Internet Engineering Task Force
INTERNET-DRAFT
Expires March 2004

L. Westberg
M. Jacobsson
M. de Kogel
S. Oosthoek
D. Partain
V. Rexhepi
P. Wallentin
Ericsson

G. Karagiannis
University of Twente

September 2003

# Resource Management in Diffserv On DemAnd (RODA) PHR
draft-westberg-rmd-od-phr-04.txt

## Status of this Memo

Abstract

The purpose of this draft is to present the Resource Management in Diffserv (RMD) On DemAnd (RODA) Per Hop Reservation (PHR) protocol. The RODA PHR protocol is used on a per-hop basis in a Differentiated Services (Diffserv) domain and extends the Diffserv Per Hop Behavior (PHB) with resource provisioning and control.

Table of Contents

1 Introduction ..... 4
2 Terminology ..... 4
3 RODA PHR functionality ..... 4
4 RODA PHR protocol operation ..... 6
4.1 RODA PHR Protocol Messages ..... 6
4.1.1 PHR_Resource_Request ..... 6
4.1.2 PHR_Refresh_Update ..... 7
4.1.3 PHR_Release_Request ..... 7
4.2 RODA PHR Normal operation ..... 7
4.3 Fault handling operation ..... 9
5 PHR message formats ..... 9
5.1 Message Format in IPv4 ..... 11
5.2 Message Format in IPv6 ..... 13
6 Adaptation for load sharing ..... 15
7 Tunneling ..... 16
8 Security considerations ..... 16
9 References ..... 16
10 Acknowledgments ..... 18
11 Authors' Addresses ..... 18

## 1. Introduction

The current definition of Diffserv [RFC2475] does not contain a simple and scalable solution to the problem of resource provisioning and control. The Resource Management in Diffserv (RMD) On DemAnd (RODA) Per Hop Reservation (PHR) protocol presented in this document operates in an edge-to-edge Diffserv domain extending the Per Hop Behavior (PHB) functionality with resource provisioning and control. The RODA PHR is a unicast edge-to-edge protocol that is applied in a Diffserv domain and aims at extreme simplicity and low cost of implementation along with good scaling properties. The RODA PHR protocol operates on a hop-by-hop basis on all nodes, both edge and interior, located in an edge-to-edge Diffserv domain. This PHR protocol can be applied in Diffserv domains that use either IPv4 [RFC791] or IPv6 [RFC2460].

The Resource Management in Diffserv (RMD) Framework document [RMD-frame] specifies how a PHR can interoperate with a Per Domain Reservation (PDR) protocol. A PDR scheme represents the resource reservation in the Diffserv domain, and it is implemented only at the boundary of the domain (in the edge nodes).

## 2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Furthermore, all new terms used in this draft are defined in [RMD-frame].

## 3. RODA PHR functionality

The RODA PHR protocol performs the following functions:

* The RODA PHR installs and maintains one reservation state per PHB, i.e., per DSCP, in all the nodes located in the communication path from the ingress node up to the egress node. This state represents the number of currently reserved resource units that are signalled by the PHR protocol for the admitted incoming flows. Thus, the ingress node generates a PHR signalling message for each incoming flow, which signals only the resource units

requested by this particular flow. These resource units, if reserved, are added to the currently reserved resources per PHB and therefore they will become a part of the per-PHB reservation state. The per-PHB reservation states can be created and maintained by combination of the reservation soft state and explicit release principles.

When the reservation soft state principle is used, a finite lifetime is set for the length of the reservation. These reservations are then maintained by sending periodic PHR refresh messages. The length of the refresh period MUST be the same throughout the Diffserv domain and SHOULD be configurable. If this reservation state does not receive a PHR refresh message within a refresh period, reserved resources associated with this PHR message will be released automatically. The reserved resources for a particular flow can also be explicitly released from a PHB reservation state by means of PHR release message. 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 will also reduce the number of periodic refresh messages. Furthermore, each node has to maintain a threshold per PHB that specifies the maximum number of reservable resource units. This threshold could, for example, be statically configured.

