP-RMF described in [RFC5974]):

* 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. Note that this is needed if a bidirectional scenario is used; see Section 4.6.2.

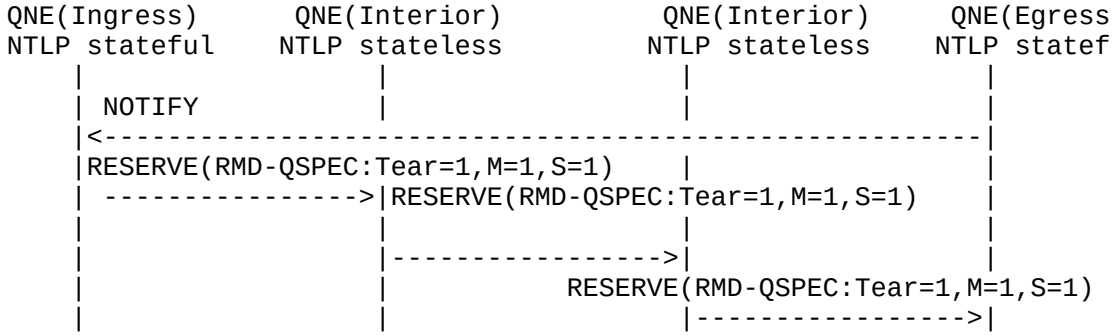

Figure 12: Basic operation during RMD-explicit release procedure triggered by NOTIFY used by the RMD-QOSM

Note that if the values of the <M> and <S> parameters included in the PHR container carried by a intra-domain tear RESERVE(RMD-QOSM) are set as ((<M>=0 and <S>=1) or (<M>=0 and <S>=0) or (<M>=1 and <S>=1)),

then the <Max Admitted Hops> value SHOULD NOT be compared to the <Admitted Hops> value and the value of the <K> field MUST NOT be set. Any QNE Edge or QNE Interior node that receives the intra-domain tear RESERVE MUST check the <K> field included in the PHR container. If the <K> field is "0", then the traffic class state (PHB) has to be identified, using the <PHB Class> parameter, and the requested resources included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter have to be released.

This can be achieved by subtracting the amount of RMD traffic class requested resources, included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> 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 field is used during the release procedure, as explained in the introductory part of Section 4.6.1.5. Afterwards, the QNE Egress node MUST terminate the tear intra-domain RESERVE(RMD-QSPEC) message.

The RMD-specific release procedure that is triggered by an intra-domain RESPONSE message with an <M>=1 and <S>=0 PDR container (see Section 4.6.1.2) generates an intra-domain tear RESERVE message that uses the combination of the <Max Admitted Hops> and <Admitted_Hops> fields to calculate and specify when the <K> value carried by the "PHR Container" can be set. When the <K> field is set, then the "PHR Container" and the RMD-QOSM <QoS Desired> carried by an intra-domain tear RESERVE MUST NOT be processed.

The RMD-specific explicit release procedure that uses the combination of <Max Admitted Hops>, <Admitted_Hops> and <K> fields to release resources/bandwidth in only a part of the RMD domain, is denoted as RMD partial release procedure.

This explicit release procedure can be used, for example, during unsuccessful reservation (see Section 4.6.1.2). When the RMD-QOSM/QoS-NSLP signaling model functionality of a QNE Ingress node receives a PDR container with values <M>=1 and <S>=0, of type "PDR_Reservation_Report", it MUST start an RMD partial release procedure.

In this situation, after the REQUIRED bandwidth is released from the RMD-QOSM traffic class state at the QNE Ingress, an intra-domain RESERVE (RMD-QOSM) message has to be generated. 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 value of the <M> parameter of the PHR container MUST be set to "1".

* the RESERVE message MAY include a PDR container.

* the value of the <Max Admitted Hops> carried by the "PHR Container" MUST be set equal to the <Max Admitted Hops> value carried by the "PDR Container" (with <M>=1 and <S>=0) carried by the received intra-domain RESPONSE message that triggers the release procedure.

Any QNE Edge or QNE Interior node that receives the intra-domain tear RESERVE has to check the value of the <K> field in the "PHR Container" before releasing the requested resources.

If the value of the <K> field is "1", then all the QNEs located downstream, including the QNE Egress, MUST NOT process the carried "PHR Container" and the RMD-QOSM <QoS Desired> object by the intra-domain tearing RESERVE.

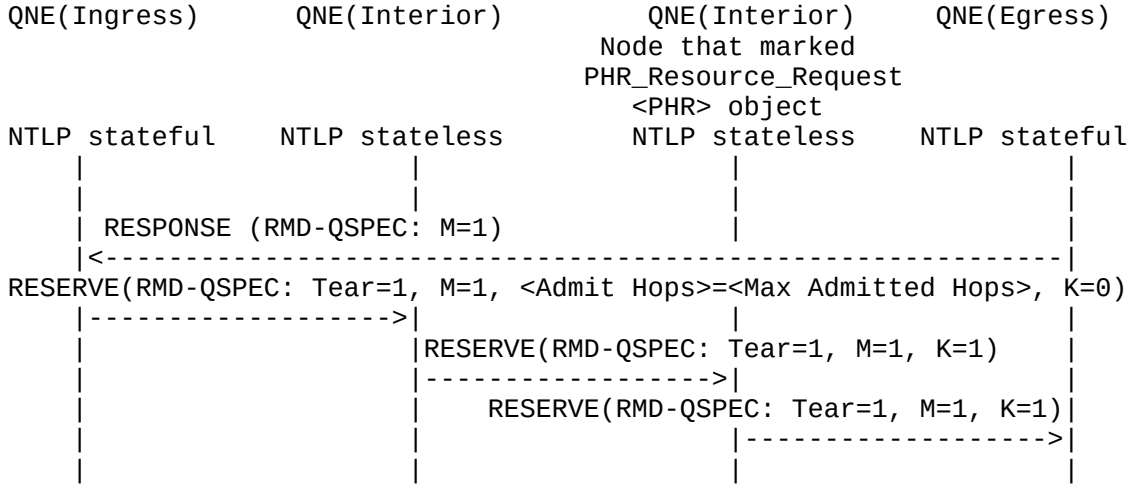

Figure 13: Basic operation during RMD explicit release procedure triggered by RESPONSE used by the RMD-QOSM

If the <K> field value is "0", any QNE Edge or QNE Interior node that receives the intra-domain tear RESERVE can release the resources by subtracting the amount of RMD traffic class requested resources, included in the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> 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 field is used during the release procedure as explained in the introductory part of Section 4.6.1.5.

Furthermore, the QNE MUST perform the following procedures.

If the values of the <M> and <S> parameters included in the "PHR_Release_Request" PHR container are (<M=1> and <S>=0) then the <Max Admitted Hops> value MUST be compared with the calculated <Admitted Hops> value. Note that each time that the intra-domain tear RESERVE is processed and before being forwarded by a QNE, the <Admitted Hops> value included in the PHR container is increased by one.

When these two values are equal, the intra-domain RESERVE(RMD-QSPEC) that is forwarded further towards the QNE Egress MUST set the <K> value of the carried "PHR Container" to "1".

The reason for doing this is that the QNE node that is currently processing this message was the last QNE node that successfully processed the RMD-QOSM <QoS Desired>) 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> and <Hop_U> parameters of the "PHR_Resource_Request".

Finally, note that the QNE Egress node MUST terminate the intra-domain RESERVE(RMD-QSPEC) message.

Moreover, note that the above described RMD partial release procedure applies to the situation that the QNE Edges maintain a per-flow QoS-NSLP reservation state.

When the QNE Edges maintain aggregated intra-domain QoS-NSLP operational states and a severe congestion occurs, then the QNE Ingress MAY receive an end-to-end NOTIFY message (see Section 4.6.1.6) with a PDR container that carries the <M>=0 and <S>=1 fields and a bandwidth value in the <PDR Bandwidth> parameter included in a "PDR_Congestion_Report" container. Furthermore, the same end-to-end NOTIFY message carries an <INFO-SPEC> object with the following values:

Error severity class: Informational
Error code value: Congestion situation

The end-to-end session associated with this NOTIFY message maintains the BOUND-SESSION-ID of the bound aggregated session; see Section 4.3.1. The RMD-QOSM at the QNE Ingress MUST start an RMD modification procedures (see Section 4.6.1.4) that uses the RMD explicit release procedure, described above in this section. In particular, the RMD explicit release procedure releases the bandwidth value included in the <PDR Bandwidth> parameter, within the "PDR_Congestion_Report" container, from the reserved bandwidth associated with the aggregated intra-domain QoS-NSLP operational state.

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 rerouted 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

This procedure applies to all RMD scenarios that use an RMD 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 setting <S>=1 (<M>=1) marking the refresh messages and by setting the <O> field in the "PHR_Refresh_Update" container, carried by the intra-domain RESERVE message.

Note that the overload situation can be detected by using the example given in Appendix A.1. In this situation, when the given signaled_overload_rate parameter given in Appendix A.1 is higher than 0, the value of the <Overload> field is set to "1". The calculation

of this is given in Appendix A.1 and denoted as the signaled_overload_rate parameter. The flows can be terminated by the RMD release procedure described in Section 4.6.1.5.

The intra-domain RESPONSE message that is sent by the QNE Egress towards the QNE Ingress will contain a PDR container with a Parameter ID = 26, i.e., "PDR_Congestion_Report". The values of the <M>, <S>, and <O> fields of this container SHOULD be set equal to the values of the <M>, <S>, and <O> fields, respectively, carried by the "PHR_Refresh_Update" container. Part of the flows, corresponding to the <O>, 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.

Furthermore, note that the above functionalities also apply to the scenario in which 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 detection and marking/re-marking functionality described in this section applies to NSIS-aware and NSIS-unaware nodes. This means however, that the "not NSIS-aware" nodes MUST be configured such that they can detect the congestion/severe congestion situations and re-mark packets in the same way the "NSIS-aware" nodes do.

The Interior node detecting severe congestion re-marks data packets passing the node. For this re-marking, 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 an "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, e.g., due to flow-based ECMP-enabled (Equal Cost Multiple Paths) routing, the Egress node either does not detect packets passed a severely

congested node or erroneously detects packets that actually did not pass the severely 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".

In this document, note that the terms "marked packets" or "marked bytes" refer to the "encoded DSCP". The terms "unmarked packets" or "unmarked bytes" represent 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 an RMD domain use the same, fixed-measurement interval, say T seconds, which MUST be preconfigured.

It is RECOMMENDED that the total number of additional (local and experimental) DSCPs needed for severe congestion handling within an RMD domain SHOULD be as low as possible, and it SHOULD NOT exceed the limit of 8. 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 that can be supported by RMD-QOSM.

An example of a re-marking procedure can be found in Appendix A.1.

4.6.1.6.2.2. Operation in the Egress Nodes

When the QNE Edges maintain a per-flow intra-domain QoS-NSLP operational state (see Sections 4.3.2 and 4.3.3), then the following procedure is followed. 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, 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. This is because, otherwise, the QNE Egress will not have any information on which flows or groups of flows were affected by the severe congestion.

When the RMD domain supports the "affected DSCP" marking, 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, the Egress MAY not generate and maintain the Ingress/Egress pair aggregated reservation states. Note that these aggregated reservation states MAY not be associated with aggregated intra-domain QoS-NSLP operational states.

The Ingress/Egress pair aggregated reservation 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 by either the RESERVE message that carries the QSPEC associated with the end-to-end QoS Model, e.g.,, <Preemption Priority> and <Defending Priority> parameter or using a locally defined policy. The priority value is kept in the reservation states (see Section 4.3), which might be used during admission control and/or severe congestion handling procedures. 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 belong 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.2.

For the flows (sessions) that have to be terminated, the QNE Egress node generates and sends an end-to-end 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 (see QoS-NSLP-RMF API described in [RFC5974]):

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

* the <INFO-SPEC> object MUST include information that notifies that the end-to-end flow MUST be terminated. This information is as follows:

Error severity class: Informational
Error code value: Congestion situation

When the QNE Edges maintain a per-aggregate intra-domain QoS-NSLP operational state (see Section 4.3.1), the QNE Edge has to calculate, per each aggregate intra-domain QoS-NSLP operational state, the total bandwidth that has to be terminated in order to solve the severe congestion. The total bandwidth to be released is calculated in the same way as in the situation in which the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states. Note that for the aggregated sessions that are affected, the QNE Egress node generates and sends one end-to-end NOTIFY message to the QNE Ingress node (its upstream stateful QoS-NSLP peer) to indicate the severe congestion in the communication path. Note that this end-to-end NOTIFY message is associated with one of the end-to-end sessions that is bound to the aggregated intra-domain QoS-NSLP operational state.

The non-default values of the objects contained in the NOTIFY message MUST be set by the QNE Egress node in the same way as the ones used by the end-to-end NOTIFY message described above for the situation that the QNE Egress maintains a per-flow intra-domain operational state. In addition to this, the end-to-end NOTIFY MUST carry the RMD-QSPEC, which contains a PDR container with a Parameter ID = 26, i.e., "PDR_Congestion_Report". The value of the <S> SHOULD be set. Furthermore, the value of the <PDR Bandwidth> parameter MUST contain the bandwidth associated with the aggregated QoS-NSLP operational state, which has to be released.

Furthermore, the number of end-to-end sessions that have to be terminated will be calculated as in the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states. Similarly for each, to be terminated, ongoing flow, the Egress will notify the Ingress in the same way as in the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states.

Note that the QNE Egress SHOULD restore the original <DSCP> values of the re-marked packets; otherwise, multiple actions for the same event might occur. However, this value MAY be left in its re-marking form if there is an SLA agreement between domains that a downstream domain handles the re-marking problem.

An example of a detailed severe congestion operation in the Egress Nodes can be found in Appendix A.2.

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), and blocking their packets 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 that carries the "Congestion situation" error code.

When the QNE Ingress node receives the end-to-end NOTIFY message, it associates this NOTIFY message with its bound intra-domain session (see Sections 4.3.2 and 4.3.3) 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.

