L. Westberg
A. Bhargava
A. Bader
Ericsson
G. Karagiannis
University of Twente

November 14, 2007

LC-PCN: The Load Control PCN Solution
draft-westberg-pcn-load-control-02

Status of this Memo

Copyright Notice

Abstract

There is an increased interest of simple and scalable resource provisioning solution for Diffserv network. The Load Control PCN (LC-PCN) addresses the following issues:

- o Admission Control for real time data flows in stateless Diffserv Domains

- o Flow Termination: Termination of flows in case of exceptional events, such as severe congestion after re-routing.

Admission control in a Diffserv stateless domain is a combination of:

- o Probing, whereby a probe packet is sent along the forwarding path in a network to determine whether a flow can be admitted based upon the current congestion state of the network

- o Admission Control based on data marking, whereby in congestion situations the data packets are marked to notify the PCN-egress-node that a congestion occurred on a particular PCN-ingress-node to PCN-egress-node path.

The scheme provides the capability of controlling the traffic load in the network without requiring signaling or any per-flow processing in the PCN-interior-nodes. The complexity of Load Control is kept to a minimum to make implementation simple.

Table of Contents

1. Introduction

The amount of traffic carried on the Internet is now greater than the traffic on the world's telephony network. Still, Internet-based communication services generate less income than plain old telephony services. Enabling value-added services over the Internet is therefore crucial for service providers. One significant class of such value-added services requires real-time packet transportation. It can be expected that these real-time services will be popular as they replicate or are natural extensions of existing communication services like telephony. Exact and reliable resource management (e.g., admission control) is essential for achieving high utilization in networks with real-time transportation capabilities. The problem is difficult mainly due to scalability issues.

With the introduction of differentiated services (DS) [RFC2475], it is now possible to provide large scale, real-time services. The basic idea of DiffServ is that, rather than classifying packets at each router, packets are only classified at the edge devices. The result - the required packet treatment - is stored and carried in the packet headers, and core routers can carry out appropriate scheduling.

The current definition of DiffServ, however, does not contain any simple, scalable solution to the problem of resource provisioning and control. A number of approaches to solving the problem already exist [RFC3175], [Berson97], [Stoica99], [Bernet99]. The scheme presented in this document does not require any state aggregation and aims at extreme simplicity and low cost of implementation along with good scaling properties. Load control operates edge-to-edge in a DS domain, or between two RSVP or NSIS capable routers, where only the edge devices keep flow state and do per-flow processing. The main purpose of Load Control is to provide a simple and scalable solution to the resource provisioning problem.

The original Load Control concept, submitted in April 2000, [Westberg00], has been developed further to a signaling concept named Resource Management in Diffserv. RMD was incorporated by NSIS working group, where the protocol details were worked out for using NSIS as external protocol [RMD]. Recently new drafts have been submitted aiming to standardize new Diffserv PHB that provides controlled load services in Diffserv domains [CL-PHB], [CL-ARCH], [Babi07], [Char07]. These concepts are very similar to the original two-bit marking scheme of Load Control.

This document aims to develop a common framework that could be used both with RSVP and NSIS external protocols.

The remainder of this draft is structured as follows. After the terminology in Section 2, we give an overview of the LC-PCN in Section 3. In Section 4 we give a detailed description of the LC-PCN. Section 5 discusses security issues.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119. The terms specified in [Eard07] are used.

3. LC-PCN Overview

Load Control PCN (LC-PCN) is achieved by two actions: Admission Control based on probing and/or Flow Termination. The LC-PCN can be applied within either a single PCN domain, see Figure 1, or multiple neighboring PCN domains, when a trust relationship exists between these multiple PCN domains.

```
     PCN-Ingress-Node                                  PCN-Egress-Node
                        (PCN-Interior-Nodes; I-Nodes)
                            |          |            |
                            |          |            |
                            V          V            V
     +-------+   Data +------+      +------+       +------+     +------+
     |-------|--------|------|------|------|-------|------|---->|------|
     |       |   Flow |      |      |      |       |      |     |      |
     |Ingress|        |I-Node|      |I-Node|       |I-Node|     |Egress|
     |       |        |      |      |      |       |      |     |      |
     +-------+        +------+      +------+       +------+     +------+
              =================================================>
              <=================================================
                                    Signaling
```

Figure 1: Actors in the LC-PCN

3.1. Admission control based on probing

The admission control function based on probing can be used to implement a simple measurement-based admission control within a PCN domain. In the PCN-interior-nodes thresholds are set for the traffic belonging to different PHBs in the measurement based admission control function. In this scenario an IP packet is used as a probe packet, meaning that the DSCP field in the header of the IP packet is re-marked when the measured PHB throughput rate exceeds a predefined congestion threshold, i.e, PCN_lower_rate.In addition to this the PCN_ingress_node has to set the Router Alert IP option on the probe packet. In this way all the PCN_interior_node will have to observe the received probe packets. Thus if a PCN_interior_node receives a probe packet then, due to the Router Alert option it has to handle it differently then the user packets.

The PCN_interior_node has to PCN_mark the probe packet if it is operating in Admission Control state (or Flow Termination state). Otherwise the probe packet remains unmarked.

In this way the data packets are marked to notify the PCN-egress-node that a congestion has occurred on a particular PCN-ingress-node to PCN-egress-node path.

If no probing is used, the request for admission can be accomplished by using an external to PCN, signaling protocol. In this case when the request, carried by the external to PCN signaling protocol arrives at a PCN_egress_node that operates in admission control state then the request is rejected. If it operates in Normal state it is accepted.

If probing is used, the request for admission is accomplished by using a probe packet. In this case when the probe arrives at a PCN_egress_node and it is PCN_marking encoded is rejected. Otherwise is accepted.

Note that by using probing, the ECMP (Equal Cost Multi Path) problem that is associated with the admission control feature can be, to a certain degree, solved by being able to identify which flows are passing through the congested node. Note that the ECMP problem is related to the fact that flows that are not passing through a congested PCN-interior-node can belong to an aggregate that detects a congestion.

Any measures that are taken on such flows will not solve the congestion problem, since such flows are not contributing and causing the congestion in the PCN-interior-node.