* Detection and notification of severe congestion. Severe congestion can be considered as an undesirable state which may occur as a result of a route change or a link failure. 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 re-routed traffic will have to follow a new path. Nodes located on this new path may become overloaded, since they suddenly might need to support more traffic than their capacity. All nodes MUST be able to identify a severe congestion situation. The RODA PHR protocol provides the means of informing other nodes of the congestion situation on a hop-by-hop basis.

* Stores a pre-configured threshold value on maximal allowable resource units per PHB.

* Adaptation to load sharing. Load sharing allows interior nodes to take advantage of multiple routes to the same

destination by sending via some or all of these available routes. The PHR protocol has to adapt to load sharing once it is used.

* Transport of transparent PDR messages. The PHR protocol may encapsulate and transport PDR messages sent from an ingress node to an egress node.

## 4. RODA PHR protocol operation

There are two main RODA PHR protocol operations:

* normal operation, which refers to the situation when no performance degradation problems are occurring in the network.

* fault handling, which refers to the situations when there are performance degradation problems in the network, such as route or link failures. These situations may result in severe congestion occurrence or loss of PHR messages.

### 4.1. RODA PHR Protocol Messages

In RODA, three PHR protocol messages are specified: the "PHR_Resource_Request", the "PHR_Refresh_Update" and the "PHR_Resource_Release". All pass through the same nodes as the actual traffic will pass through.

#### 4.1.1. PHR_Resource_Request

The "PHR_Resource_Request" is used to initiate or update the PHB reservation state on all nodes located on the communication path between the ingress and egress nodes according to an external QoS Request. This state represents the number of currently reserved resource units that are signalled by the "PHR_Resource_Request" for the admitted incoming flows. Thus, the ingress node generates for each new incoming flow a "PHR_Resource_Request" message, which signals only the resource units requested by this particular flow. These resource units, if reserved, are added to the currently reserved resources per PHB and therefore they will become a part of the per PHB reservation state. Furthermore, the "PHR_Resource_Request" message does not refresh any existing soft

state reservation.

#### 4.1.2. PHR_Refresh_Update

The "PHR_Refresh_Update" is used to refresh the PHB reservation soft state on all nodes located on the communication path between the ingress and egress nodes according to a resource reservation request that was successfully processed by the PHR functionality during a previous refresh period. Note that when the reservation soft state principle is used, a finite lifetime is set for the length of the reservation. These reservations are then maintained by sending periodic "PHR_Refresh_Update" messages. The length of the refresh period MUST be the same throughout the Diffserv domain and SHOULD be configurable. If this reservation state does not receive a "PHR_Refresh_Update" message within a refresh period, reserved resources associated with this PHR message will be automatically released.

#### 4.1.3. PHR_Release_Request

The "PHR_Release_Request" is used to explicitly release reserved resources for a particular flow from a PHB reservation state. Any node that receives a "PHR_Resource_Release" signalling message must identify the DSCP and release the requested resources associated with it. This can be achieved by subtracting the amount of PHR requested resources, included in the "Requested Resources" field, from the total reserved amount of resources stored in the PHB reservation state. The usage of "PHR_Release_Request" enables the instantaneous release of the resources independently of the length of the refresh period. This allows a longer refresh period, which will also reduce the number of periodic "PHR_Refresh_Update" messages.

### 4.2. RODA PHR Normal operation

All nodes SHOULD process the "PHR_Refresh_Update" message with a higher priority than the "PHR_Resource_Request" message. The detailed RODA PHR message format is described in Section 5 below. Any node that receives a RODA PHR message (a "PHR_Resource_Request" or a "PHR_Refresh_Update" message) MUST identify the DSCP of these signalling messages and, if possible, reserve the requested units of resources contained in the "Requested Resources" field of these signalling messages. If this can be accomplished then the node

reserves the requested resources by adding the requested on-demand units of resources to the total amount of reserved units associated with that DSCP.

Otherwise, these messages are marked, which means setting the "M" bit to "1". Moreover, in this case the node SHOULD include the number of previous interior nodes that successfully reserved the resources which were signalled by this "PHR_Resource_Request" into this "PHR_Resource_Request". This number is identified by the TTL (Time-To-Live) value included in the IP header of the received "PHR_Resource_Request" message. Note that each time that an IP packet passes a node, its TTL value is decreased by one. Moreover, if the TTL value of the packets becomes zero, then the packet is released.