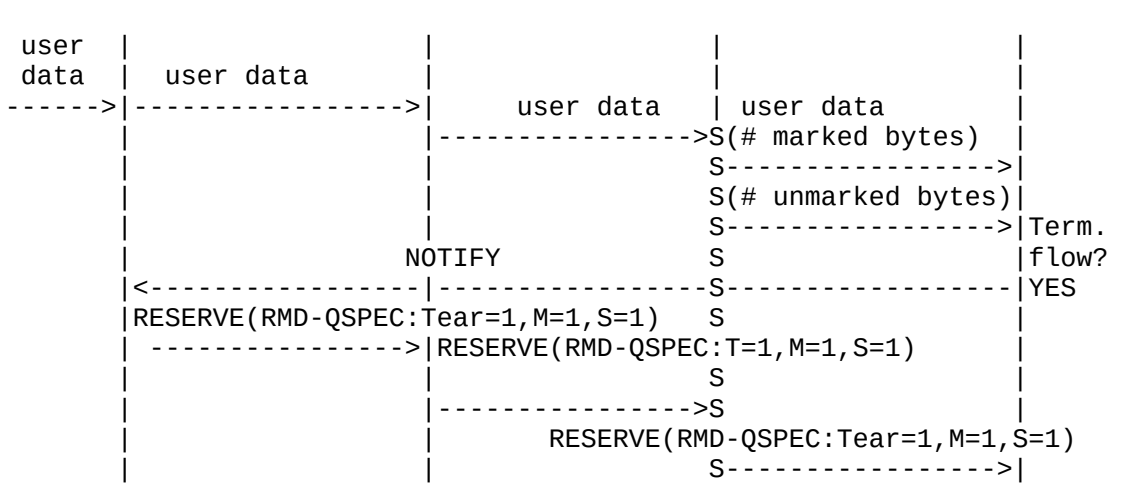

Figure 14: RMD severe congestion handling

Note that the above functionality applies to the RMD reservation-based (see Section 4.3.3) and to both measurement-based admission control methods (i.e., congestion notification based on probing and the NSIS measurement-based admission control; see Section 4.3.2).

In the case that the QNE Edges support aggregated intra-domain QoS-NSLP operational states, the following actions take place. The QNE Ingress MAY receive an end-to-end NOTIFY message with a PDR container that carries an <S> marked and a bandwidth value in the <PDR

Bandwidth> parameter included in a "PDR_Congestion_Report" container. Furthermore, the same end-to-end NOTIFY message carries an <INFO-SPEC> object with the "Congestion situation" error code.

When the QNE Ingress node receives this end-to-end NOTIFY message, it associates the NOTIFY message with the aggregated intra-domain QoS-NSLP operational state via the BOUND-SESSION-ID information included in the end-to-end per-flow QoS-NSLP operational state, see Section 4.3.1.

The RMD-QOSM at the QNE Ingress node by using the total bandwidth value to be released included in the <PDR Bandwidth> parameter MUST reduce the bandwidth associated and reserved by the RMD aggregated session. This is accomplished by triggering the RMD modification for aggregated reservations procedure described in Section 4.6.1.4.

In addition to the above, the QNE Ingress MUST select a number of inter-domain (end-to-end) flows (sessions) that MUST be terminated. This is accomplished in the same way as in the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states.

The terminated end-to-end sessions are selected from the end-to-end sessions bound to the aggregated intra-domain QoS-NSLP operational state. Note that the end-to-end session associated with the received end-to-end NOTIFY message that notified the severe congestion MUST also be selected for termination.

For the flows (sessions) that have to be terminated, the QNE Ingress node generates and sends an end-to-end NOTIFY message upstream towards the sender (QNI). The values carried by this message are:

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

* the <INFO-SPEC> object MUST include information that notifies that the end-to-end flow MUST be terminated. This information is as follows:

  Error severity class: Informational
  Error code value: Congestion situation

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.

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. In this way, it is ensured that the end-to-end RESERVE (probe) packet passed through the node that it is congested. This feature is very useful when ECMP-based routing is used to detect only flows that are passing through the congested router.

When a (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 do not need to be NSIS-aware nodes and they do not need to process the NSIS functionality of NSIS messages. Note that the "not NSIS-aware" nodes MUST be configured such that they can detect the congestion/severe congestion situations and re-mark packets in the same way the "NSIS-aware" nodes do.

Using standard functionalities, congestion notification thresholds are set for the traffic that belongs 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 all data packets and 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, the data packets are also re-marked. An example of the detailed operation of this procedure is given in Appendix A.2.

4.6.1.7.3. Operation in Egress Nodes

As emphasized 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 (RMF) reservation state for each supported PHB. Note that this aggregated reservation state does not require that an aggregated intra-domain QoS-NSLP operational state is needed also.

Appendix A.4 contains an example of 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 unsuccessful. In particular, it will generate an <INFO-SPEC> object of:

Error severity class: Transient Failure
Error code value: Reservation failure

The QSPEC that was carried by the end-to-end RESERVE that belongs 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 local RMD-QSPEC <TMOD-1> 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 the QNE Egress SHOULD restore the original <DSCP> values of the re-marked packets; otherwise, multiple actions for the same event might occur. However, this value MAY be left in its re-marking form if there is an SLA agreement between domains that a downstream domain handles the re-marking problem. Note that the break <B> flag carried by the end-to-end RESERVE message MUST NOT be set.

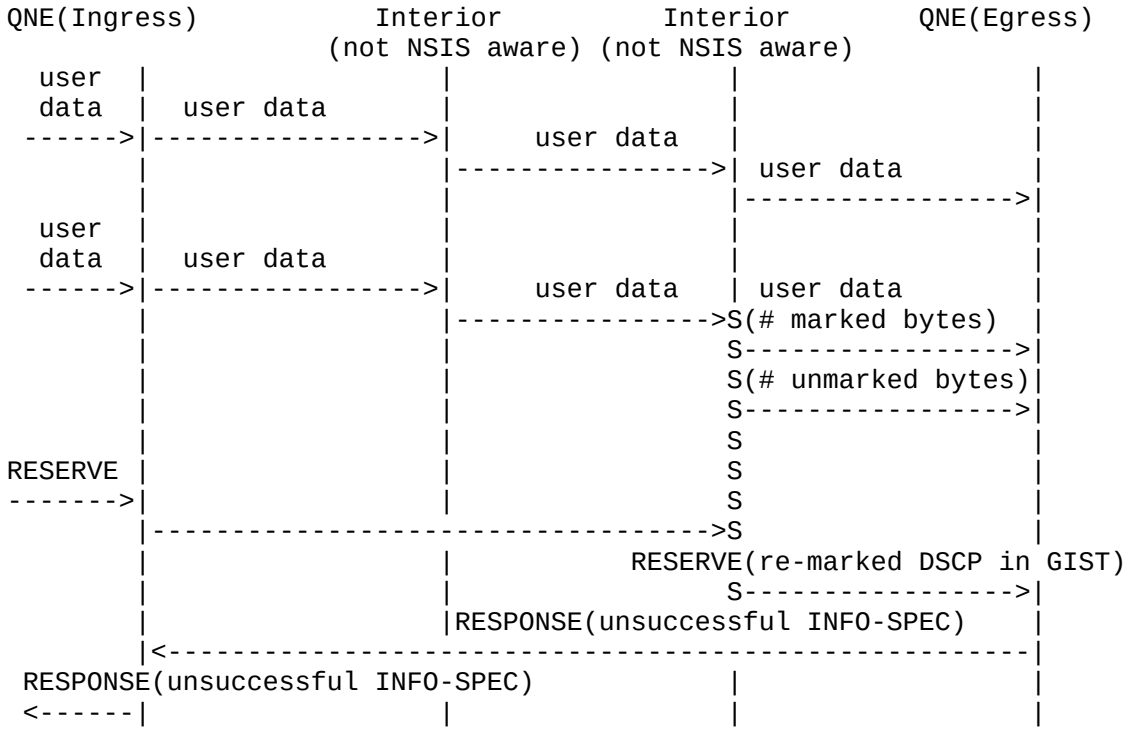

Figure 15: Using RMD congestion notification function for admission control based on probing

4.6.2. Bidirectional Operation

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

* Successful and unsuccessful reservation (Section 4.6.2.1);
* Refresh reservation (Section 4.6.2.2);
* Modification of aggregated reservation (Section 4.6.2.3);
* Release procedure (Section 4.6.2.4);
* Severe congestion handling (Section 4.6.2.5);
* Admission control using congestion notification based on probing (Section 4.6.2.6).

It is important to emphasize that the content of this section is used for the specification of the following RMD-QOSM/QoS-NSLP signaling schemes, when basic unidirectional operation is assumed:

* "per-flow congestion notification based on probing";

* "per-flow RMD NSIS measurement-based admission control",

* "per-flow RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

* "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure;

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by the RMD-QOSM refresh" procedure;

* "per-aggregate RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" procedure.

For more details, please see Section 3.2.3.

In particular, the functionality described in Sections 4.6.2.1, 4.6.2.2, 4.6.2.3, 4.6.2.4, and 4.6.2.5 applies to the RMD reservation-based and NSIS measurement-based admission control methods. The described functionality in Section 4.6.2.6 applies to the admission control procedure that uses the congestion notification based on probing. The QNE Edge nodes maintain either per-flow QoS-NSLP operational and reservation states or aggregated QoS-NSLP operational and reservation states.

RMD-QOSM assumes that asymmetric routing MAY be applied in the RMD domain. 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 bidirectional 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. Note that if this is not the case, then the bidirectional procedure could be managed and maintained by nodes located outside the RMD domain, by using other procedures than the ones defined in RMD-QOSM.

This (intra-domain) bidirectional sender&sender procedure can then be applied between the QNE Edge (QNE Ingress and QNE Egress) nodes of the RMD QoS signaling model. In the situation in which a security association exists between the QNE Ingress and QNE Egress nodes (see Figure 15), and the QNE Ingress node has the REQUIRED <Peak Data Rate-1 (p)> values of the local RMD-QSPEC <TMOD-1> parameters for both directions, i.e., QNE Ingress towards QNE Egress and QNE Egress

towards QNE Ingress, then the QNE Ingress MAY include both <Peak Data Rate-1 (p)> values of the local RMD-QSPEC <TMOD-1> 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 <BOUND-SESSION-ID> objects at the QNE Ingress and QNE Egress, which will be carried by bound end-to-end RESERVE messages.

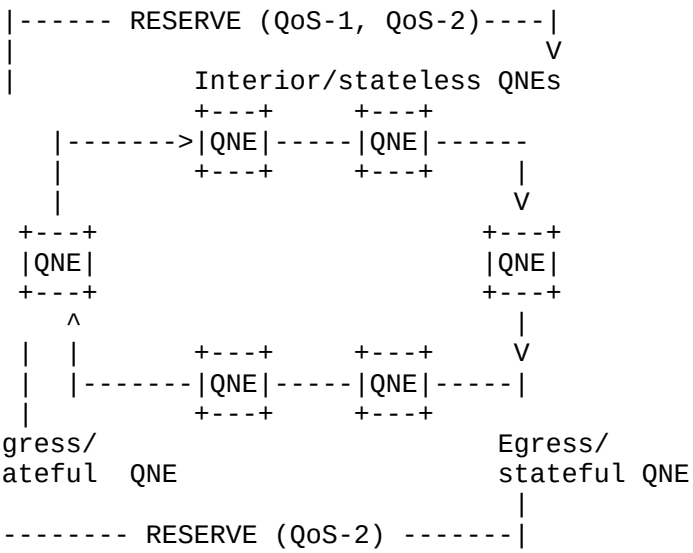

Figure 16: The intra-domain bidirectional reservation scenario in the RMD domain

Note that it is RECOMMENDED that the QNE implementations of RMD-QOSM process the QoS-NSLP signaling messages with a higher priority than data packets. This can be accomplished as described in Section 3.3.4 in [RFC5974] and the QoS-NSLP-RMF API [RFC5974].

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.

When the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states, the end-to-end RESERVE message carries two BOUND-SESSION-IDs. One BOUND-SESSION-ID carries the SESSION-ID of the tunneled intra-domain (per-flow) session that is using a Binding_Code with value set to code (Tunneled and end-to-end sessions). Another

BOUND-SESSION-ID carries the SESSION-ID of the bound bidirectional end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Bidirectional sessions).

When the QNE Edges maintain aggregated intra-domain QoS-NSLP operational states, the end-to-end RESERVE message carries two BOUND-SESSION-IDs. One BOUND-SESSION-ID carries the SESSION-ID of the tunneled aggregated intra-domain session that is using a Binding_Code with value set to code (Aggregated sessions). Another BOUND-SESSION-ID carries the SESSION-ID of the bound bidirectional end-to-end session. The Binding_Code associated with this BOUND-SESSION-ID is set to code (Bidirectional sessions).

The intra-domain and end-to-end QoS-NSLP operational states are initiated/modified depending on the binding type (see Sections 4.3.1, 4.3.2, and 4.3.3).

If no security association exists between the QNE Ingress and QNE Egress nodes, the bidirectional reservation for the sender&sender scenario in the RMD domain SHOULD use the scenario specified in [RFC5974] as "bidirectional reservation for sender&sender scenario". This is because in this scenario, the RESERVE message sent from the QNE Ingress to QNE Egress does not have to carry the QoS parameters needed for the "Egress towards Ingress" direction (QoS-2).

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 <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameters for both directions, i.e., QNE Ingress towards QNE Egress and QNE Egress towards QNE Ingress.

According to Section 3.2.3, it is specified that only the "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" scheme MUST be implemented within one RMD domain. However, all RMD QNEs supporting this specification MUST support the combination the "per-flow RMD reservation-based" in combination with the "severe congestion handling by proportional data packet marking" scheme. If the RMD QNEs support more RMD-QOSM schemes, then the operator of that RMD domain MUST preconfigure all the QNE Edge nodes within one domain such that the <SCH> field included in the "PHR Container" (Section 4.1.2) and the "PDR Container" (Section 4.1.3) will always use the same value, such that within one RMD domain, only one of the below described RMD-QOSM schemes is used at a time.

All QNE nodes located within the RMD domain MUST read and interpret the <SCH> field included in the "PHR Container" before processing all the other <PHR Container> payload fields. Moreover, all QNE Edge nodes located at the boarder of the RMD domain, MUST read and interpret the <SCH> field included in the "PDR container" before processing all the other <PDR Container> payload fields.

4.6.2.1. Successful and Unsuccessful Reservations

This section describes the operation of the RMD-QOSM where an RMD Intra-domain bidirectional reservation operation, see Figure 16 and Section 4.6.2, is either successfully or unsuccessfully accomplished.