3.2. Flow Termination

The Flow Termination function is able to terminate flows in case of exceptional events, such as severe congestion after re-routing. The exceptional event, or severe congestion can be detected using a DSCP remarking approach where the PCN_marking is proportional to the excess rate. In particular, the PCN-interior-nodes packets using the PCN_marking DSCP, whenever the measured PHB throughput rate exceeds a pre-configured throughput threshold denoted as PCN_upper_rate.

The PCN-egress-nodes can use the remarked PCN_marking DSCP packets to calculate the fraction of throughput or bandwidth that does exceed PCN_upper_rate_egress. The PCN_Affected_marking DSCP is used to mark all packets that are passing through an PCN-interior-node that is either in Flow Termination state and are not PCN_marking DSCP encoded. In this way an ECMP solution can be provided for the Flow Termination state. The PCN-egress-node can then, in combination with the PCN-ingress-node, sender of the traffic and the support of the PCN domain(s), reduce the generated rate, by terminating ongoing flows, until the excess rate drops below PCN_upper_rate_egress.

3.3. Common PCN node configurations

The PCN-interior-nodes, see Figure 1, which are supporting the LC-PCN, must perform the following functionalities:

(1) Meter + (2) Marking Action: the PCN-interior-nodes must be configured with a meter and marking function that measures and remarks bytes that are out of a configured traffic profile (e.g., bandwidth threshold) for a corresponding PHB traffic class, to provide an indication of a potential resource limitation to a PCN-egress-node. The traffic profile can be set according to an engineered bandwidth limitation based on pre-configured thresholds or based on a capacity limitation of specific PHBs. By using an algorithm that calculates the rate of bytes that are out of profile, say signaled_remarked_bytes; a special number of bytes, i.e., signaled_remarked_bytes/N, are remarked to a second DSCP, denoted in this example as PCN_marking DSCP, that receives the same PHB as the original DSCP (where N is equal or greater than 1). Another type of encoding that is used, is the PCN_Affected_marking DSCP, which is used to mark all packets that are passing through an PCN-interior-node in Flow Termination state and the arriving packets are not PCN_marking DSCP encoded.

The PCN_marking DSCP and PCN_Affected_marking DSCP are defined to be used only locally within the PCN domain. "N" is a pre-configured parameter used to indicate the proportionality between the measured out of profile bytes and the remarked bytes. If "N" is used in the algorithm, then it must have the same value in all Diffserv nodes that use this mechanism. As previously mentioned, N is higher or equal to 1 (N >= 1).

(3) Packet Classification + (4) Scheduling: The PCN-interior-node SHOULD be configured to consider that the packets marked either with the original DSCP or with the PCN_marking DSCP or Affected_ marking DSCP SHOULD receive the same per hop behavior treatment. However, packets that are marked with the PCN_marking DSCP, may be classified to enter a different and larger virtual queue than the packets marked

with either the original DSCP or PCN_Affected_marking DSCP. This can ensure that the dropping probability of PCN_marking DSCP remarked packets is lower than the dropping probability of original DSCP remarked packets. This classification can be accomplished by using the packet classification function, while the way of how the packets are treated in the virtual queues is accomplished using the scheduling function. Note that the original DSCP marked packets and their associated PCN_marking DSCP packets get the same forwarding behavior. The main difference is related to the fact that the PCN_marking DSCP packets get a lower dropping probability compared to the original_DSCP packets. This is because the marking information carried by the PCN_marking DSCP packets has a higher significance for the operation of the resource unavailability algorithm compared to the marking information carried by the original_DSCP packets.

The two virtual queues, one for the original_DSCP and another one for PCN_marking DSCP marked packets can, for example, be implemented by using one Drop Tail physical queue and by maintaining queuing information and also one queuing threshold for each of the virtual queues. The physical queue uses the same scheduling algorithm, but the length of each of the virtual queue defines the packet dropping probability of a virtual queue. The classification of packets SHOULD be based on either the DSCP or on a combination of IP header fields including the DSCP.

When the LC-PCN is applied in multiple neighboring PCN domains where a trust relationship exists between these multiple PCN domains and a packet is received by the edge router of another trusted domain (new PCN domain, that might be managed by another operator), remarking of the original DSCP, PCN_marking DSCP and PCN_Affected_marking DSCP to other DSCPs, say original new_DSCP, PCN_marking new_DSCP and PCN_Affected_marking new_DSCP might be necessary. This is because the neighbor PCN operator may use different Diffserv Mapping schemes.

PCN_upper_rate is configured in all PCN-interior-nodes and it can be calculated in the following way:

PCN_upper_rate = Maximum PHB capacity - Termination_offset_rate

Maximum PHB capacity is the maximum link capacity that is supported by a PCN node.

The Termination_offset_rate is an absolute rate value that should be set equal into all PCN_interior_nodes. The Termination_offset_rate can also be equal to 0.

Note that this value is used by PCN_interior_nodes to calculate their PCN_upper_rate and is also used during the situation that a

PCN_interior_node is in flow termination state and it receives PCN_marked packets. This situation occurs when more than one PCN-interior-nodes located on same communication path, are simultaneously operating in the admission control state or flow termination state. The Termination_offset_rate is needed due to the following fact. Consider the fact that when the measured PHB rate exceeds the "Maximum PHB capacity" then the packets belonging to the given PHB will be either dropped or set to another PHB. In multiple severe congestion situations solving the severe congestion on a severe congestion PCN_Interior_node, further away than the PCN_egress_node, say severe_congestion_point_1, it could cause the situation that the severe congestion on a PCN_Interior_node located on the same path and closer to the PCN_egress_node, say severe_congestion_point_2, will be solved without marking the excess rate measured at severe_congestion_point_2. This is however true only if the measured PHB rate on severe_congestion_point_1 does not exceed the "Maximum PHB capacity". This is due to the fact that before the severe_congestion_point_1 goes into flow termination it generates a measured PHB rate that it does not exceed the value equal to ("Maximum PHB capacity"- Termination_offset_rate) and in flow termination state it generates a measured PHB rate not higher than "Maximum PHB capacity". Thus if the excess rate on severe_congestion_point_1 is higher than "Maximum PHB capacity" then this it is not seen by severe_congestion_point_2 but, due to the principle of marking, it will be seen by the PCN_egress_nodes.