Thus, if the ingress node is able to initialize the TTL value included in the IP header of any "PHR_Resource_Request" message sent towards the egress node then any interior node will be able to find out how many nodes before it, processed this PHR message. The node will copy the TTL value included in the IP header of the received "PHR_Resource_Request" message into the "PDR encapsulated data" field. Moreover, the node MUST set the "T" field value to "1". This PHR message will be sent towards the egress node.

Any "M" marked (the "M" bit is 1) "PHR_Resource_Request" messages that arrives in an interior node are not processed and are forwarded untouched.

Any "PHR_Refresh_Update" message, whether it is marked or not, is always processed, but marked bits are not changed.

When a node receives a "PHR_Release_Request" message it MUST identify the DSCP and estimate the refresh period where it last signalled the resource usage (where it last processed a "PHR_Refresh_Update").

This MAY be done by, for example (see [MaPo01]), giving the opportunity to an ingress node to calculate the time lag, say T_lag, between the last sent "PHR_Refresh_Update" message and the "PHR_Release_Request" message. The value of this time lag (T_lag), is first normalized to the length of the refresh period, say T_period. In other words the ratio between this time lag, T_lag, and the length of the refresh period, T_period, is calculated. This ratio is then introduced into the "Delta T" field of the "PHR_Release_Request".

When a node receives this "PHR_Release_Request" message it will have to store its arrival time. Then it will calculate the time

difference, say Tdiff, between this arrival time and the start of the current refresh period, T_period. Furthermore, this node will have to derive the value of the time lag, T_lag, from the "Delta T" field. This can be found by multiplying the value included in the "Delta T" field 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 "Requested Resources" field of the "PHR_Release_Request" message, but not below zero.

### 4.3. Fault handling operation

When a node detects this situation it MUST inform the egress node by setting the "S" field of any received PHR message to "1" and sending this message towards the egress node. In the situation that this cannot be done, operational management solutions, such as Simple Network Management Protocol (SNMP) notifications SHOULD be used.

Moreover, when an interior node detects this situation, it SHOULD notify the egress node by using DSCP remarking of user data packets that are passing through the node. Proportionally to the detected overload, the interior node will remark a number of user data packets which are passing through a severe congested interior node and are associated to a certain PHB, into a domain specific DSCP (see [RFC2474]). [RMD-frame] describes a severe congestion handling procedure which uses the DSCP remarked packets and solves the severe congestion situation.

Any "S" marked (the "S" bit is 1) "PHR_Resource_Request" messages that arrives in an interior node are not processed and are forwarded untouched. Any "PHR_Refresh_Update" message, whether it is marked or not, is always processed, but marked bits are not changed.

## 5. PHR message formats

The PHR protocol information is carried in:

* an IP header Options field, as defined in the [RFC791], when IPv4 is used

* an option field encoded into the Hop-by-Hop Options Extended Header, as defined in [RFC2460], when IPv6 is used

We denote this IP Option field as the RODA PHR option.

### 5.1. Message Format in IPv4

The RODA PHR protocol messages used in IPv4 Diffserv domains are represented by the combination of the DSCP field and the contents of an IPv4 option header field [RFC791]. This IPv4 option header field has the following format. Note that the contents of the PDR (per-domain reservation) encapsulated data are simply opaque data to the PHR and are not processed by the PHR. Please see [RMD-frame] for a description of PDR functionality.

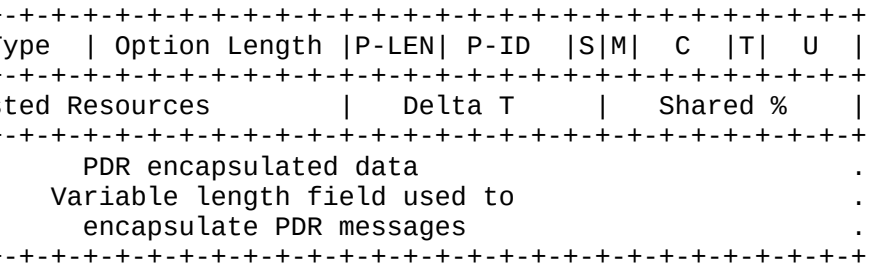


Figure 1: PHR Option field in the IPv4 Option header field

Option Type — 8-bit identifier of the type of option. The semantics of this field are specified in [RFC791].