The bidirectional 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 bidirectional successful reservation procedure with the combination of two unidirectional successful reservations accomplished in opposite directions are as follows. Note also that the intra-domain and end-to-end QoS-NSLP operational states generated and maintained by the end-to-end RESERVE messages contain, compared to the unidirectional reservation scenario, a different BOUND-SESSION-ID data structure (see Sections 4.3.1, 4.3.2, and 4.3.3). In this scenario, 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 intra-domain RESERVE (RMD-QSPEC): "forward" message used for the bidirectional successful reservation procedure and a RESERVE (RMD-QSPEC) message used for the unidirectional successful reservation are as follows (see the QoS-NSLP-RMF API described in [RFC5974]):

* the <RII> object MUST NOT be included in the message. This is because no RESPONSE message is REQUIRED.

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

* the PDR container is also included in the RESERVE(RMD-QSPEC): "forward" message. The value of the Parameter ID is "20", 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 bidirectional reservation and is set to "1".

* the <PDR Bandwidth> 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 bidirectional successful reservation procedure and a RESERVE(RMD-QSPEC) message used for the unidirectional successful reservation are as follows:

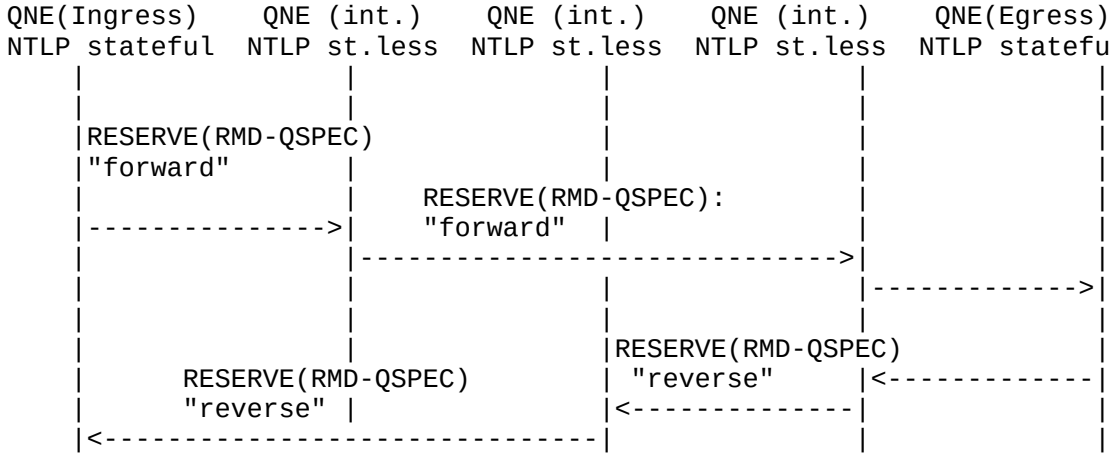

Figure 17: Intra-domain signaling operation for successful bidirectional reservation

* the <RII> object is not included in the message. This is because no RESPONSE message is REQUIRED;

* the value of the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter is set equal to the value of the <PDR Bandwidth> 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 bidirectional reservation and is set to "1";

* the PDR container is included into the RESERVE(RMD-QSPEC): "reverse" message. The value of the Parameter ID is "23", i.e., "PDR_Reservation_Report";

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

Figures 18 and 19 show the flow diagrams used in the case of an unsuccessful bidirectional reservation. In Figure 18, the QNE that is not able to support the requested <Peak Data Rate-1 (p)> value of local RMD-QSPEC <TMOD-1> is located in the direction QNE Ingress towards QNE Egress. In Figure 19, the QNE that is not able to support the requested <Peak Data Rate-1 (p)> value of local RMD-QSPEC <TMOD-1> is located in the direction QNE Egress towards QNE Ingress. The main differences between the bidirectional unsuccessful procedure shown in Figure 18 and the bidirectional 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., the path from the QNE Ingress towards the QNE Egress.

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

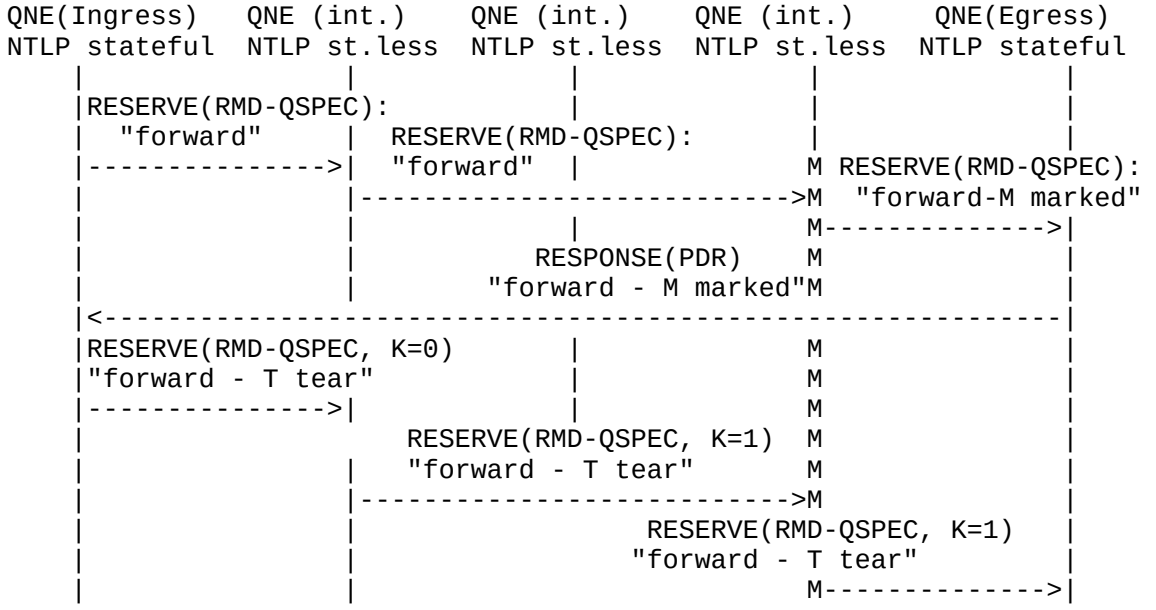

Figure 18: Intra-domain signaling operation for unsuccessful bidirectional reservation (rejection on path QNE(Ingress) towards QNE(Egress))

The operation for this type of unsuccessful bidirectional reservation is similar to the operation for unsuccessful unidirectional reservation, shown in Figure 9.

The main differences between the bidirectional unsuccessful procedure shown in Figure 19 and the in bidirectional 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., the path from the QNE Egress towards the QNE Ingress.

* the QNE node that is not able to support the requested <Peak Data Rate-1 (p)> value of local RMD-QSPEC <TMOD-1> MUST mark the <M> bit, i.e., set to value "1", the RESERVE(RMD-QSPEC): "reverse".

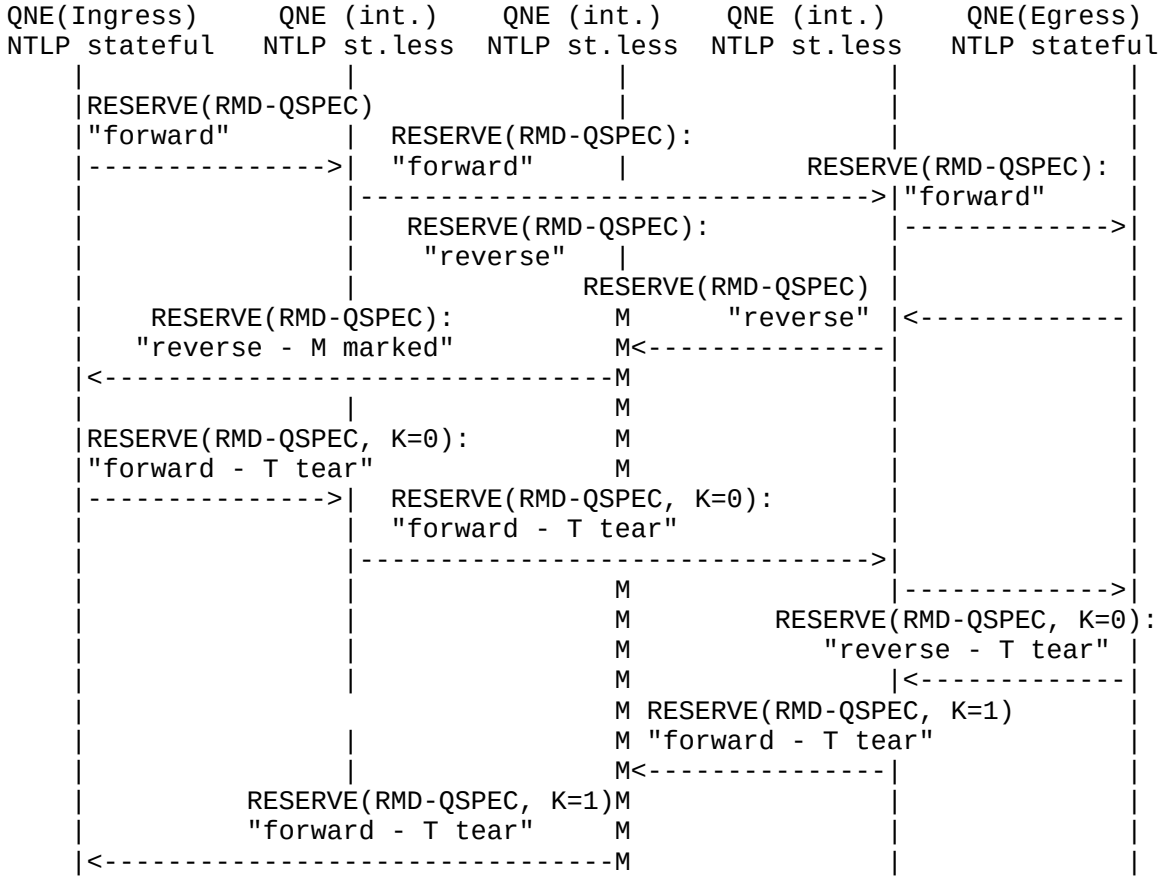

Figure 19: Intra-domain signaling normal operation for unsuccessful bidirectional reservation (rejection on path QNE(Egress) towards QNE(Ingress)

* 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 RESERVE(RMD-QSPEC): "forward - T tear" message. This message carries a "PHR_Release_Request" and "PDR_Release_Request" control information. This message is sent to the QNE Egress node. The QNE Egress node uses the information contained in the "PHR_Release_Request" and the "PDR_Release_Request" control info containers to generate a RESERVE(RMD-QSPEC): "reverse - T tear" message that is sent towards the QNE Ingress node.

4.6.2.2. Refresh Reservations

This section describes the operation of the RMD-QOSM where an RMD intra-domain bidirectional refresh reservation operation is accomplished.

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

Note that the RMD traffic class refresh periods used by the bound bidirectional 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 bidirectional refresh procedure and a RESERVE(RMD-QSPEC): "forward" message used for the bidirectional successful reservation procedure are as follows:

* the value of the Parameter ID of the PHR container is "19", i.e., "PHR_Refresh_Update".

* the value of the Parameter ID of the PDR container is "21", i.e., "PDR_Refresh_Request".

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

* the value of the Parameter ID of the PHR container is "19", i.e., "PHR_Refresh_Update".

* the value of the Parameter ID of the PDR container is "24", i.e., "PDR_Refresh_Report".

4.6.2.3. Modification of Aggregated Intra-Domain QoS-NSLP Operational Reservation States

This section describes the operation of the RMD-QOSM where RMD intra-domain bidirectional QoS-NSLP aggregated reservation states have to be modified.

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 <Peak Data Rate-1 (p)> field local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired>), which is sent together with "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" associated with the local RMD-QSPEC <TMOD-1> parameter MUST be marked. In this situation, the RMD-specific operation for unsuccessful reservation will be applied (see Section 4.6.2.1). Note that the value of the <PDR Bandwidth> parameter, which is sent within a "PDR_Reservation_Request" container, represents the increase of the reserved resources in the "reverse" direction.

* When the modification request requires a decrease of the reserved resources, the QNE Ingress node MUST include this value into the <Peak Data Rate-1 (p)> field of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired>). Subsequently, an RMD release procedure SHOULD be accomplished (see Section 4.6.2.4). Note that the value of the <PDR Bandwidth> parameter, which is sent within a "PDR_Release_Request" container, represents the decrease of the reserved resources in the "reverse" direction.

4.6.2.4. Release Procedure

This section describes the operation of the RMD-QOSM, where an RMD intra-domain bidirectional 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, how the release of the reservation is accomplished is similar to the RMD release procedure used for the intra-domain unidirectional reservations (see Section 4.6.1.5 and Figures 18 and 19).

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

* the value of the Parameter ID of the PHR container is "18", i.e."PHR_Release_Request";

* the value of the Parameter ID of the PDR container is "22", i.e., "PDR_Release_Request";

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

* the value of the Parameter ID of the PHR container is "18", 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 RMD intra-domain bidirectional 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 Bidirectional 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 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 an RMD intra-domain bidirectional reservation procedure. Note that the detection and marking/re-marking functionality described in this section and used by Interior nodes, applies to NSIS-aware but also to NSIS-unaware nodes. This means however, that the "not NSIS-aware" Interior nodes MUST be configured such that they can detect the congestion situations and re-mark packets in the same way as the Interior "NSIS-aware" nodes do.

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" or "reverse" path. Note that when a severe congestion situation occurs, e.g., on 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.5.

Furthermore, a special case of this operation is associated with the severe congestion situation occurring simultaneously on the forward and reverse paths. An example of this operation is given in Appendix A.6.

Simulation results associated with these procedures can be found in [DiKa08].