Therefore, the severe_congestion_point_2 has to consider the incoming_PCN_marked_rate from severe_congestion_point_1 in its marking algorithm only for measured PHB rates higher than the PCN_upper_rate (associated with severe_congestion_point_1) and lower or equal to the PCN_upper_rate + Termination_offset_rate. The severe_congestion_point_2 can compute the Termination_offset_rate used by the previous severe congestion point by using a variable that is the same in the whole PCN domain.

PCN_lower_rate is configured in all PCN-interior-nodes and is calculated in the following way:

PCN_lower_rate = PCN_upper_rate - Admission_offset_rate

The Admission_offset_rate is an absolute rate value and it is equal in all PCN_interior_nodes and PCN_egress_nodes.

The Admission_offset_rate and Termination_offset_rate are required in order to provide a solution for the situation that more than one PCN-interior-nodes located on same communication path, are simultaneously operating in the Admission Control or Flow Termination state, respectivelly.

The Admission_offset_rate and Termination_offset_rate are required in order to provide a solution for the situation that more than one PCN-interior-nodes located on same communication path, are simultaneously operating in the admission control state or flow termination state, respectively.

It is however, considered that SLA agreements exist between the operator(s) of these PCN domains, thus also the remarking rules followed in each PCN domain are known. Note that the PCN nodes used in the neigbouring PCN domains should use the same classification, meter & marking actions as described above.

3.4. Configuration of edge nodes

The edges must maintains aggregated states that encompass several flows/calls. The size of the aggregates should be large enough to ensure that new flows/calls belong to aggregates where ongoing calls provide feedback for admission control decisions. In addition to this the edges must maintain per flow states.

When the PCN-egress-nodes, receive the remarked PCN_marking DSCP packets, the rate of the received PCN_marking DSCP bytes, per each flow aggregate, is measured. Note that the calculated rate has to be multiplied with the parameter "N", above, in order to calculate the real rate of overload, say signaled_overload_rate. This rate can be used to provide handling decisions on the Admission Control and Flow Termination functionality. Two types of handling decisions could be supported.

For admission control, the PCN-egress-node can maintain at least one threshold, say PCN_lower_rate_egress. Then if the calculated rate of remarked PCN_marking DSCP bytes is higher than PCN_lower_rate_egress, i.e., signaled_overload_rate > PCN_lower_rate_egress, then the PCN-egress-node can use this information to provide the basis of call admission decisions for new flows. The detailed specification of this algorithm is given in Section 4.1.4.

One way to calculate the PCN_lower_rate_egress threshold that defines when a PCN_egress_node goes into the admission control state that is to monitor when the PCN_egress_node receives a PCN_marked packet. That will mean that at least one intermediate PCN_interior_node started to be in congested state and thus the egress node transition from Normal state to admission control state. We use a fraction of the received PCN_marking encoded packets to be realistic. The value of PCN_lower_rate_egress is calculated as follows:

PCN_lower_rate_egress = A * Admission_offset_rate, where 0 < A < 1
Typically, factor A should be set low around 1%.

If the PCN domain supports probing then the PCN-ingress-node is configured such that when it receives a request for reservation message, it generates a probe packet that is sent within the PCN domain. The probe packet should use the same flow ID and DSCP value as the ones used by the data packets associated with the request for reservation message. Furthermore, the probe packet MUST enable the Router Alert Option.

If the PCN-ingress-node receives a response that notifies that the probe was successfully processed, then the reservation request is admitted. Otherwise it is rejected. Both situations are notified to the sender of the flow.

If no probing is used within the PCN domain, the request for admission can be accomplished by using an external to PCN signaling protocol. In this case when the request arrives at a PCN_egress_node that operates in admission control operation/state then the request is rejected. If it operates in Normal operation/state is accepted.

When the Flow Termination procedure is also supported, then at least two pre-configured bandwidth thresholds are used, i.e., PCN_lower_rate_egress and PCN_upper_rate_egress, with PCN_upper_rate_egress > PCN_lower_rate_egress.

But how will the PCN_egress_node change state from Admission Control state to Flow Termination state. Two solutions are provided below that specify how the PCN_egress_node can transition from Admission control state to Flow Termination state. First solution: if the PCN_interior_nodes use the PCN_Affected_marking encoding only during flow termination for the packets that are passing through the severe congested node, but without being PCN_marked, then the PCN_egress_node can change to flow termination state when it receives PCN_Affected_marked packets. The transition from flow termination state to normal state occurs when the PCN_egress_node does not receive any PCN_Affected_marked packets. Second solution: In order to explain this, it is important to note that each PCN_interior_node, that is in Admission Control state, can PCN_mark packets up to Admission_offset_rate. Furthermore, if a PCN_interior_node receives incoming PCN_marked packets and is in the Admission Control state, will not remark any packets if the excess rate is equal or lower than the incoming_PCN_marking_rate. If we consider the situation where no ECMP occurs and that all flows belonging to the same ingress-egress pair will use the same path from PCN_ingress to PCN_egress, this would mean that when the PCN_egress_node receives an excess rate equal to a fraction of the Admission_offset_rate i.e. F * Admission_offset_rate, where 1 >= F > A, it would transition from Admission Control state to Flow Termination state. Note that F can be preconfigured and depends on the network topology. Thus in this case the second threshold, is calculated as follows:

PCN_upper_egress_rate = PCN_lower_egress_rate + F * Admission_offset_rate. However, there are some special/corner cases, that mainly occur when different congestion points (admission control congested PCN_interior_nodes) on the same path are not simultaneously starting to be congested. Therefore we use the multicongestion_error parameter to identify the error bound that occurs due to these special cases. Note that this error bound can be e.g., predefined ones off line by the operator, by studying the network topology and/or studying how often such corner cases could occur and/or doing off line measurements. Therefore, the PCN_upper_rate_egress can be calculated as follows:

```
PCN_upper_rate_egress = PCN_lower_rate_egress +
     F * Admission_offset_rate +/- multicongestion_error
```

Note that when the PCN_Affected_marking is applied in whole PCN domain, then the first solution described above SHOULD be selected, otherwise the second solution described above SHOULD be selected.