Option Length — 8-bit field. This is specified in [RFC791] and represents the length of the Option-Data field of this option, in octets. The option data field consists of all fields included in the option field of the IPv4 header and are placed after the "Option Length" field.

P-LEN (PHR length) — 3-bit field. This specifies the length in octets of the specific PHR information data included in the "Option-Data" field. This information does not include the encapsulated PDR information.

The value 0 specifies that this IP option field contains only PDR data and no PHR data. The PDR data MUST begin on the next 32-bit word boundary after the P-LEN field (after the first "unused" field). In this case, the sender MUST set the "S", "M", "C", and "unused" fields to 0. The P-ID MUST have the value 1.

If a receiver receives a packet with a P-LEN value of 0, it MUST ignore the values in the "S", "M", "C", and "unused" fields.

P-ID (PHR type) 4-bit field. This specifies the PHR type. For the RODA PHR, the value MUST be 1.

S (Severe Congestion) 1-bit field. The sender MUST set the "S" field to 0. This field is set to 1 by an interior or edge node when a severe congestion situation occurs.

M (Marked) 1-bit field. The sender MUST set the "M" field to 0. This field is set to 1 by an interior or edge node when the node cannot satisfy the "Requested Resources" value.

C (Message type) 3-bit field. This field specifies the type of the PHR message.

| C | Description |
|---|---|
| 0 | Reserved |
| 1 | "PHR_Resource_Request" |
| 2 | "PHR_Refresh_Update" |
| 3 | "PHR_Release_Request" |
| 4-7 | Unused |

T (TTL active) 1-bit field. The ingress node (i.e., sender) MUST set the "T" field to 0. This field MAY be set to "1" by a node when the node will have to include the TTL value from the header of the IP packet into the "PDR encapsulated data" field.

U A 3-bit field that is currently unused. Reserved for future PHR extensions.

Requested Resources 16-bit field. This field specifies the requested number of units of resources to be reserved by a node. The unit is not necessarily a simple bandwidth value. It may be defined in terms of any resource unit (e.g., effective bandwidth) to support statistical multiplexing at message level.

| | |
|---|---|
| Delta T | 8 bit field. The value of this field MAY be set by any ingress node into (only) "PHR_Resource_Release" messages. It specifies a percentage that represents the ratio between a time lag, say T_lag, and the length of the refresh period, say T_period. Where, T_lag represents the difference between the departure time of the previous sent "PHR_Refresh_Update" message and the departure time of the "PHR_Resource_Release" message. T_period represents the length of the refresh period. This information MAY be used by any node during an explicit release procedure. |
| Shared % (Shared path percentage) | 8 bit field. This value MAY be used to specify if a load sharing situation occurred on a communication or not. The ingress node sets this value to 100. If load sharing occurred in a node then the node will divide the shared percentage value to the number of equal cost paths. |
| PDR encapsulated data | PDR encapsulated information data. This field is only processed by the edge nodes. |

## 5.2. Message Format in IPv6

The PHR protocol messages used in IPv6 Diffserv domains are represented by the combination of the DSCP field and the contents of an option field of a IPv6 Hop-by-Hop header option [RFC2460]. This IPv6 option field has the following format.