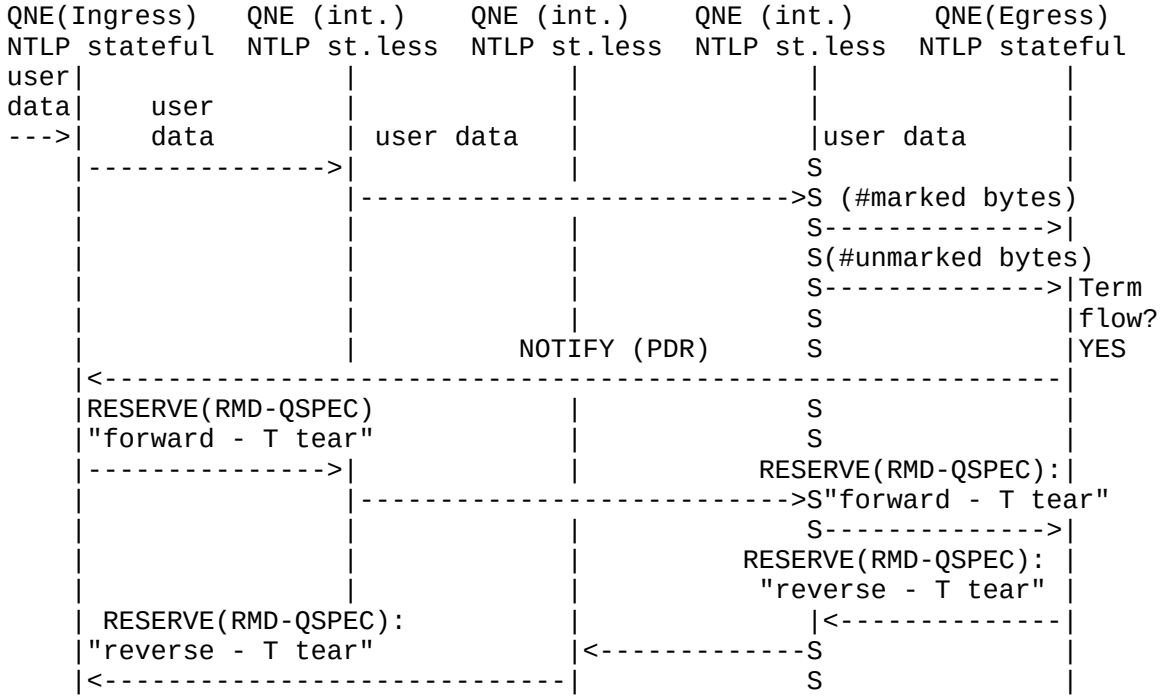

Figure 20: Intra-domain RMD severe congestion handling for bidirectional reservation (congestion on path QNE(Ingress) towards QNE(Egress))

Figure 20 shows the scenario in which the severely congested node is located in the "forward" path. The QNE Egress node has to generate an end-to-end NOTIFY (PDR) message. In this way, the QNE Ingress will be able to receive the (#marked and #unmarked) that were measured by the QNE Egress node on the congested "forward" path. Note that in this situation, it is assumed that the "reverse" path is not congested.

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 shows the scenario in which the severely congested node is located in the "reverse" path. Note that in this situation, it is assumed that the "forward" path is not congested. The main difference between this scenario and the scenario shown in Figure 20 is that no end-to-end NOTIFY (PDR) message has to be generated by the QNE Egress node.

This is because now the severe congestion occurs on the "reverse" path and the QNE Ingress node receives the (#marked and #unmarked) user data passing through the severely congested "reverse" path. 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.

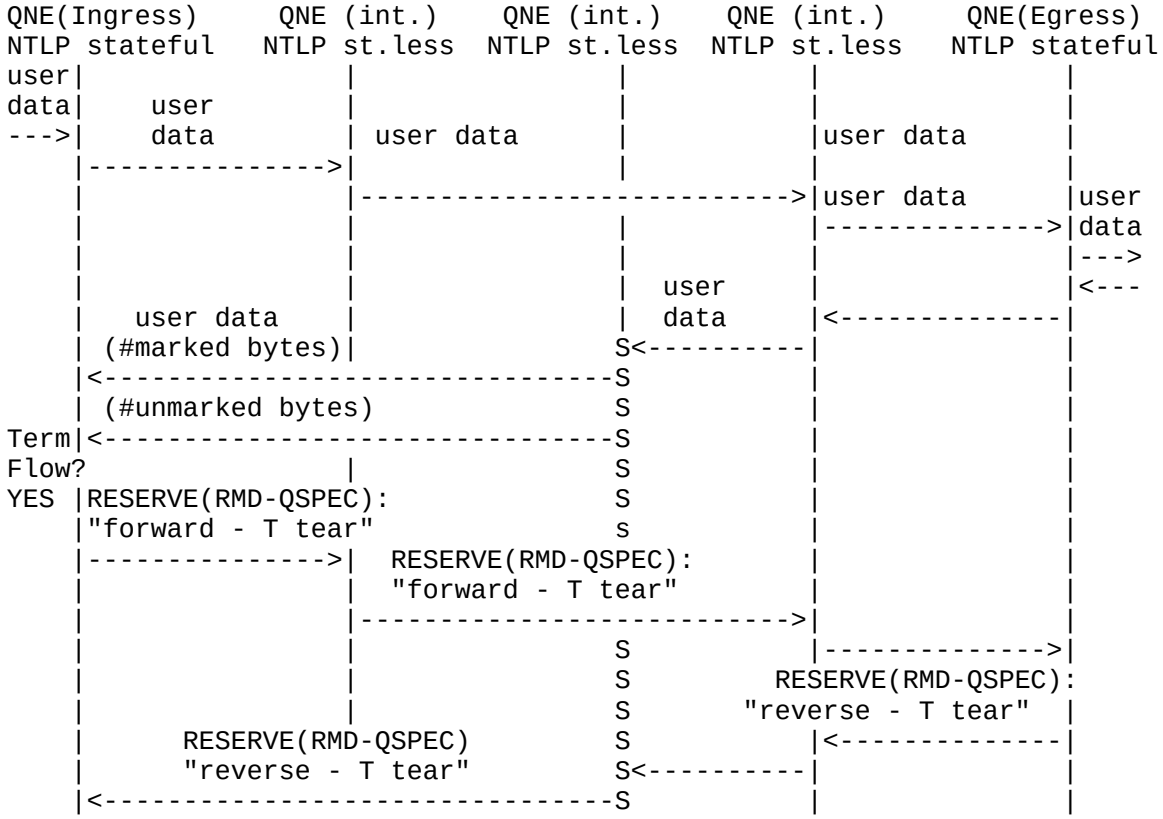

Figure 21: Intra-domain RMD severe congestion handling for bidirectional reservation (congestion on path QNE(Egress) towards QNE(Ingress))

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 RMD intra-domain bidirectional reservations are supported.

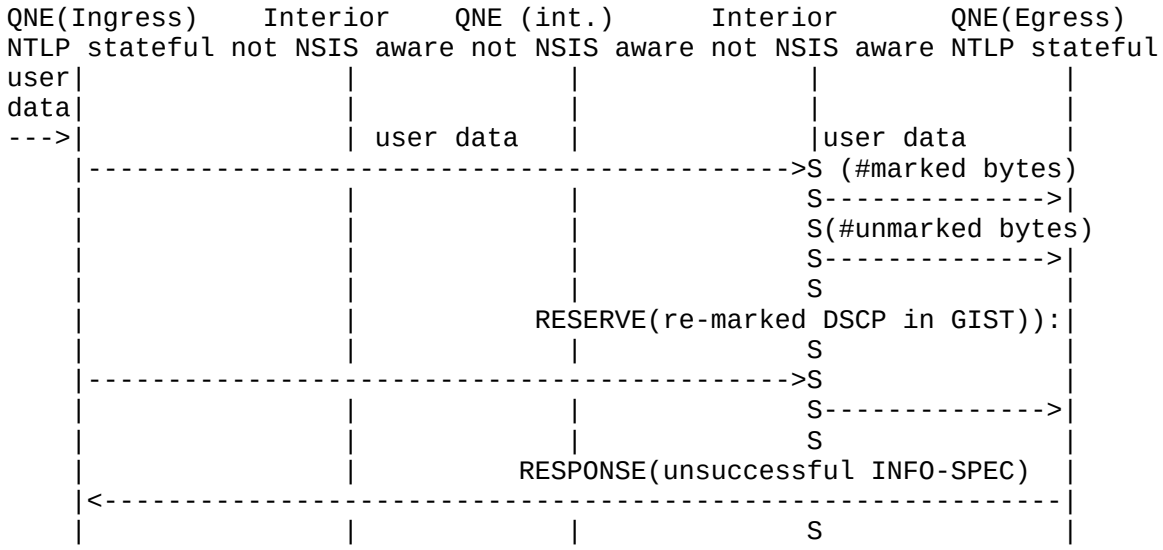

Figure 22: Intra-domain RMD congestion notification based on probing for bidirectional admission control (congestion on path from QNE(Ingress) towards QNE(Egress))

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 shows the scenario in which the severely congested node is located in the "forward" path. The functionality of providing admission control is the same as that described in Section 4.6.1.7, Figure 15.

Figure 23 shows the scenario in which the congested node is located in the "reverse" path. The probe RESERVE message sent in the "forward" direction will not be affected by the severely congested node, while the <DSCP> value in the IP header of any packet of the "reverse" direction flow and also of the GIST message that carries the probe RESERVE message sent in the "reverse" direction will be re-marked 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.

Note that the "not NSIS-aware" Interior nodes MUST be configured such that they can detect the congestion/severe congestion situations and re-mark packets in the same way as the Interior "NSIS-aware" nodes do.

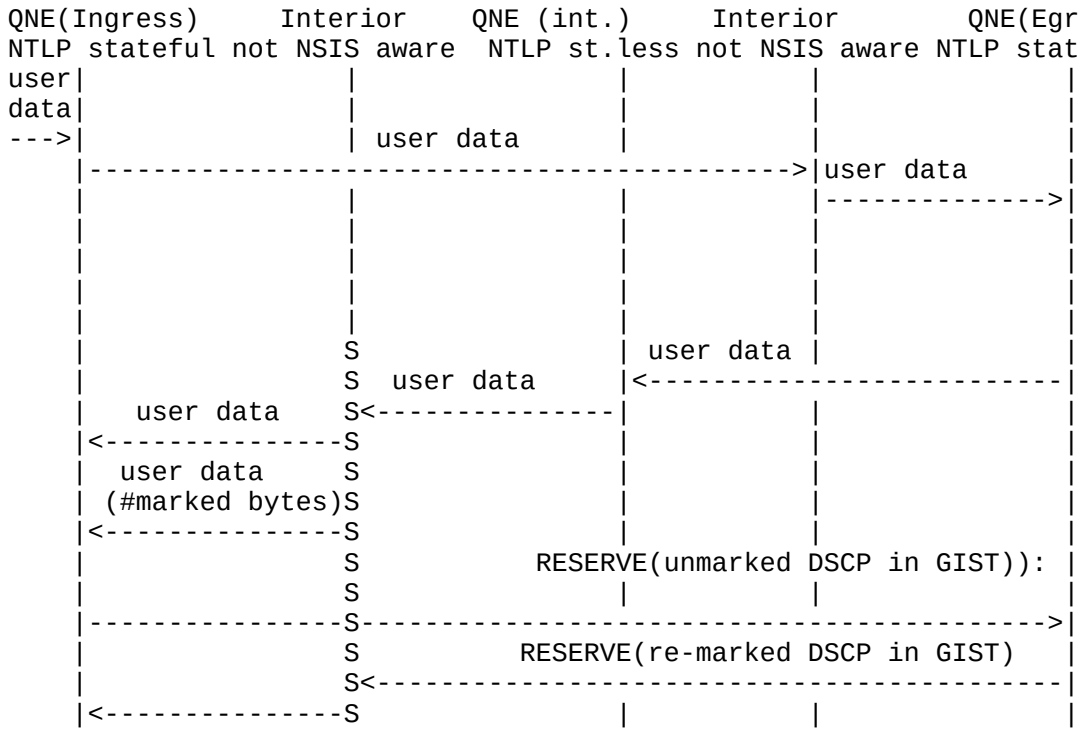

Figure 23: Intra-domain RMD congestion notification for bidirectional admission control (congestion on path QNE(Egress) towards QNE(Ingress))

4.7. Handling of Additional Errors

During the QSPEC processing, additional errors MAY occur. The way in which these additional errors are handled and notified is specified in [RFC5975] and [RFC5974].

5. Security Considerations

5.1. Introduction

A design goal of the RMD-QOSM protocol is to be "lightweight" in terms of the number of exchanged signaling message and the amount of state established at involved signaling nodes (with and without reduced-state operation). A side effect of this design decision is

to introduce second-class signaling nodes, namely QNE Interior nodes, that are restricted in their ability to perform QoS signaling actions. Only the QNE Ingress and the QNE Egress nodes are allowed to initiate certain signaling messages.

Moreover, RMD focuses on an intra-domain deployment only.

The above description has the following implications for security:

1) QNE Ingress and QNE Egress nodes require more security and fault protection than QNE Interior nodes because their uncontrolled behavior has larger implications for the overall stability of the network. QNE Ingress and QNE Egress nodes share a security association and utilize GIST security for protection of their signaling messages. Intra-domain signaling messages used for RMD signaling do not use GIST security, and therefore they do not store security associations.

2) The focus on intra-domain QoS signaling simplifies trust management and reduces overall complexity. See Section 2 of RFC 4081 for a more detailed discussion about the complete set of communication models available for end-to-end QoS signaling protocols. The security of RMD-QOSM does not depend on Interior nodes, and hence the cryptographic protection of intra-domain messages via GIST is not utilized.

It is important to highlight 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 end-to-end (E2E) signaling exchange, then the RESERVE message is created by a node outside the RMD domain and will subsequently travel further (e.g., to the data receiver). Such an exchange is shown in Figure 3. As such, an evaluation of an RMD's security always has to be seen as a combination of the two signaling sessions, (1) and (2) of Figure 24. Note that for the E2E message, such as the RESERVE and the RESPONSE message, a single "hop" refers to the communication between the QNE Ingress and the QNE Egress since QNE Interior nodes do not participate in the exchange.