The PCN-egress-node should operate in the following way.

When the PCN-egress-node operates in flow termination state, then the PCN- egress-node can calculate the amount of excess rate above this threshold, see Section 4.2.3.

By using this excess rate, the PCN-egress-node can support the below options:

- o identify ongoing flows, that are part of the aggregate, to be terminated and send Flow Termination notifications to these ongoing sessions towards the PCN-ingress-node
- o send the measured value(s) of the excess rate towards the PCN-ingress-node

The "PCN_Affected_marking DSCP" encoding is used to mark all packets that are passing through an PCN-interior-node that is operating in Flow Termination state and are not "PCN_marking DSCP" encoded. The PCN-egress-node uses the received "PCN_Affected_marking DSCP" packets to identify which flows have passed through one or more PCN-Interior-Nodes that operate in Flow Termination state. In this way an ECMP solution can be provided for the Flow Termination state.

If the PCN-ingress-node, due to the Flow Termination congestion situation, receives flow termination notifications for certain flows, it will have to terminate these flows within the PCN domain and send flow termination notifications towards the sender of these flows. The PCN-ingress-node, up to the moment that the severe congestion situation is solved, it will also have to stop admitting new flows that could be incorporated within the aggregated state that is affected by the severe congestion situation. Furthermore, the PCN-ingress-node uses the received measured excess rate to resize the aggregated reservation state.

4. LC-PCN detailed description

This section describes the details of the used LC-PCN algorithms. Section 4.1 and 4.2 describe the "Admission control based on probing" and "Flow Termination" scenario, respectively, for the situation that the end-to-end sessions are using unidirectional reservations. Sections 4.3 and 4.4 are describing the two algorithms for the situation that the end-to-end sessions are using bi-directional reservations.

4.1. Admission control based on probing for unidirectional flows

The admission control function based on probing can be used to implement a simple measurement-based admission control within a PCN domain. At PCN-interior-nodes along the data path PCN_lower_rate are set in the measurement based admission control function for the traffic belonging to different PHBs.

4.1.1. Operation in PCN-ingress-nodes

After a trigger event, e.g., the PCN-ingress-node receives a reservation request message, the PCN-ingress-node can do the following:

If the PCN domain supports probing, then the PCN_ingress_node sends a probe packet, see Figure 2, towards the PCN-egress-node. Note that the probe packet should use the same flow ID information and DSCP value as the data packets associated with the received reservation request message. The probe packet SHOULD set a Router Alert Option. If the PCN-ingress-node receives a response that notifies that the probe was successfully processed, then the reservation request is admitted. Otherwise it is rejected. Both situations have to be notified to the sender of the flow.

If the PCN domain does not support probing, then the reservation request message belonging to the external signaling protocol can be used during the admission control process. If the PCN-ingress-node receives a response that notifies that the reservation request message belonging to the external signaling protocol was successfully processed, then the reservation request is admitted. Otherwise it is rejected.Both situations have to be notified to the sender of the flow.

4.1.2. Operation in PCN-interior-nodes

Using standard functionalities admission control thresholds, i.e., PCN_lower_rate, are set for the traffic belonging to different PHBs, see Section 3.

When the PCN_interior_node operates in Admission Control state and the PCN_lower_rate is exceeded then the DSCP field of data packets are proportionally to the excess rate re-marked, using the PCN_marking DSCP, see event A, in Figure 4. Furthermore, when probing is used and when the PCN_interior_node operates in admission control state and it receives a probe packet, this probe packet MUST be remarked using the PCN_mark DSCP encoding. Note that the probe packet will be processed by the PCN_interior_node since it carries a Router Alert Option.

An example of the detailed operation of this procedure is described below.

The predefined PCN_lower_rate, see Section 3.3 and Section 4.2.2 is set according to, and usually less than, an engineered bandwidth limitation, i.e., real admission threshold, based on e.g. agreed Service Level Agreement or a capacity limitation of specific links. The difference between the PCN_lower_rate and the engineered bandwidth limitation, i.e., real 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. Note that this difference is used at the PCN-egress-node to trigger the situation that the PCN-egress-node operates in the admission control state. This is due to the fact that data packets associated with an admitted session have not yet arrived, while allows the admission control process available at the PCN-egress-node to interpret the signaling information and reject new calls before reaching congestion. Note that in the situation when the data rate is higher than the preconfigured congestion notification rate, also data packets are re-marked to PCN_marking DSCP.

During admission control the interior node calculates, per traffic class (PHB), the incoming rate that is above PCN_lower_rate, denoted as signaled_overload_rate, in the following way:

o before queuing and eventually dropping the packets, at the end of each measurement interval of T seconds, the PCN-interior-node should count the total number of original DSCP, PCN_marking DSCP and PCN_Affected_marking DSCP bytes received, denote this number as total_received_bytes. Note that there are situations when more than one PCN-interior-nodes in the same communication path become admission control congested and operate in Admission Control state. Therefore, any PCN-interior-node located behind a PCN-interior-node that operates in Admission Control state may receive PCN_marking DSCP and PCN_Affected_marking DSCP bytes.

Then the PCN-interior-node calculates the current estimated overloaded rate, say signaled_overload_rate, by using the following equation:

```
signaled_overload_rate =
   ((total_received_bytes) / T) - PCN_lower_rate)
```

To provide reliable estimation of the encoded information several techniques can be used, see [AtLi01], [AdCa03], [ThCo04], [AnHa06].

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

```
IF (measured PHB rate > PCN_lower_rate) AND
   (measured PHB rate =< PCN_upper_rate)
THEN
 {
   IF (incoming_PCN_marking_rate <> 0) AND
      (incoming_PCN_marking_rate <= Admission_offset_rate)
   THEN
    { signaled_remarked_bytes =
        ((signaled_overload_rate -
         incoming_PCN_marking_rate) * T) / N
    }
   ELSE IF (incoming_PCN_marking_rate = 0)
   THEN signaled_remarked_bytes =
          signaled_overload_rate * T / N
   ELSE IF (incoming_PCN_marking_rate >
             Admission_offset_rate)
   THEN signaled_remarked_bytes = 0
  }
```

Where the "incoming_PCN_marking_rate" is calculated as follows:

```
incoming_PCN_marking_rate =
   (received number of "PCN_marking" DSCP during T) * N)/T
```

When incoming remarked bytes are dropped, the operation of the admission control 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 PCN_marking DSCP and unmarked (original DSCP and PCN_Affected_marking DSCP) bytes, see Section 3.3.