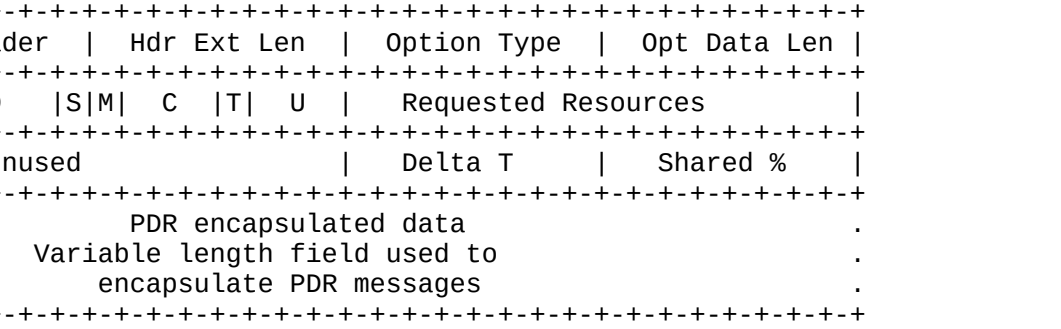


Figure 2: PHR Option field in the IPv6 Hop-by-Hop Header Option

| | |
|---|---|
| Next Header | 8-bit selector. This is specified in [RFC2460] and identifies the type of header immediately following the Hop-by-Hop Options header. |
| Hdr Ext Len | 8-bit field. This is specified in [RFC2460] and represents the length of the Hop-by-Hop Options header in 8-octet units, not including the first 8 octets. |
| Option Type | 8-bit identifier of the type of option. The semantics of this field are specified in [RFC2460]. |
| Opt Data Len | 8-bit field. This is specified in [RFC2460] and represents the length in octets of the Option Data field of this option. The option data field consists of all fields included in the Hop-by-Hop header option and placed after the "Opt Data Len" field. |
| P-LEN (PHR length) | 3-bit field. The semantics of this field are identical to the field in the IPv4 option.<br><br>Just as for IPv4, the value 0 specifies that this IP option field contains only PDR data and no PHR data. The PDR data MUST begin on the next 32-bit word boundary after the P-LEN field (after the first "Requested Resources" field). In this case, the sender MUST set the "S", "M", "C", "unused", and "Requested Resources" fields to 0. The P-ID MUST have the value 1.<br><br>If a receiver receives a packet with a P-LEN value of 0, it MUST ignore the values in the "S", "M", "C", and "unused" fields. |
| U | A 3-bit field that is currently unused. Reserved for future PHR extensions. |
| UNUSED | A 16-bit field that is currently unused. Reserved for future PHR extensions. |
| PDR encapsulated data | a variable length field that contain PDR encapsulated information data. This field is only processed by the edge nodes. |

The "Requested Resources", "P-LEN", "P-ID", "S", "M" and "C", "T",

"Delta T" and "Shared %" fields in Figure 2 are identical to those shown in Figure 1.

## 6. Adaptation for load sharing

Due to load sharing (see e.g., [RFC2676]), a node may cycle between different routes in order to balance the load. This will imply that the traffic (user) data will not follow exactly the same paths as the PHR messages used to reserve or refresh the transport resources used by this traffic (user) data. As such, interior nodes MUST be able to observe when a load sharing situation occurs.

It is recommended that interior and edge nodes SHOULD forward the PHR messages in such a way that they will follow the same forwarding path as the traffic (user) data associated with these PHR messages. When this cannot be done, we propose use of the same solutions as the multi-path route solutions proposed in Section 1.4.6 of [RFC3175]. These are:

* the data may be tunneled from the ingress to egress node using technologies such as IP-in-IP, GRE (Generic Routing Encapsulation), MPLS (Multiple Label Protocol Switching) label-switched paths, and so on.

* measurement could be used to determine what proportion of traffic for a given reservation travels along each of the load sharing paths, thereby verifying that there is sufficient bandwidth for the reservation.

* by reserving the total capacity of the route down each load sharing path.

In case a network domain is using a routing protocol which is applying an equal cost load sharing principle, any interior node SHOULD be able to know the number, e.g., "N", of multiple equal cost paths that the routing protocol will use to provide the load sharing principle. Subsequently, for each arrived PHR message which is affected by the load sharing principle, the interior node SHOULD be able to create "N" number of PHR messages of identical type as the original one. Each of these generated PHR messages SHOULD contain in its "Requested Resources" field a value equal to the requested resources value which was included in the "Requested Resources" field of the original PHR message divided by the number of equal cost paths, i.e., "N". Moreover, each of these generated PHR messages

SHOULD also contain in its "Shared %" field a new value that is calculated by dividing the shared percentage value, included in the "Shared %" field of the original PHR message, by the number of equal cost paths, i.e., "N".