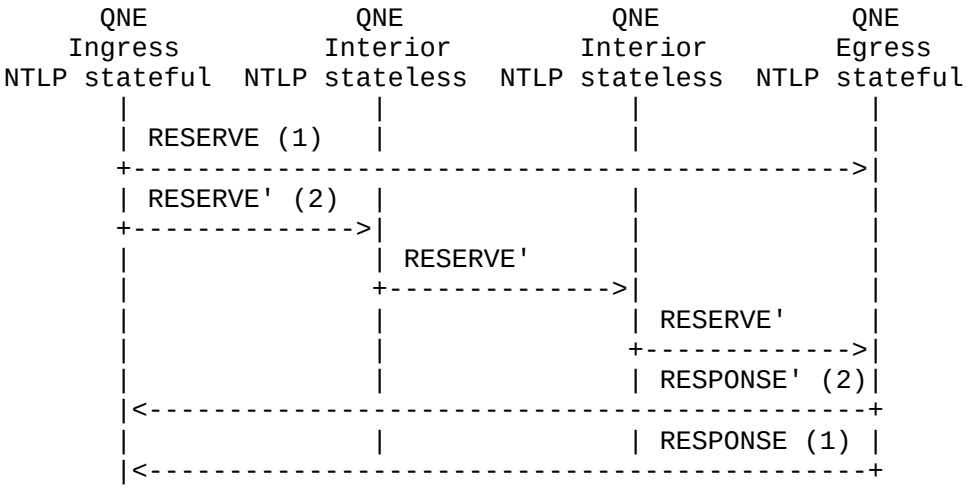

Figure 24: RMD message exchange

Authorizing quality-of-service reservations is accomplished using the Authentication, Authorization, and Accounting (AAA) framework and the functionality is inherited from the underlying NSIS QoS NSLP, see [RFC5974], and not described again in this document. As a technical solution mechanism, the Diameter QoS application [RFC5866] may be used. The end-to-end reservation request arriving at the Ingress node will trigger the authorization procedure with the backend AAA infrastructure. The end-to-end reservation is typically triggered by a human interaction with a software application, such as a voice-over-IP client when making a call. When authorization is successful then no further user initiated QoS authorization check is expected to be performed within the RMD domain for the intra-domain reservation.

5.2. Security Threats

In the RMD-QOSM, the Ingress node constructs both end-to-end and intra-domain signaling messages based on the end-to-end message initiated by the sender end node.

The Interior nodes within the RMD network ignore the end-to-end signaling message, but they process, modify, and forward the intra-domain signaling messages towards the Egress node. In the meantime, resource reservation states are installed, modified, or deleted at each Interior node along the data path according to the content of each intra-domain signaling message. The Edge nodes of an RMD network are critical components that require strong security protection.

Therefore, they act as security gateways for incoming and outgoing signaling messages. Moreover, a certain degree of trust has to be placed into Interior nodes within the RMD-QOSM network, such that these nodes can perform signaling message processing and take the necessary actions.

With the RMD-QOSM, we assume that the Ingress and the Egress nodes are not controlled by an adversary and the communication between the Ingress and the Egress nodes is secured using standard GIST security, (see Section 6 of [RFC5971]) mechanisms and experiences integrity, replay, and confidentiality protection.

Note that this only affects messages directly addressed by these two nodes and not any other message that needs to be processed by intermediaries. The <SESSION-ID> object of the end-to-end communication is visible, via GIST, to the Interior nodes. In order to define the security threats that are associated with the RMD-QOSM, we consider that an adversary that may be located inside the RMD domain and could drop, delay, duplicate, inject, or modify signaling packets.

Depending on the location of the adversary, we speak about an on-path adversary or an off-path adversary, see also RFC 4081 [RFC4081].

5.2.1. On-Path Adversary

The on-path adversary is a node, which supports RMD-QOSM and is able to observe RMD-QOSM signaling message exchanges.

1) Dropping signaling messages

An adversary could drop any signaling messages after receiving them. This will cause a failure of reservation request for new sessions or deletion of resource units (bandwidth) for ongoing sessions due to states timeout.

It may trigger the Ingress node to retransmit the lost signaling messages. In this scenario, the adversary drops selected signaling messages, for example, intra-domain reserve messages. In the RMD-QOSM, the retransmission mechanism can be provided at the Ingress node to make sure that signaling messages can reach the Egress node. However, the retransmissions triggered by the adversary dropping messages may cause certain problems. Therefore, disabling the use of retransmissions in the RMD-QOSM-aware network is recommended, see also Section 4.6.1.1.1.

2) Delaying Signaling Messages

Any signaling message could be delayed by an adversary. For example, if RESERVE' messages are delayed over the duration of the refresh period, then the resource units (bandwidth) reserved along the nodes for corresponding sessions will be removed. In this situation, the Ingress node does not receive the RESPONSE within a certain period, and considers that the signaling message has failed, which may cause a retransmission of the "failed" message. The Egress node may distinguish between the two messages, i.e., the delayed message and the retransmitted message, and it could get a proper response.

However, Interior nodes suffer from this retransmission and they may reserve twice the resource units (bandwidth) requested by the Ingress node.

3) Replaying Signaling Messages

An adversary may want to replay signaling messages. It first stores the received messages and decides when to replay these messages and at what rate (packets per second).

When the RESERVE' message carried an <RII> object, the Egress will reply with a RESPONSE' message towards the Ingress node. The Ingress node can then detect replays by comparing the value of <RII> in the RESPONSE' messages with the stored value.

4) Injecting Signaling Messages

Similar to the replay-attack scenario, the adversary may store a part of the information carried by signaling messages, for example, the <RSN> object. When the adversary injects signaling messages, it puts the stored information together with its own generated parameters (RMD-QSPEC <TMOD-1> parameter, <RII>, etc.) into the injected messages and then sends them out. Interior nodes will process these messages by default, reserve the requested resource units (bandwidth) and pass them to downstream nodes.

It may happen that the resource units (bandwidth) on the Interior nodes are exhausted if these injected messages consume too much bandwidth.

5) Modifying Signaling Messages

On-path adversaries are capable of modifying any part of the signaling message. For example, the adversary can modify the <M>, <S>, and <O> parameters of the RMD-QSPEC messages. The Egress node will then use the SESSION-ID and subsequently the <BOUND-SESSION-ID>

objects to refer to that flow to be terminated or set to lower priority. It is also possible for the adversary to modify the RMD-QSPEC <TMOD-1> parameter and/or <PHB Class> parameter, which could cause a modification of an amount of the requested resource units (bandwidth) changes.

5.2.2. Off-Path Adversary

In this case, the adversary is not located on-path and it does not participate in the exchange of RMD-QOSM signaling messages, and therefore is unable to eavesdrop signaling messages. Hence, the adversary does not know valid <RII>s, <RSN>s, and <SESSION-ID>s. Hence, the adversary has to generate new parameters and constructs new signaling messages. Since Interior nodes operate in reduced-state mode, injected signaling messages are treated as new once, which causes Interior nodes to allocate additional reservation state.

5.3. Security Requirements

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, must not interfere with QNE Interior nodes. Off-path nodes (off-path with regard to the path taken by a particular signaling message exchange) must 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.

5.4. Security Mechanisms

An important security mechanism that was built into RMD-QOSM was the ability to tie the end-to-end RESERVE and the RESERVE' messages together using the BOUND-SESSION-ID and to allow the Ingress node to match the RESERVE' with the RESPONSE' by using the <RII>. These mechanisms enable the Edge nodes to detect unexpected signaling messages.

We assume that the RESERVE/RESPONSE is sent with hop-by-hop channel security provided by GIST and protected between the QNE Ingress and the QNE Egress. GIST security mechanisms MUST be used to offer authentication, integrity, and replay protection. Furthermore, encryption MUST be used to prevent an adversary located along the path of the RESERVE message from learning information about the session that can later be used to inject a RESERVE' message.

The following messages need to be mapped to each other to make sure that the occurrence of one message is not without the other:

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' relate to each other. 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' was created by the QNE Ingress node since the intra-domain session, which sent the RESERVE', is bound to an end-to-end session via the <BOUND-SESSION-ID> value included in the intra-domain QoS-NSLP operational state maintained at the QNE Egress.

The RESERVE and the RESERVE' message are tied together using the BOUND-SESSION-ID(s) maintained by the intra-domain and end-to-end QoS-NSLP operational states maintained at the QNE Edges (see Sections 4.3.1, 4.3.2, and 4.3.3). 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 rerouting 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 and generate an error message.

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 <O> 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, maintained by the intra-domain QoS-NSLP operational state, to refer to a flow that might be terminated. The aspect of intermediate nodes initiating messages for severe congestion handling is for further study.

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'.

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.

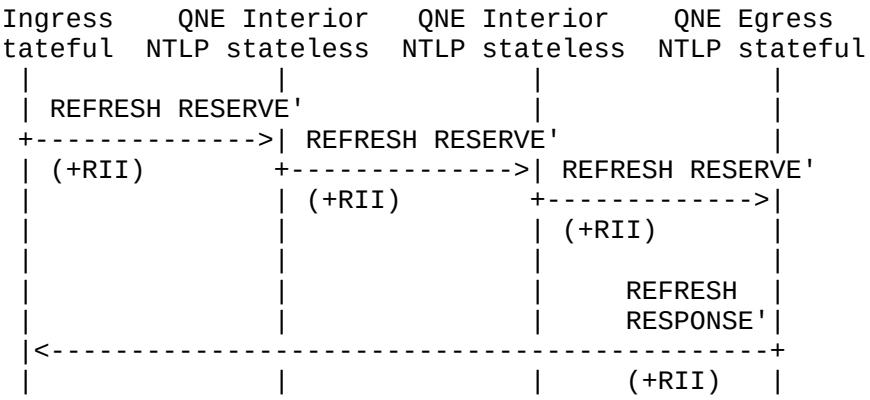

Figure 25: RMD REFRESH message exchange

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/rerouting case where the path through the network changes and the set of nodes that are along a path changes over time.

6. IANA Considerations

This section defines additional codepoint assignments in the QSPEC Parameter ID registry, in accordance with BCP 26 [RFC5226].

6.1. Assignment of QSPEC Parameter IDs

This document specifies the following QSPEC containers in the QSPEC Parameter ID registry created in [RFC5975]:

<PHR_Resource_Request> (Section 4.1.2 above, ID=17)

<PHR_Release_Request> (Section 4.1.2 above, ID=18)

<PHR_Refresh_Update> (Section 4.1.2 above, ID=19)

<PDR_Reservation_Request> (Section 4.1.3 above, ID=20)

<PDR_Refresh_Request> (Section 4.1.3 above, ID=21)

<PDR_Release_Request> (Section 4.1.3 above, ID=22)

<PDR_Reservation_Report> (Section 4.1.3 above, ID=23)

<PDR_Refresh_Report> (Section 4.1.3 above, ID=24)

<PDR_Release_Report> (Section 4.1.3 above, ID=25)

<PDR_Congestion_Report> (Section 4.1.3 above, ID=26)

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, T. Sealy, H. Chang, and J. de Waal.

8. References

8.1. Normative References

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

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

[RFC5971] Schulzrinne, H. and R. Hancock, "GIST: General Internet Signaling Transport", RFC 5971, October 2010.

[RFC5974] Manner, J., Karagiannis, G., and A. McDonald, "NSIS Signaling Layer Protocol (NSLP) for Quality-of-Service Signaling", RFC 5974, October 2010.

[RFC5975] Ash, G., Bader, A., Kappler C., and D. Oran, "QSPEC Template for the Quality-of-Service NSIS Signaling Layer Protocol (NSLP)", RFC 5975, October 2010.

8.2. Informative References

[AdCa03] Adler, M., Cai, J.-Y., Shapiro, J. K., Towsley, D., "Estimation of congestion price using probabilistic packet marking", Proc. IEEE INFOCOM, pp. 2068-2078, 2003.

[AnHa06] Lachlan L. H. Andrew and Stephen V. Hanly, "The Estimation Error of Adaptive Deterministic Packet Marking", 44th Annual Allerton Conference on Communication, Control and Computing, 2006.

[AtLi01] Athuraliya, S., Li, V. H., Low, S. H., Yin, Q., "REM: active queue management", IEEE Network, vol. 15, pp. 48-53, May/June 2001.

[Chan07] H. Chang, "Security support in RMD-QOSM", Masters thesis, University of Twente, 2007.

[CsTa05] Csaszar, A., Takacs, A., Szabo, R., Henk, T., "Resilient Reduced-State Resource Reservation", Journal of Communication and Networks, Vol. 7, No. 4, December 2005.

[DiKa08] Dimitrova, D., Karagiannis, G., de Boer, P.-T., "Severe congestion handling approaches in NSIS RMD domains with bi-directional reservations", Journal of Computer Communications, Elsevier, vol. 31, pp. 3153-3162, 2008.

[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.

[Part94] C. Partridge, Gigabit Networking, Addison Wesley Publishers (1994).

[RFC1633] Braden, R., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[RFC2215] Shenker, S. and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements", RFC 2215, September 1997.

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

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

[RFC2998] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J., and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, November 2000.

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

[RFC3726] Brunner, M., Ed., "Requirements for Signaling Protocols", RFC 3726, April 2004.

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

[RFC4127] Le Faucheur, F., Ed., "Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering", RFC 4127, June 2005.

[RFC4081] Tschofenig, H. and D. Kroeselberg, "Security Threats for Next Steps in Signaling (NSIS)", RFC 4081, June 2005.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5866] Sun, D., Ed., McCann, P., Tschofenig, H., Tsou, T., Doria, A., and G. Zorn, Ed., "Diameter Quality-of-Service Application", RFC 5866, May 2010.

[RFC5978] Manner, J., Bless, R., Loughney, J., and E. Davies, Ed., "Using and Extending the NSIS Protocol Family", RFC 5978, October 2010.

[RMD1] Westberg, L., et al., "Resource Management in Diffserv (RMD): A Functionality and Performance Behavior Overview", IFIP PfHSN 2002.

[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.

[TaCh99] P. P. Tang, T-Y Charles Tai, "Network Traffic Characterization Using Token Bucket Model", IEEE Infocom 1999, The Conference on Computer Communications, no. 1, March 1999, pp. 51-62.

[ThCo04] Thommes, R. W., Coates, M. J., "Deterministic packet marking for congestion packet estimation" Proc. IEEE Infocom, 2004.

Appendix A. Examples

A.1. Example of a Re-Marking Operation during Severe Congestion in the Interior Nodes

This appendix describes an example of a re-marking operation during severe congestion in the Interior nodes.