When the measured PHB throughput rate is higher than PCN_upper_rate, see Figure 4, then it is considered that the operation PCN-interior-node has moved to the Flow Termination state.

4.1.3. Operation in PCN-egress-nodes

When the operation state of the ingress/egress pair aggregate in the PCN_egress_node is in the Admission Control state (see Figure 4 and Section 4.2.3), then the implementation of this algorithm is accomplished using the received data packets that are marked using the PCN_marking DSCP encoding. In this case, during a measurement interval T, the PCN-egress-node measures the input_PCN_marking_bytes by counting, during the interval T, the PCN_marking bytes.

The incoming_PCN_marking_rate can be then calculated as follows:

```
incoming_PCN_marking_rate =
   N * input_PCN_marking_bytes / T
```

To provide reliable estimation of the encoded information several techniques can be used, see [AtLi01], [AdCa03], [ThCo04], [AnHa06].

If the incoming_PCN_marking_rate is higher than a preconfigured PCN_lower_rate_egress (see Section 3.4 and Figure 4), then the communication path between PCN-ingress-node and PCN-egress-node is considered to be pre-congested.

If probing is used within the whole PCN domain, and when the probe arrives at a PCN_egress_node with PCN marking DSCP encoded then it SHOULD be rejected. If the requesting probe packet is not marked using the PCN_marking DSCP then this requesting probe SHOULD be admitted. In this way it is ensured that the 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 pre- congested router. Note that if an ingress/egress pair aggregated state is not available at the PCN_egress_node, then the PCN_egress node cannot determine whether a PCN_egress_node associated with the ingress-egress aggregate operates in normal state, admission control state or flow termination state. However, even in this case, when a probe packet arrives at the PCN-egress-node, then this request is rejected if the probe packet is PCN_marked. Otherwise (if it is not PCN_marked) it is accepted.

If probing is not used within the whole PCN domain and the request for admission can be accomplished by using an external to PCN, signaling protocol. In this case when the request arrives at a PCN_egress_node that operates in admission control state then the request is rejected. If it operates in Normal state it is accepted.

In any of the situations the PCN-egress-node will have to notify the PCN-ingress-node whether the request for reservation is admitted or rejected.

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

  user  |                  |                 |                  |
  data  |  user data       |                 |                  |
 ------>|----------------->|     user data   |                  |
        |                  |---------------->| user data        |
        |                  |                 |----------------->|
  user  |                  |                 |                  |
  data  |  user data       |                 |                  |
 ------>|----------------->|     user data   | user data        |
        |                  |---------------->S(# marked bytes)  |
        |                  |                 S----------------->|
        |                  |                 S(# unmarked bytes)|
        |                  |                 S----------------->|
        |                  |                 S                  |
request for reservation    |                 S                  |
------->|           probe packet             S                  |
        |----------------------------------->S                  |
        |                  |                 S  probe packet    |
        |                  |                 S----------------->|
        |                  |response                            |
        |<------------------------------------------------------|
 response                  |                 |                  |
 <------|                  |                 |                  |
```

Figure: 2 Admission control based on probing

4.2. Flow Termination for unidirectional flows

The Flow Termination handling method requires the following functionalities.

4.2.1. Operation in the PCN-ingress-nodes

Upon receiving the notification message sent by the PCN-egress-node, the PCN-ingress-node resolves the flow termination 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 LC-PCN traffic class (PHB). This operation is depicted in Figure 3, where the PCN-ingress- node, for each flow (session) to be terminated, receives a notification message.

When the PCN-ingress-node receives the notification message, it starts the termination of the flows within the LC-PCN domain by sending release messages.

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

  user  |                  |                 |                  |
  data  |  user data       |                 |                  |
 ------>|----------------->|     user data   | user data        |
        |                  |---------------->S(# marked bytes)  |
        |                  |                 S----------------->|
        |                  |                 S(# unmarked bytes)|
        |                  |                 S----------------->|Term.
        |             notification for termination              |flow?
        |<-----------------|-----------------S------------------|YES
           release         |                 S                  |
        | -----------------|---------------------------------->|
        |                  |                 |                  |
```

Figure: 3 LC-PCN Flow Termination handling

When the PCN-ingress-node receives the notification message that contains the to be released aggregation bandwidth, it can use it to resize the size of the aggregation size accordingly.

4.2.2. Operation in the PCN-interior-nodes

The PCN-interior-node that operates in a Flow Termination state remarks data packets passing the node. For this remarking, two additional DSCPs can be allocated for each traffic class. One DSCP can be used to indicate that the packet passed a node that operates in the Flow Termination state. This type of DSCP is denoted in this document as PCN_Affected_marking DSCP.

The use of this DSCP type eliminates the possibility that, due to e.g. ECMP (Equal Cost Multiple Paths) enabled routing, the PCN-egress-node either does not detect packets passed a node that operats in the Flow Termination state or erroneously detects packets that actually did not pass the severe congested node. Note that this type of DSCP MUST only be used if all the nodes within the PCN 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 PCN_marking.

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

To provide reliable estimation of the encoded information several techniques can be used, see [AtLi01], [AdCa03], [ThCo04], [AnHa06].

It is RECOMMENDED that the total number of additional (local and experimental) DSCPs needed for flow termination handling within an PCN domain should be as low as possible and it should not exceed the limit of 8.

An example of a remarking procedure is given below. Per supported PHB, the PCN-interior-node can support the operation States depicted in Figure 4, when the admission control based on probing signaling scheme is used in combination with this flow termination type.

```
                 ---------------------------------------------
                |        event B                              |
                |                                             V
             ----------             -------------           ----------
            | Normal   |  event A  | Admission   | event B | Flow      |
            |  state   |---------->| Control     |-------->|Termination|
            |          |           |  state      |         |  state    |
             ----------             -------------           ----------
              ^  ^                       |                     |
              |  |      event C          |                     |
              |   -----------------------                      |
              |         event D                                |
               ------------------------------------------------
```

Figure 4: States of operation, flow termination with congestion notification based on probing

The terms used in Figure 4 are:

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

Flow Termination state: it represents the state related to a certain PHB when the PCN-interior-node is severely congested and ongoing flows need to be terminated in order to solve this congestion.