## 7. Tunneling

When PHR messages are tunneled within the RMD Diffserv domain, the tunneling messages MUST include the PHR option field.

## 8. Security considerations

The general security and tunneling considerations stated in Section 6 of [RFC2475] and [RMD-frame] also apply to this PHR.

In addition, unlike Differentiated Services PHBs, the RODA PHR allows the edge nodes to reserve bandwidth or other QoS parameters dynamically. This flexibility makes it more vulnerable to erroneous reservations and sabotage. In order to keep functioning properly, the edge nodes MUST be certain that any flow reserving bandwidth in the network is authorized to do this and only up to that flow's agreed upon limit. If the edge node detects erroneous or malicious behavior, it MUST police that flow to the agreed upon limits or reject it entirely.

Because of the soft state principle used, the PHR can recover relatively easily from incorrect reservations. Thus it is quite safe to deploy the RODA PHR in a well-controlled network with trustworthy edge nodes.

In order to prevent abuse of the QoS capabilities of the core network, the ingress nodes SHOULD filter any PHR or PDR related header information coming from the outside before sending it through the core network. Whether this information needs to be preserved and later re-inserted or if it should be discarded from the packet or if the entire packet should be discarded is an open issue.

## 9. References

[MaPo01] Marquetant, A., Pop, O., Szabo, R., Dinnyes, G., Turanyi, Z., "Novel enhancements to load control

a soft state, lightweight admission control protocol", QofIS'2000 - 2nd International Workshop on Quality of future Internet Services, September 2001.

[RMD-frame] Karagiannis, G., Rexhepi, V., Westberg, L., Partain, D., Oosthoek, S., Jacobsson, M., Szabo, R., Wallentin, P., "Resource Management in Diffserv Framework", Internet draft, February 2002 (work in progress).

[RFC791] DARPA INTERNET PROGRAM PROTOCOL SPECIFICATION, "Internet Protocol", IETF RFC 791, September 1981.

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

[RFC2205] Braden, R., Zhang, L., Berson, S., Herzog, A., Jamin, S., "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", IETF RFC 2205, 1997.

[RFC2460] Deering, S., Hinden, R., "Internet Protocol, Version 6 (IPv6) Specification", IETF RFC 2460, December 1998.

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

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

[RFC2676] Apostolopoulos, G., Willians, D., Kamat, S., Guerin, R., Orda, A., Przygienda, T., "QoS Routing Mechanisms and OSPF Extensions", IETF Experimental RFC 2676, August 1999.

[RFC2859] Fang, W., Seddigh, N., Nandy, B., "A Time Sliding Window Three Colour Marker (TSWTCM)", IETF Experimental RFC 2859, June 2000.

[RFC3175] Baker, F., Iturralde, C. Le Faucher, F., Davie, B.,

"Aggregation of RSVP for IPv4 and IPv6 Reservations",
IETF RFC 3175, 2001.

## 10. Acknowledgments

Thanks to Robert Szabo and Geert Heijenk for reviewing this draft and providing useful input.

## 11. Authors' Addresses

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

Martin Jacobsson
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Martin.Jacobsson@eln.ericsson.se

Georgios Karagiannis
University of Twente
P.O. BOX 217
7500 AE Enschede
The Netherlands
EMail: karagian@cs.utwente.nl

Simon Oosthoek
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Simon.Oosthoek@eln.ericsson.se

David Partain
Ericsson Radio Systems AB

P.O. Box 1248
SE-581 12 Linkoping
Sweden
EMail: David.Partain@ericsson.com

Vlora Rexhepi
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Vlora.Rexhepi@eln.ericsson.se

Pontus Wallentin
Ericsson Radio Systems AB
P.O. Box 1248
SE-581 12 Linkoping
Sweden
EMail: Pontus.Wallentin@era.ericsson.se

Marcel de Kogel
Ericsson EuroLab Netherlands B.V.
5121 ML Rijen
The Netherlands
EMail: Marcel.de.Kogel@eln.ericsson.se