Per supported PHB, the Interior node can support the operation states depicted in Figure 26, when the per-flow congestion notification based on probing signaling scheme is used in combination with this severe congestion type. Figure 27 depicts the same functionality when the per-flow congestion notification based on probing scheme is not used in combination with the severe congestion scheme. The description given in this and the following appendices, focuses on the situation where: (1) the "notified DSCP" marking is used in congestion notification state, and (2) the "encoded DSCP" and "affected DSCP" markings are used in severe congestion state. In this case, the "notified DSCP" marking is used during the congestion notification state to mark all packets passing through an Interior node that operates in the congestion notification state. In this way, and in combination with probing, a flow-based ECMP solution can be provided for the congestion notification state. The "encoded DSCP" marking is used to encode and signal the excess rate, measured at Interior nodes, to the Egress nodes. The "affected DSCP" marking is used to mark all packets that are passing through a severe congested node and are not "encoded DSCP" marked.

Another possible situation could be derived in which both congestion notification and severe congestion state use the "encoded DSCP" marking, without using the "notified DSCP" marking. The "affected DSCP" marking is used to mark all packets that pass through an Interior node that is in severe congestion state and are not "encoded DSCP" marked. In addition, the probe packet that is carried by an intra-domain RESERVE message and pass through Interior nodes SHOULD be "encoded DSCP" marked if the Interior node is in congestion notification or severe congestion states. Otherwise, the probe packet will remain unmarked. In this way, an ECMP solution can be provided for both congestion notification and severe congestion states. The"encoded DSCP" packets signal an excess rate that is not only associated with Interior nodes that are in severe congestion state, but also with Interior nodes that are in congestion notification state. The algorithm at the Interior node is similar to the algorithm described in the following appendix sections. However, this method is not described in detail in this example.

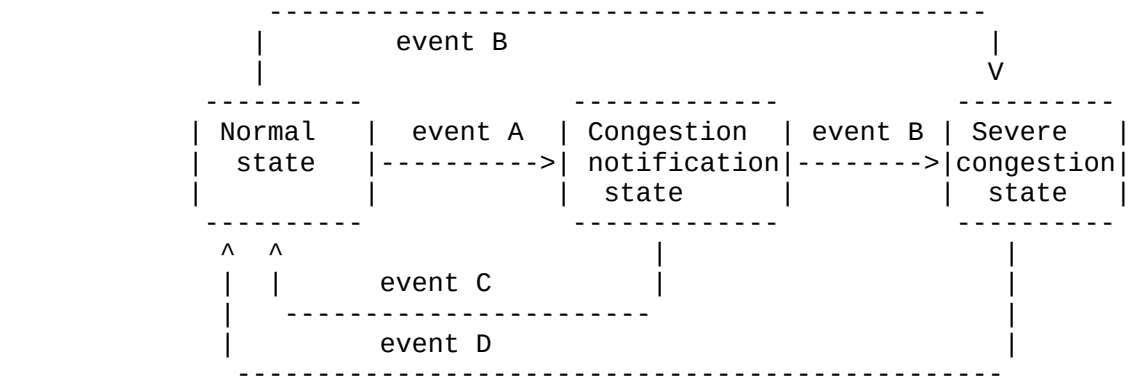

Figure 26: States of operation, severe congestion combined with congestion notification based on probing

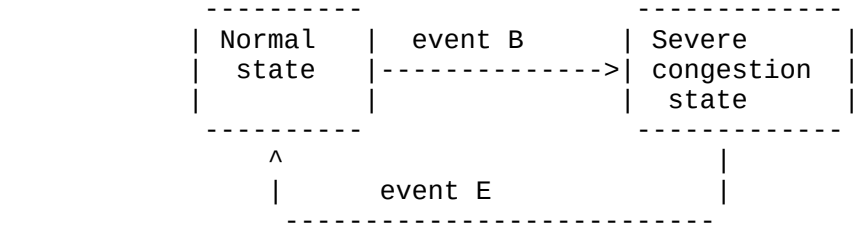

Figure 27: States of operation, severe congestion without congestion notification based on probing

The terms used in Figures 26 and 27 are:

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

Severe congestion state: represents the state in which the Interior node is severely congested related to a certain PHB. It is important to emphasize that one of the targets of the severe congestion state solution to change the severe congestion state behavior directly to the normal state.

Congestion notification: state in which 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 and lower than the "severe congestion detection". This threshold is used by the congestion notification based on probing scheme, see Sections 4.6.1.7 and 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 or equal to the "congestion notification detection" threshold.

event D: this event occurs when the incoming PHB rate is lower than or equal to the "severe_congestion_restoration" threshold. It is important to emphasize that this even supports one of the targets of the severe congestion state solution to change the severe congestion state behavior directly to the normal state.

event E: this event occurs when the incoming PHB rate is lower than or equal to 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, that is:

"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 re-marked bytes received by the severe congested node, denote this number as total_received_bytes. Note that there are situations in which more than one Interior node in the same path become severely congested. Therefore, any Interior node located behind a severely congested node MAY receive marked bytes.

When the "severe congestion detection" threshold per PHB is set equal to the maximum capacity allocated to one PHB used by the RMD-QOSM, it means that if the maximum capacity associated to a PHB is fully utilized and a packet belonging to this PHB arrives, then it is assumed that the Interior node will not forward this packet downstream.

In other words, this packet will either be dropped or set to another PHB. Furthermore, this also means that after the severe congestion situation is solved, then the ongoing flows will be able to send their associated packets up to a total rate equal to the maximum capacity associated with the PHB. Therefore, when more than one Interior node located on the same path will be severely congested and when the Interior node receives "encoded DSCP" marked packets, it means that an Interior node located upstream is also severely congested.

When the "severe congestion detection" threshold per PHB is set equal to the maximum capacity allocated to one PHB, then this Interior node MUST forward the "encoded DSCP" marked packets and it SHOULD NOT consider these packets during its local re-marking process. In other words, the Egress should see the excess rates encoded by the different severely congested Interior nodes as independent, and therefore, these independent excess rates will be added.

When the "severe congestion detection" threshold per PHB is not set equal to the maximum capacity allocated to one PHB, this means that after the severe congestion situation is solved, the ongoing flows will not be able to send their associated packets up to a total rate equal to the maximum capacity associated with the PHB, but only up to the "severe_congestion_threshold". When more than one Interior node located on the same communication path is severely congested and when one of these Interior node receives "encoded_DSCP" marked packets, this Interior node SHOULD NOT mark unmarked, i.e., either "original DSCP" or "affected DSCP" or "notified DSCP" encoded packets, up to a rate equal to the difference between the maximum PHB capacity and the "severe congestion threshold", when the incoming "encoded DSCP" marked packets are already able to signal this difference. In this case, the "severe congestion threshold" SHOULD be configured in all Interior nodes, which are located in the RMD domain, and equal to:

```
"severe_congestion_threshold" =
   Maximum PHB capacity - threshold_offset_rate
```

The threshold_offset_rate represents rate and SHOULD have the same value in all Interior nodes.

* 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)

To provide a reliable estimation of the encoded information, several techniques can be used; see [AtLi01], [AdCa03], [ThCo04], and [AnHa06]. Note that since marking is done in Interior nodes, the decisions are made at Egress nodes, and the termination of flows is 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 severely 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 severely congested Interior node, an additional trip time from the Ingress node to the severely 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 were 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 overreaction 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 interval, 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 re-marked to satisfy the signaled overload rate: signaled_remarked_bytes, are calculated using the following pseudocode:

```
IF severe_congestion_threshold <> Maximum PHB capacity
THEN
 {
  IF (incoming_encoded-DSCP_rate <> 0) AND
     (incoming_encoded-DSCP_rate =< termination_offset_rate)
  THEN
     { signaled_remarked_bytes =
      = ((signaled_overload_rate - incoming_encoded-DSCP_rate)*T)/N
     }
  ELSE IF (incoming_encoded-DSCP_rate > termination_offset_rate)
  THEN signaled_remarked_bytes =
      = ((signaled_overload_rate - termination_offset_rate)*T)/N
  ELSE IF (incoming_encoded-DSCP_rate =0)
  THEN signaled_remarked_bytes =
      = signaled_overload_rate*T/N
  }
 ELSE signaled_remarked_bytes =  signaled_overload_rate *T/N

 Where the incoming "encoded DSCP" rate is calculated as follows:

 incoming_encoded-DSCP_rate =
  = (received number of "encoded_DSCP" during T) * N)/T;
```

The signal_remarked_bytes also represents the number of the outgoing packets (after the dropping stage) that MUST be re-marked, 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 can be calculated and represented. N can be equal to or higher than 1.

Note that when incoming re-marked 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 used by a node that is congested due to a severe congestion situation, then all the outgoing packets that are not marked (i.e., by using the "encoded DSCP") have to be re-marked using the "affected DSCP" marking.

The "encoded DSCP" and the "affected DSCP" marked packets (when applied in the whole RMD domain) are propagated to the QNE Edge nodes.

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 re-marking of the same PHB is used (see Section 4.6.1.7). This additional DSCP is denoted in this document as "notified DSCP". When an Interior node operates in the severe congested state (see Figure 27), and receives "notified DSCP" packets, these packets are considered to be unmarked packets (but not "affected DSCP" packets). This means that during severe congestion, also the "notified DSCP" packets can be re-marked and encoded as either "encoded DSCP" or "affected DSCP" packets.

A.2. Example of a Detailed Severe Congestion Operation in the Egress Nodes

This appendix describes an 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 Appendix A.1. The definition of the events, see below, is however different than the definition of the events given in Figures 26 and 27:

* event A: when the Egress receives a predefined rate of "notified DSCP" marked bytes/packets, event A is activated (see Sections 4.6.1.7 and A.4). The predefined rate of "notified DSCP" marked bytes is denoted as the congestion notification detection threshold. Note this congestion notification detection threshold can also be zero, meaning that the event A is activated when the Egress node, during an interval T, receives at least one "notified DSCP" packet.

* 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).

* event C: this event occurs when the rate of incoming "notified DSCP" packets decreases below the congestion notification detection threshold. In the situation that the congestion notification detection threshold is zero, this will mean that event C is activated when the Egress node, during an interval T, does not receive any "notified DSCP" marked packets.

* event D: this event occurs when the Egress, during an interval T, does not receive packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain). Note that when "notified DSCP" is applied in the whole RMD domain for the support of congestion notification, this event could cause the following change in operation state.

  When the Egress, during an interval T, does not receive (1) packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain) and (2) it does NOT receive "notified DSCP" marked packets, the change in the operation state occurs from the severe congestion state to normal state.

  When the Egress, during an interval T, does not receive (1) packets marked as either "encoded DSCP" or "affected DSCP" (when "affected DSCP" is applied in the whole RMD domain) and (2) it does receive "notified DSCP" marked packets, the change in the operation state occurs from the severe congestion state to the congestion notification state.

* event E: this event occurs when the Egress, during an interval T, 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.

The Edge nodes are able to support severe congestion handling by: (1) identifying which flows were affected by the severe congestion and (2) selecting and terminating some of these flows such that the quality of service of the remaining flows is recovered.

The "encoded DSCP" and the "affected DSCP" marked packets (when applied in the whole RMD domain) are received by the QNE Edge node.

The QNE Edge nodes keep per-flow state and therefore they can translate the calculated bandwidth to be terminated, to number of flows. The QNE Egress node records the excess rate and the identity of all the flows, arriving at the QNE Egress node, with "encoded DSCP" and with "affected DSCP" (when applied in the whole RMD domain); only these flows, which are the ones passing through the severely congested Interior node(s), are candidates for termination. The excess rate is calculated by measuring the rate of all the "encoded DSCP" data packets that arrive at the QNE Egress node. The measured excess rate is converted by the Egress node, by multiplying it by the factor N, which was used by the QNE Interior node(s) to encode the overload level.

When different priority flows are supported, all the low priority flows that arrived at the Egress node are terminated first. Next, all the medium priority flows are stopped and finally, if necessary, even high priority flows are chosen. Within a priority class both "encoded DSCP" and "affected DSCP" are considered before the mechanism moves to higher priority class. Finally, for each flow that has to be terminated the Egress node, sends a NOTIFY message to the Ingress node, which stops the flow.

Below, this algorithm is described in detail.

First, when the Egress operates in the severe congestion state, the total amount of re-marked 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 reservation aggregate. This bandwidth represents the severely 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 "encoded DSCP" marked bytes that arrive at the Egress, during one measurement interval T, N is defined as in Sections 4.6.1.6.2.1 and A.1. The term denoted as terminated_bandwidth is a temporal variable representing the total bandwidth that has to be terminated, belonging to the same PHB traffic class. The terminate_flow_bandwidth (priority_class) is the total 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 QNE Egress node records the identity of the QNE Ingress node that forwarded each flow, the total_congested_bandwidth and the identity of all the flows, arriving at the QNE Egress node, with "encoded DSCP" and "affected DSCP" (when applied in whole RMD domain). This ensures that only these flows, which are the ones passing through the severely overloaded QNE Interior node(s), are candidates for termination. The selection of the flows to be terminated is described in the pseudocode that is given below, which is realized by the function denoted below as calculate_terminate_flows().

The calculate_terminate_flows() function uses the <terminate_bandwidth_class> value and translates this bandwidth value to number of flows that have to be terminated. Only the "encoded DSCP" flows and "affected DSCP" (when applied in whole RMD domain) flows, which are the ones passing through the severely overloaded Interior node(s), are candidates for termination.

After the flows to be terminated are selected, the <sum_bandwidth_terminate(priority_class)> value is calculated that is the sum of the bandwidth associated with the flows, belonging to a certain priority class, which will certainly 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 approximately 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 reservation aggregates, then the above algorithm is performed for each Ingress/Egress pair reservation aggregate.

Finally, for each flow that has to be terminated, the QNE Egress node sends a NOTIFY message to the QNE Ingress node to terminate the flow.

A.3. Example of a Detailed Re-Marking Admission Control (Congestion Notification) Operation in Interior Nodes

This appendix describes an example of a detailed re-marking admission control (congestion notification) operation in Interior nodes. The predefined congestion notification threshold, see Appendix A.1, is set according to, and usually less than, an engineered bandwidth limitation, i.e., admission threshold, e.g., based on a 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, which 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, data packets are also 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 A.1):

* using different <DSCP> values (re-marked <DSCP> values). The re-marked 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 Appendix

A.1, then the node SHOULD re-mark all the unmarked bytes passing through the node using the "notified DSCP". Note that this method can only be applied if all nodes in the RMD domain use the "notified" DSCP marking. In this way, probe packets that will pass through the Interior node that operates in congestion notification state are also encoded using the "notified DSCP" marking.