Admission Control state: state where the load is relatively high, close to the level when pre-congestion can occur

event A: this event occurs when the incoming measured PHB rate is higher than the admission control threshold, i.e., PCN_lower_rate, see Section 4.1, 4.3.

event B: this event occurs when the incoming measured PHB rate is higher than the flow termination threshold, i.e., PCN_upper_rate.

event C: this event occurs when the incoming measured PHB rate is lower or equal to the admission control threshold, i.e., PCN_lower_rate.

event D: this event occurs when the incoming measured PHB rate is lower or equal to the flow termination threshold, PCN_upper_rate.

During flow termination the PCN-interior-node calculates, per traffic class (PHB), the incoming measured PHB rate that is above the flow termination threshold, i.e., denoted in Section 3.3 as PCN_upper_rate, denoted as signaled_overload_rate, in the following way:

- o A PCN-interior-node that operates in Flow Termination state should take into account that packets might be dropped. Therefore, before queuing and eventually dropping packets, the PCN-interior-node should count, per interval T, the total number of original DSCP, PCN_marking DSCP and PCN_Affected_marking DSCP bytes received by the PCN-interior-node that operates in Flow Termination state. Denote this number as total_received_bytes. Note that there are situations when more than one PCN-interior-nodes in the same communication path become severe congested and can operate in Flow Termination state. Therefore, any PCN-interior-node located behind a PCN-interior-node that operates in Flow Termination state, may receive PCN_marking DSCP and PCN_Affected_marking DSCP marked bytes.

- o 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 same method as desribed in Section 4.1.2., see below: measured_overload_rate = ((total_received_bytes) / T) - PCN_upper_rate)

However, the main difference between calculating the signaled overload_rate during Admission Control and Flow Termination is that during the flow termination situation since marking is done in PCN-interior-nodes, the decisions are made at PCN-egress-nodes, and termination of flows are performed by PCN-ingress-nodes, there is a significant delay until the overload information is learned by the PCN-ingress-nodes, see Section 6 of [CsTa05]. The delay consists of the trip time of data packets from the PCN-interior-node that operates in Flow Termination state to the PCN-egress-node, the measurement interval, i.e., T, and the trip time of the notification signaling messages from PCN-egress-node to PCN-ingress-node. Moreover, until the overload decreases at the PCN-interior-node that operates in Flow Termination state, an additional trip time from the PCN-ingress-node to this PCN-interior-node must expire. This is because immediately before receiving the flow termination notification, the PCN-ingress-node 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 PCN-ingress-node to the PCN-interior-node. Without considering the above, PCN-interior-nodes would continue marking the packets until the measured utilization falls below the flow termination threshold. In this way, at the end more flows will be terminated than necessary, i.e., an over-reaction takes place. [CsTa05] provides a solution to this problem, where the PCN-interior-nodes use a sliding window memory to keep track of the signaling overload in a couple of previous measurement intervals. At the end of a measurement intervals, T, before encoding and signaling the overloaded rate as PCN_marking 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 flow termination 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 PCN-egress-node, i.e., signaled_overload_rate is calculated as follows:

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

signaled_overload_rate = measured_overload_rate - Sum_Mi,

where Sum_Mi is calculated as above.
```

Next, the sliding memory is updated as follows:

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

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

```
IF (measured PHB rate > PCN_upper_rate)
THEN
{
  IF (incoming_PCN_marking_rate <> 0) AND
      (incoming_PCN_marking_rate =< Termination_offset_rate)
  THEN
    { signaled_remarked_bytes =
          ((signaled_overload_rate -
           incoming_PCN_marking_rate) * T) / N
    }
  ELSE IF (incoming_PCN_marking_rate =0)
  THEN signaled_remarked_bytes = signaled_overload_rate * T / N
  ELSE IF (incoming_PCN_marking_rate >
              Termination_offset_rate)
  THEN signaled_remarked_bytes =
            ((signaled_overload_rate - Termination_offset_rate)*T)/N
}
```

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

Note that in order to process an overload situation higher than 100% of the maintained PCN_upper_rate all the nodes within the PCN domain must be configured and maintain a scaling parameter, e.g., N used in the above equation, which in combination with the PCN_marking DSCP encoded bytes, e.g., signaled_remarked_bytes, such a high overload situation can be calculated and represented. N can be equal or higher than 1.

Note that when incoming remarked bytes are dropped, the operation of the flow termination 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, see Section 3.3.

All the outgoing packets that are not marked (i.e., by using the PCN_marking DSCP) have to be remarked using the PCN_Affected_marking DSCP.

4.2.3. Operation in the PCN-egress-nodes

When the operation state of the ingress/egress pair aggregate in the PCN_egress_node is the flow termination, see Figure 4, then the implementation of this algorithm is accomplished in the following way.

The PCN-egress-node node applies a predefined policy to solve the flow termination situation, by selecting a number of inter-domain (end-to-end) flows that should be terminated, or forwarded in a lower priority queue.

Some flows, belonging to the same PHB traffic class might get other priority than other flows belonging to the same PHB traffic class. It is considered that this difference in priority can be notified by a signalling protocol and that the edges can store and maintain the priority information releted to each of the end-to-end flows. The terminated flows are selected from the flows having the same PHB traffic class as the PHB of the marked (as PCN_marking DSCP) and PCN_Affected_marking DSCP (when applied in the complete PCN domain) packets and that are belonging to the same ingress/egress pair aggregate.

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

The states of operation in PCN-egress-nodes are similar to the ones described in Section 4.2.2. The definition of the events, see below, is however different than the definition of the events given in Figure 4.

- event A: the PCN-egress-node measures the rate of the incoming "PCN_marking" encoded packets, i.e., incoming_PCN_marking_rate, and compare it with a predefined PCN_lower_rate_egress and to a

PCN_upper_rate_egress in the PCN- egress-node, see Section 3.4. When the incoming_PCN_marking_rate, is higher than the PCN_lower_rate_egress but lower or equal to the flow termination threshold, i.e., PCN_upper_rate_egress then event_A is activated.

- event B: this event is activated depending on which of the solutions described in Section 3.4 are applied at the PCN_egress_node. If the PCN_Affected_marking is used within whole PCN domain, then event B occurs when the PCN_egress_node receives at least one packet that is associated with the ingress/egress aggregate and is PCN_Affected_marking encoded. If the PCN_Affected_marking is not used within whole PCN domain then event B is activated when the incoming_PCN_marking_rate received by the PCN-egress- node is higher than the PCN_upper_rate_egress, see Section 3.4.

- event C: this event occurs when the incoming_PCN_marking_rate received by the PCN-egress-node is lower or equal to PCN_lower_rate_egress, see Section 3.4.

- event D: this event is activated depending on which of the solutions described in Section 3.4 are applied at the PCN_egress_node. If the PCN_Affected_marking is used within whole PCN domain, then event D occurs when the PCN_egress_node does not receives any PCN_affected_marked packets within a predefined amount of time, e.g., one measurement period. If the PCN_Affected_marking is not used within whole PCN domain then event D occurs when the incoming_PCN_marking_rate received by the PCN- egress-node is lower or equal to PCN_upper_rate_egress, see Section 3.4.

An example of the algorithm for calculation of the number of flows associated with each priority class that have to be terminated is explained by the pseudocode below. First, when the PCN-egress-node operates in the flow termination state then the total amount of remarked (PCN_marking DSCP marked) rate, per ingress/egress pair reservation aggregate, associated with the PHB traffic class, say incoming_PCN_marking_rate, is calculated. This rate represents the flow termination bandwidth, per ingress/egress pair, that should be terminated. Note that the below algorithm is performed for each ingress/egress pair reservation aggregate. The incoming_PCN_marking_rate can be then calculated as follows:

```
incoming_PCN_marking_rate =
  N * input_PCN_marking_bytes / T
```

To provide reliable estimation of the encoded information several techniques can be used, see [AtLi01], [AdCa03],[ThCo04], [AnHa06].

If the incoming_congestion_rate is higher than a preconfigured PCN_upper_rate_egress, see Section 3.4 and Figure 4, then it is considered that at least one PCN-interior-node located on a communication path between PCN-ingress-node and PCN-egress-node is considered to operate in the Flow Termination state. The incoming_PCN_marking_rate can be calculated as follows:

```
incoming_PCN_marking_rate =
  N * input_PCN_marking_bytes / T
```

Where, input_PCN_marking_bytes represents the number of marked bytes that arrive at the PCN-egress-node, during one measurement interval T, N is defined as in Section 3.3 and 4.2.1. The term denoted as terminated_bandwidth is a temporal variable representing the total bandwidth that have to be terminated, belonging to the same PHB traffic class. The terminate_flow_bandwidth(priority_class) is the total of bandwidth associated with flows of priority class equal to priority_class. The parameter priority_class is an integer fulfilling

0 < priority_class =< Maximum_priority.

Note that if the PCN domain does not support priority differentiation then the variable Maximum_priority SHOULD be equal to 0.

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

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

For the end-to-end flows (sessions) that have to be terminated, the

PCN-egress-node generates and sends notification message to the PCN-ingress-node to indicate the flow termination in the communication path. Furthermore, for the aggregated sessions that are affected, the PCN-egress-node sends within a notify message that contains the To be released bandwidth, associated with the aggregated reservation state. Note that PCN-egress-node should restore the original DSCP values of the remarked packets, otherwise multiple actions for the same event might occur. However, this value MAY be left in its remarking form if there is an SLA agreement between domains that a downstream domain handles the remarking problem.

4.3. Admission control based on probing for bi-directional flows

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

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

user|                |             |              |               |
data|                |             |              |               |
--->|                | user data   |              |user data      |
    |--------------------------------------------->S (#marked bytes)
    |                |             |              S-------------->|
    |                |             |              S(#unmarked bytes)
    |                |             |              S-------------->|
    |                |             |              S               |
    |                |           probe(re-marked DSCP)            |
    |                |             |              S               |
    |--------------------------------------------->S               |
    |                |             |              S-------------->|
    |                |             |              S               |
    |                |          response(unsuccessful)           |
    |<------------------------------------------------------------|
    |                |             |              S               |
```

Figure 5: Admission control based on probing
for bi-directional admission control (pre-congestion on
path from PCN-ingress-node towards PCN-egress-node)

This procedure is similar to the admission control procedure described in Section 4.1, for the situation that the PCN domain supports probing. The main difference is related to the location of the PCN-interior-ndoe that operates in admission control state, i.e., "forward" path (i.e., path between PCN-ingress-node towards PCN-egress-node) or "reverse" path (i.e., path between PCN- egress-node towards PCN-ingress-node). Figure 5 shows the scenario where the pre-congested PCN-interior-node is located in the "forward" path. The functionality of providing admission control is the same as the one described in Section 4.1, Figure 2. Figure 6 shows the scenario where the pre-congested PCN-interior-node is located in the "reverse" path. The probe packet sent in the "forward" direction will not be affected by the pre-congested PCN-interior-node, while the DSCP value in the IP header of any packet of the "reverse" direction flow and also of the probe packet that carries the sent in the "reverse" direction will be remarked by the pre-congested node. The PCN-ingress-node is in this way notified that a pre-congestion situation occurred in the network and therefore it is able to reject the new initiation of the reservation.