* Using the "encoded DSCP" marking for congestion notification and severe congestion. This method is not described in detail in this example appendix.

A.4. Example of a Detailed Admission Control (Congestion Notification) Operation in Egress Nodes

This appendix describes an 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 maintains the Ingress/Egress pair aggregate. When the operation state of the Ingress/Egress pair aggregate is the "congestion notification", see Appendix A.2, then the implementation of the algorithm depends on how the congestion notification situation is notified to the Egress. As mentioned in Appendix A.3, 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 the "encoded DSCP" bytes.

Below only the detail description of the first method is given.

The incoming congestion_rate can be then calculated as follows:

```
incoming_congestion_rate = 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. Note that the pre-congestion notification threshold can be set to "0". In this

case, the Egress node will operate in congestion notification state at the moment that it receives at least one "notified DSCP" encoded packet.

When the Egress node operates in "congestion notification" state and if the end-to-end RESERVE (probe) arrives at the Egress, then this request SHOULD be rejected. Note that this happens only when the probe packet is either "notified DSCP" or "encoded DSCP" marked. In this way, it is ensured that the end-to-end RESERVE (probe) packet passed through the node that is congested. This feature is very useful when ECMP-based routing is used to detect only flows that are passing through the congested router.

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 message is unmarked. Otherwise (if the packet is either "notified DSCP" or "encoded DSCP" marked), it is rejected.

A.5. Example of Selecting Bidirectional Flows for Termination during Severe Congestion

This appendix describes an example of selecting bidirectional flows for termination during severe congestion.

When a severe congestion occurs, e.g., in the forward path, and when the algorithm terminates flows to solve the severe congestion in the 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.

A.6. Example of a Severe Congestion Solution for Bidirectional Flows Congested Simultaneously on Forward and Reverse Paths

This appendix describes an example of a severe congestion solution for bidirectional flows congested simultaneously on forward and reverse paths.

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 in forward and reverse directions, which MAY affect the same bidirectional flows.

When the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states, the steps can be the following, see Figure A.3. Consider that the Egress node selects a number of bidirectional flows to be terminated. In this case, the Egress will send, for each bidirectional flow, 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 Figures 26 and 27), then the Ingress operates in the following way:

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

* 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". This bandwidth can be calculated by the sum of the bandwidth values associated with all the end-to-end sessions that received a (severe congestion) NOTIFY message.

* Furthermore, using the received marked packets (from the reverse path) the Ingress will calculate, using the algorithm used by an Egress and described in Appendix A.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".

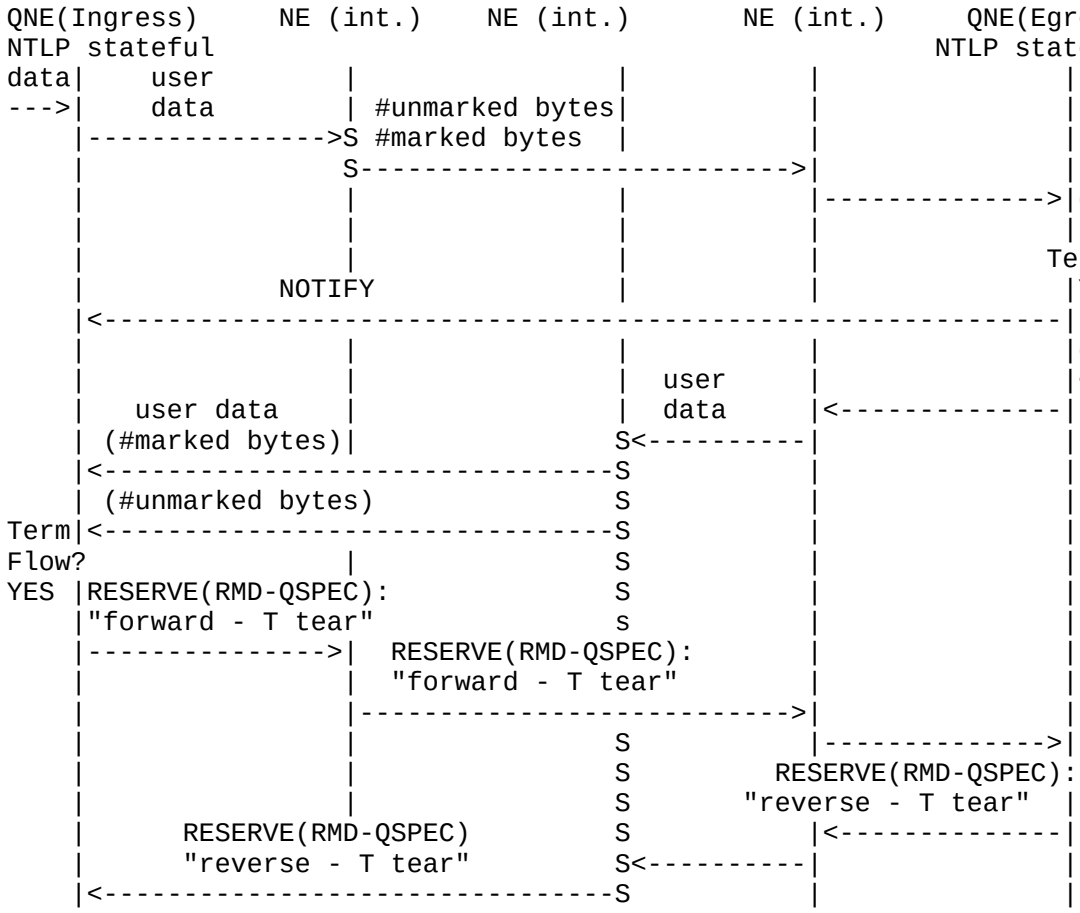

Figure 28: Intra-domain RMD severe congestion handling for bidirectional reservation (congestion in both forward and reverse direction)

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 experienced a severe congestion in the forward path and received a (severe congestion) NOTIFY message.

* If possible, the Ingress SHOULD terminate unidirectional flows that use the same Egress-Ingress reverse direction communication path to satisfy the release of a total bandwidth up equal to the "additional_reverse_bandwidth", see Appendix A.5.

* If the number of REQUIRED unidirectional flows (to satisfy the above issue) is not available, then a number of bidirectional 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 bandwidth equal up to the "additional_reverse_bandwidth". Note that using the guidelines given in Appendix A.5, 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.

When the QNE Edges maintain aggregated intra-domain QoS-NSLP operational states, the steps can be the following.

* The Egress calculates the bandwidth to be terminated using the same method as described in Section 4.6.1.6.2.2. The Egress includes this bandwidth value in a <PDR Bandwidth> within a "PDR_Congestion_Report" container that is carried by the end-to-end NOTIFY message.

* The Ingress receives the NOTIFY message and reads the <PDR Bandwidth> value included in the "PDR_Congestion_Report" container. Note that this value is denoted as "notify_reverse_bandwidth" in the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states, but is calculated differently. The variables "marked_reverse_bandwidth" and "additional_reverse_bandwidth" are calculated using the same steps as explained for the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP states.

* Regarding the termination of flows that use the same Egress-Ingress reverse direction communication path, the Ingress can follow the same procedures as the situation that the QNE Edges maintain per-flow intra-domain QoS-NSLP operational states.

The RMD-aggregated (reduced-state) reservations maintained by the Interior nodes, can be reduced in the "forward" and "reverse" directions by using the procedure described in Section 4.6.2.3 and including in the <Peak Data Rate-1 (p)> value of the local RMD-QSPEC <TMOD-1> parameter of the RMD-QOSM <QoS Desired> field carried by the forward intra-domain RESERVE

the value equal to <notify_reverse_bandwidth> and by including the <additional_reverse_bandwidth> value in the <PDR Bandwidth> parameter within the "PDR_Release_Request" container that is carried by the same intra-domain RESERVE message.

A.7. Example of Preemption Handling during Admission Control

This appendix describes an example of how preemption handling is supported during admission control.

This section describes the mechanism that can be supported by the QNE Ingress, QNE Interior, and QNE Egress nodes to satisfy preemption during the admission control process.

This mechanism uses the preemption building blocks specified in [RFC5974].

A.7.1. Preemption Handling in QNE Ingress Nodes

If a QNE Ingress receives a RESERVE for a session that causes other session(s) to be preempted, for each of these to-be-preempted sessions, then the QNE Ingress follows the following steps:

Step_1:

The QNE Ingress MUST send a tearing RESERVE downstream and add a BOUND-SESSION-ID, with <Binding_Code> value equal to "Indicated session caused preemption" that indicates the SESSION-ID of the session that caused the preemption. Furthermore, an <INFO-SPEC> object with error code value equal to "Reservation preempted" has to be included in each of these tearing RESERVE messages.

The selection of which flows have to be preempted can be based on predefined policies. For example, this selection process can be based on the MRI associated with the high and low priority sessions. In particular, the QNE Ingress can select low(er) priority session(s) where their MRI is "close" (especially the target IP) to the one associated with the higher priority session. This means that typically the high priority session and the to-be-preempted lower priority sessions are following the same communication path and are passing through the same QNE Egress node.

Furthermore, the amount of lower priority sessions that have to be preempted per each high priority session, has to be such that the requested resources by the higher priority session SHOULD be lower or equal than the sum of the reserved resources associated with the lower priority sessions that have to be preempted.

Step_2:

For each of the sent tearing RESERVE(s) the QNE Ingress will send a NOTIFY message with an <INFO-SPEC> object with error code value equal to "Reservation preempted" towards the QNI.

Step_3:

After sending the preempted (tearing) RESERVE(s), the Ingress QNE will send the (reserving) RESERVE, which caused the preemption, downstream towards the QNE Egress.

A.7.2. Preemption Handling in QNE Interior Nodes

The QNE Interior upon receiving the first (tearing) RESERVE that carries the <BOUND-SESSION-ID> object with <Binding_Code> value equal to "Indicated session caused preemption" and an <INFO-SPEC> object with error code value equal to "Reservation preempted" it considers that this session has to be preempted.

In this case, the QNE Interior creates a so-called "preemption state", which is identified by the SESSION-ID carried in the preemption-related <BOUND-SESSION-ID> object. Furthermore, this "preemption state" will include the SESSION-ID of the session associated with the (tearing) RESERVE. Subsequently, if additional tearing RESERVE(s) are arriving including the same values of BOUND-SESSION-ID and <INFO-SPEC> objects, then the associated SESSION-IDs of these (tearing) RESERVE message will be included in the already created "preemption state". The QNE will then set a timer, with a value that is high enough to ensure that it will not expire before the (reserving) RESERVE arrives.

Note that when the "preemption state" timer expires, the bandwidth associated with the preempted session(s) will have to be released, following a normal RMD-QOSM bandwidth release procedure. If the QNE Interior node will not receive all the to-be-preempted (tearing) RESERVE messages sent by the QNE Ingress before their associated (reserving) RESERVE message arrives, then the (reserving) RESERVE message will not reserve any resources and this message will be "M" marked (see Section 4.6.1.2). Note that this situation is not a typical situation. Typically, this situation can only occur when at least one of (tearing) the RESERVE messages is dropped due to an error condition.

Otherwise, if the QNE Interior receives all the to-be-preempted (tearing) RESERVE messages sent by the QNE Ingress, then the QNE Interior will remove the pending resources, and make the new reservation using normal RMD-QOSM bandwidth release and reservation procedures.

A.7.3. Preemption Handling in QNE Egress Nodes

Similar to the QNE Interior operation, the QNE Egress, upon receiving the first (tearing) RESERVE that carries the <BOUND-SESSION-ID> object with the <Binding_Code> value equal to "Indicated session caused preemption" and an <INFO-SPEC> object with error code value equal to "Reservation preempted", it considers that this session has to be preempted. Similar to the QNE Interior operation the QNE Egress creates a so called "preemption state", which is identified by the SESSION-ID carried in the preemption-related <BOUND-SESSION-ID> object. This "preemption state" will store the same type of information and use the same timer value as specified in Appendix A.7.2.

Subsequently, if additional tearing RESERVE(s) are arriving including the same values of BOUND-SESSION-ID and <INFO-SPEC> objects, then the associated SESSION-IDs of these (tearing) RESERVE message will be included in the already created "preemption state".

If the (reserving) RESERVE message sent by the QNE Ingress node arrived and is not "M" marked, and if all the to-be-preempted (tearing) RESERVE messages arrived, then the QNE Egress will remove the pending resources and make the new reservation using normal RMD-QOSM procedures.

If the QNE Egress receives an "M" marked RESERVE message, then the QNE Egress will use the normal partial RMD-QOSM procedure to release the partial reserved resources associated with the "M" marked RESERVE (see Section 4.6.1.2).

If the QNE Egress will not receive all the to-be-preempted (tearing) RESERVE messages sent by the QNE Ingress before their associated and not "M" marked (reserving) RESERVE message arrives, then the following steps can be followed:

* If the QNE Egress uses an end-to-end QOSM that supports the preemption handling, then the QNE Egress has to calculate and select new lower priority sessions that have to be terminated. How the preempted sessions are selected and signaled to the downstream QNEs is similar to the operation specified in Appendix A.7.1.

* If the QNE Egress does not use an end-to-end QOSM that supports the preemption handling, then the QNE Egress has to reject the requesting (reserving) RESERVE message associated with the high priority session (see Section 4.6.1.2).

Note that typically, the situation in which the QNE Egress does not receive all the to-be-preempted (tearing) RESERVE messages sent by the QNE Ingress can only occur when at least one of the (tearing) RESERVE messages are dropped due to an error condition.

A.8. Example of a Retransmission Procedure within the RMD Domain

This appendix describes an example of a retransmission procedure that can be used in the RMD domain.

If the retransmission of intra-domain RESERVE messages within the RMD domain is not disallowed, then all the QNE Interior nodes SHOULD use the functionality described in this section.

In this situation, we enable QNE Interior nodes to maintain a replay cache in which each entry contains the <RSN>, <SESSION-ID> (available via GIST), <REFRESH-PERIOD> (available via the QoS NSLP [RFC5974]), and the last received "PHR Container" <Parameter ID> carried by the RMD-QSPEC for each session [RFC5975]. Thus, this solution uses information carried by <QoS-NSLP> objects [RFC5974] and parameters carried by the RMD-QSPEC "PHR Container". The following phases can be distinguished:

Phase 1: Create Replay Cache Entry

When an Interior node receives an intra-domain RESERVE message and its cache is empty or there is no matching entry, it reads the <Parameter ID> field of the "PHR Container" of the received message. If the <Parameter ID> is a PHR_RESOURCE_REQUEST, which indicates that the intra-domain RESERVE message is a reservation request, then the QNE Interior node creates a new entry in the cache and copies the <RSN>, <SESSION-ID> and <Parameter ID> to the entry and sets the <REFRESH-PERIOD>.

By using the information stored in the list, the Interior node verifies whether or not the received intra-domain RESERVE message is sent by an adversary. For example, if the <SESSION-ID> and <RSN> of a received intra-domain RESERVE message match the values stored in the list then the Interior node checks the <Parameter ID> part.

If the <Parameter ID> is different, then:

Situation D1: <Parameter ID> in its own list is PHR_RESOURCE_REQUEST, and <Parameter ID> in the message is PHR_REFRESH_UPDATE;

Situation D2: <Parameter ID> in its own list is PHR_RESOURCE_REQUEST or PHR_REFRESH_UPDATE, and <Parameter ID> in the message is PHR_RELEASE_REQUEST;

Situation D3: <Parameter ID> in its own list is PHR_REFRESH_UPDATE, and <Parameter ID> in the message is PHR_RESOURCE_REQUEST;

For Situation D1, the QNE Interior node processes this message by RMD-QOSM default operation, reserves bandwidth, updates the entry, and passes the message to downstream nodes. For Situation D2, the QNE Interior node processes this message by RMD-QOSM default operation, releases bandwidth, deletes all entries associated with the session and passes the message to downstream nodes. For situation D3, the QNE Interior node does not use/process the local RMD-QSPEC <TMOD-1> parameter carried by the received intra-domain RESERVE message. Furthermore, the <K> flag in the "PHR Container" has to be set such that the local RMD-QSPEC <TMOD-1> parameter carried by the intra-domain RESERVE message is not processed/used by a QNE Interior node.

If the <Parameter ID> is the same, then:

Situation S1: <Parameter ID> is equal to PHR_RESOURCE_REQUEST;
Situation S2: <Parameter ID> is equal to PHR_REFRESH_UPDATE;

For situation S1, the QNE Interior node does not process the intra-domain RESERVE message, but it just passes it to downstream nodes, because it might have been retransmitted by the QNE Ingress node. For situation S2, the QNE Interior node processes the first incoming intra-domain (refresh) RESERVE message within a refresh period and updates the entry and forwards it to the downstream nodes.

If only <Session-ID> is matched to the list, then the QNE Interior node checks the <RSN>. Here also two situations can be distinguished:

If a rerouting takes place (see Section 5.2.5.2 in [RFC5974]), the <RSN> in the message will be equal to either <RSN + 2> in the stored list if it is not a tearing RESERVE or <RSN -1> in the stored list if it is a tearing RESERVE:

The QNE Interior node will check the <Parameter ID> part;

If the <RSN> in the message is equal to <RSN + 2> in the stored list and the <Parameter ID> is a PHR_RESOURCE_REQUEST or PHR_REFRESH_UPDATE, then the received intra-domain RESERVE message has to be interpreted and processed as a typical (non-tearing) RESERVE message, which is caused by rerouting, see Section 5.2.5.2 in [RFC5974].

If the <RSN> in the message is equal to <RSN-1> in the stored list and the <Parameter ID> is a PHR_RELEASE_REQUEST, then the received intra-domain RESERVE message has to be interpreted and processed as a typical (tearing) RESERVE message, which is caused by rerouting (see Section 5.2.5.2 in [RFC5974]).

If other situations occur than the ones described above, then the QNE Interior node does not use/process the local RMD-QSPEC <TMOD-1> parameter carried by the received intra-domain RESERVE message. Furthermore, the <K> parameter has to be set, see above.

Phase 2: Update Replay Cache Entry

When a QNE Interior node receives an intra-domain RESERVE message, it retrieves the corresponding entry from the cache and compares the values. If the message is valid, the Interior node will update <Parameter ID> and <REFRESH-PERIOD> in the list entry.

Phase 3: Delete Replay Cache Entry

When a QNE Interior node receives an intra-domain (tear) RESERVE message and an entry in the replay cache can be found, then the QNE Interior node will delete this entry after processing the message. Furthermore, the Interior node will delete cache entries, if it did not receive an intra-domain (refresh) RESERVE message during the <REFRESH-PERIOD> period with a <Parameter ID> value equal to PHR_REFRESH_UPDATE.

A.9. Example on Matching the Initiator QSPEC to the Local RMD-QSPEC

Section 3.4 of [RFC5975] describes an example of how the QSPEC can be Used within QoS-NSLP. Figure 29 illustrates a situation where a QNI and a QNR are using an end-to-end QOSM, denoted in this context as Z-e2e. It is considered that the QNI access network side is a wireless access network built on a generation "X" technology with QoS support as defined by generation "X", while QNR access network is a wired/fixed access network with its own defined QoS support.

Furthermore, it is considered that the shown QNE Edges are located at the boundary of an RMD domain and that the shown QNE Interior nodes are located inside the RMD domain.

The QNE Edges are able to run both the Z-e2e QOSM and the RMD-QOSM, while the QNE Interior nodes can only run the RMD-QOSM. The QNI is considered to be a wireless laptop, for example, while the QNR is considered to be a PC.

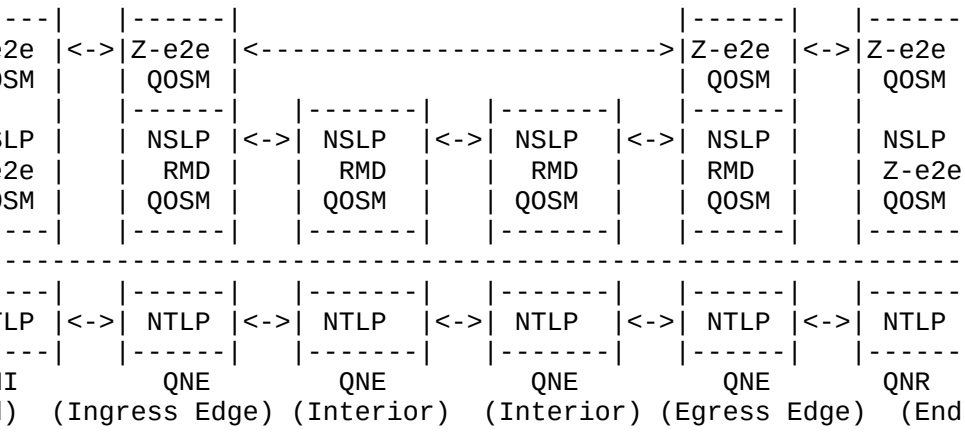

Figure 29. Example of initiator and local domain QOSM operation

The QNI sets <QoS Desired> and <QoS Available> QSPEC objects in the initiator QSPEC, and initializes <QoS Available> to <QoS Desired>. In this example, the <Minimum QoS> object is not populated. The QNI populates QSPEC parameters to ensure correct treatment of its traffic in domains down the path. Additionally, to ensure correct treatment further down the path, the QNI includes <PHB Class> in <QoS Desired>. The QNI therefore includes in the QSPEC.

<QoS Desired> = <TMOD-1> <PHB Class>
<QoS Available> = <TMOD-1> <Path Latency>

In this example, it is assumed that the <TMOD-1> parameter is used to encode the traffic parameters of a VoIP application that uses RTP and the G.711 Codec, see Appendix B in [RFC5975]. The below text is copied from [RFC5975].

> In the simplest case the Minimum Policed Unit m is the sum of the IP-, UDP- and RTP- headers + payload. The IP header in the IPv4 case has a size of 20 octets (40 octets if IPv6 is used). The UDP header has a size of 8 octets and RTP uses a 12 octet header. The

G.711 Codec specifies a bandwidth of 64 kbit/s (8000 octets/s). Assuming RTP transmits voice datagrams every 20 ms, the payload for one datagram is 8000 octets/s * 0.02 s = 160 octets.

IPv4+UDP+RTP+payload: m=20+8+12+160 octets = 200 octets
IPv6+UDP+RTP+payload: m=40+8+12+160 octets = 220 octets

The Rate r specifies the amount of octets per second. 50 datagrams are sent per second.

IPv4: r = 50 1/s * m = 10,000 octets/s
IPv6: r = 50 1/s * m = 11,000 octets/s

The bucket size b specifies the maximum burst. In this example, a burst of 10 packets is used.

IPv4: b = 10 * m = 2000 octets
IPv6: b = 10 * m = 2200 octets

In our example, we will assume that IPV4 is used and therefore, the <TMOD-1> values will be set as follows:

m = 200 octets
r = 10000 octets/s
b = 2000 octets

The <Peak Data Rate-1 (p)> and MPS are not specified above, but in our example we will assume:

p = r = 10000 octets/s
MPS = 220 octets

The <PHB Class> is set in such a way that the Expedited Forwarding (EF) PHB is used.

Since <Path Latency> and <QoS Class> are not vital parameters from the QNI's perspective, it does not raise their <M> flags.

Each QNE, which supports the Z-e2e QOSM on the path, reads and interprets those parameters in the initiator QSPEC.

When an end-to-end RESERVE message is received at a QNE Ingress node at the RMD domain border, the QNE Ingress can "hide" the initiator end-to-end RESERVE message so that only the QNE Edges process the initiator (end-to-end) RESERVE message, which then bypasses intermediate nodes between the Edges of the domain, and issues its own local RESERVE message (see Section 6). For this new local RESERVE message, the QNE Ingress node generates the local RMD-QSPEC.

The RMD-QSPEC corresponding to the RMD-QOSM is generated based on the original initiator QSPEC according to the procedures described in Section 4.5 of [RFC5974] and in Section 6 of this document. The RMD QNE Ingress maps the <TMOD-1> parameters contained in the original Initiator QSPEC into the equivalent <TMOD-1> parameter representing only the peak bandwidth in the local RMD-QSPEC.

In this example, the initial <TMOD-1> parameters are mapped into the RMD-QSPEC <TMOD-1> parameters as follows.

As specified, the RMD-QOSM bandwidth equivalent <TMOD-1> parameter of RMD-QSPEC should have:

```
r = p of initial e2e <TMOD-1> parameter
m = large;
b = large;
```

For the RMD-QSPEC <TMOD-1> parameter, the following values are calculated:

```
r = p of initial e2e <TMOD-1> parameter = 10000 octets/s
m is set in this example to large as follows:
m = MPS of initial e2e <TMOD-1> parameter = 220 octets
```

The maximum value of b = 250 gigabytes, but in our example this value is quite large. The b parameter specifies the extent to which the data rate can exceed the sustainable level for short periods of time.

In order to get a large b, in this example we consider that for a period of certain period of time the data rate can exceed the sustainable level, which in our example is the peak rate (p).

Thus, in our example, we calculate b as:

```
b = p * "period of time"
```

For this VoIP example, we can assume that this period of time is 1.5 seconds, see below:

```
b = 10000 octets/s * 1.5 seconds = 15000 octets
```

Thus, the local RMD-QSPEC <TMOD-1> values are:

```
r = 10000 octets/s
p = 10000 octets/s
m = 220 octets
b = 15000 octets
MPS = 220 octets
```

The bit level format of the RMD-QSPEC is given in Section 4.1. In particular, the Initiator/Local QSPEC bit, i.e., <I> is set to "Local" (i.e., "1") and the <Qspec Proc> is set as follows:

* Message Sequence = 0: Sender initiated
* Object combination = 0: <QoS Desired> for RESERVE and <QoS Reserved> for RESPONSE

The <QSPEC Version> used by RMD-QOSM is the default version, i.e., "0", see [RFC5975]. The <QSPEC Type> value used by the RMD-QOSM is specified in [RFC5975] and is equal to: "2".

The <Traffic Handling Directives> contains the following fields:

<Traffic Handling Directives> = <PHR container> <PDR container>

The Per-Hop Reservation container (PHR container) and the Per-Domain Reservation container (PDR container) are specified in Sections 4.1.2 and 4.1.3, respectively. The <PHR container> contains the traffic handling directives for intra-domain communication and reservation. The <PDR container> contains additional traffic handling directives that are needed for edge-to-edge communication. The RMD-QOSM <QoS Desired> and <QoS Reserved>, are specified in Section 4.1.1.

In RMD-QOSM the <QoS Desired> and <QoS Reserved> objects contain the following parameters:

<QoS Desired> = <TMOD-1> <PHB Class> <Admission Priority>
<QoS Reserved> = <TMOD-1> <PHB Class> <Admission Priority>

The bit format of the <PHB Class> (see [RFC5975] and Figures 4 and 5) and <Admission Priority> complies to the bit format specified in [RFC5975].

In this example, the RMD-QSPEC <TMOD-1> values are the ones that were calculated and given above. Furthermore, the <PHB Class>, represents the EF PHB class. Moreover, in this example the RMD reservation is established without an <Admission Priority> parameter, which is equivalent to a reservation established with an <Admission Priority> whose value is 1.

The RMD QNE Egress node updates <QoS Available> on behalf of the entire RMD domain if it can. If it cannot (since the <M> flag is not set for <Path Latency>) it raises the parameter-specific, "not-supported" flag, warning the QNR that the final latency value in <QoS Available> is imprecise.

In the "Y" access domain, the initiator QSPEC is processed by the QNR in the similar was as it was processed in the "X" wireless access domain, by the QNI.

If the reservation was successful, eventually the RESERVE request arrives at the QNR (otherwise, the QNE at which the reservation failed would have aborted the RESERVE and sent an error RESPONSE back to the QNI). If the <RII> was included in the QoS-NSLP message, the QNR generates a positive RESPONSE with QSPEC objects <QoS Reserved> and <QoS Available>. The parameters appearing in <QoS Reserved> are the same as in <QoS Desired>, with values copied from <QoS Available>. Hence, the QNR includes the following QSPEC objects in the RESPONSE message:

<QoS Reserved> = <TMOD-1> <PHB Class>
<QoS Available> = <TMOD-1> <Path Latency>

Contributors

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

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
ck technology concepts
Berlin, Germany
EMail: cornelia.kappler@cktecc.de

Hannes Tschofenig
Nokia Siemens Networks
Linnoitustie 6
Espoo 02600
Finland
EMail: Hannes.Tschofenig@nsn.com
URI: http://www.tschofenig.priv.at

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