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

user|                |                |           |               |
data|                |                |           |               |
--->|                | user data      |           |               |
    |-------------------------------------------->|user data      |user
    |                |                |           |-------------->|data
    |                |                |           |               |--->
    |                |                |           |               |user
    |                |                |           |               |data
    |                |                |           |               |<---
    |                S                | user data |               |
    |                S  user data     |<--------------------------|
    |   user data    S<---------------|           |               |
    |<---------------S                |           |               |
    |  user data     S                |           |               |
    | (#marked bytes)S                |           |               |
    |<---------------S                |           |               |
    |                S          probe(unmarked DSCP)              |
    |                S             |              |               |
    |----------------S------------------------------------------->|
    |                S         probe(re-marked DSCP)              |
    |                S<-------------------------------------------|
    |<---------------S             |              |               |
```

Figure 6: Admission control based on probing for bi-directional admission control (pre-congestion on path PCN-egress-node towards PCN-ingress-node)

4.4. Flow Termination handling for bi-directional flows

This section describes the flow termination handling operation for bi-directional flows. This flow termination handling operation is similar to the one described in Section 4.2.

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

user|                |             |              |               |
data|    user        |             |              |               |
--->|    data        | user data   |              |user data      |
    |--------------->|             |              S               |
    |                |--------------------------->S (#marked bytes)
    |                |             |              S-------------->|
    |                |             |              S(#unmarked bytes)
    |                |             |              S-------------->|Term
    |                |             |              S               |flow?
    |                |          notification (terminate)          |YES
    |<----------------------------------------------------------- |
    |release (forward)             |              S               |
    |------------------------------------------------------------>|
    |        release (reverese)    |              S               |
    |<------------------------------------------------------------|
    |                |             |              S               |
```

Figure 7: Flow termination handling for bi-directional reservation (congestion on path PCN-ingress-node towards PCN-egress-node)

This procedure is similar to the flow termination handling procedure described in Section 4.2. The main difference is related to the location of the the PCN-interior-ndoe that operates in Flow Termination state, , i.e. "forward" or "reverse" path. When a flow termination congestion occurs on e.g., in the forward path, and when the algorithm terminates flows to solve the flow termination 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 flow termination congestion point on a unidirectional path can be the following:

o the PCN-egress-node should select, if possible, first unidirectional flows instead of bidirectional flows

o the PCN-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.

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

user|                |                |           |               |
data|    user        |                |           |               |
--->|    data        | user data      |           |user data      |
    |--------------->|                |           |               |
    |                |------------------------->|user data      |user
    |                |                |           |-------------->|data
    |                |                |           |               |--->
    |                |                |  user     |               |<---
    |   user data    |                |  data     |<--------------|
    | (#marked bytes)|                S<----------|               |
    |<--------------------------------S           |               |
    | (#unmarked bytes)               S           |               |
Term|<--------------------------------S           |               |
Flow?                |                S           |               |
YES |                |                S           |               |
    |release (forward)                S           |               |
    |------------------------------------------------------------->|
    |        release (reverse)        S           |               |
    |<-------------------------------------------------------------|
    |                |                S           |               |
```

Figure 8: Flow termination handling for bi-directional reservation (flow termination congestion on path PCN-egress-node towards PCN-ingress-node)

Furthermore, a special case of this operation is associated to the Flow Termination situation occurring simultaneously on the forward and reverse paths. An example of this operation is given below. Consider that the PCN-egress-node selects a number of bi-directional flows to be terminated, see Figure 9. In this case the PCN-egress-node will send for each bi-directional flows a notification message to PCN-ingress-node. If the PCN-ingress-node receives these notification messages and its operational state (associated with reverse path) is in the Flow Termination state (see Figure 4), then the PCN-ingress-node operates in the following way:

```
PCN-ingress-node  PCN-interior-node  PCN-interior-node   PCN-egress-node

user|                |                |           |               |
data|    user        |                |           |               |
--->|    data        | #unmarked bytes|           |               |
    |--------------->S #marked bytes  |           |               |
    |                S-------------------------->|               |
    |                |                |           |-------------->|data
    |                |                |           |               |--->
    |                |                |           |             Term.?
    |            NOTIFY               |           |               |Yes
    |<-----------------------------------------------------------|
    |                |                |           |               |data
    |                |                |  user     |               |<---
    |   user data    |                |  data     |<--------------|
    | (#marked bytes)|                S<----------|               |
    |<--------------------------------S           |               |
    | (#unmarked bytes)               S           |               |
Term|<--------------------------------S           |               |
Flow?                |                S           |               |
YES |                |                S           |               |
    |release (forward)                S           |               |
    |----------------------------------------------------------->|
    |        release (reverse)        S           |               |
    |<-----------------------------------------------------------|
```

Figure 9: Flow termination handling for bi-directional reservation (flow termination congestion on both forward and reverse direction)

- For each notification message, the PCN-ingress-node should identify the bidirectional flows that have to be terminated.

- The PCN-ingress-node 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 flows that received a (flow termination) notification message.

- Furthermore, using the received marked packets (from the reverse path) the PCN-ingress-node will calculate, using the algorithm used by an PCN-egress-node and described in Section 4.2.3, the total bandwidth that has to be terminated in order to solve the flow termination congestion in the reverse path direction, say "marked_reverse_bandwidth".

- o The PCN-ingress-node then calculates the bandwidth of the additional flows that have to be terminated, say "additional_reverse_bandwidth", in order to solve the flow termination congestion in the reverse direction, by taking into account:

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

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

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

- o PCN-ingress-node terminates the flows that experienced a severe congestion in the "forward" path and received a (flow termination) notification message

- o If possible the PCN-ingress-node should terminate unidirectional flows that are using the same egress-ingress reverse direction communication path to satisfy the release of a total bandiwtdh up equal to the: "additional_reverse_bandwidth".

- o If the number of required uni-directional flows (to satisfy the above issue) is not available, then a number of bi-directional flows that are using the same egress-ingress reverse direction communication path may be selected for flow termination in order to satisfy the release of a total bandiwtdh equal up to the: "additional_reverse_bandwidth". Note that using the guidelines given in above, 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.

- o Furthermore, the PCN-egress-node includes the to be released aggregated bandwidth value in one of the notification messages.

- o The PCN-ingress-node receives this notification message and reads the value of the carried to be released aggregated bandwidth.

The size of the aggregated reservation state can be reduced in the "forward" and "reverse" by using the received to be reduced values the aggregated bandwidth in "forward" and "reverese" directions. Figure 7 shows the scenario where the severe congested node is located in the "forward" path. This scenario is very similar to the flow termination handling scenario described in Section 4.2. The difference is related to the release procedure, which is accomplished in both directions "forward" and "reverse". Figure 8 shows the scenario where the severe congested node is located in the "reverse" path. The main difference between this scenario and the scenario shown in Figure 7 is that no notification messages have to be generated by the PCN-egress-node. This is because the (#marked and #unmarked) user data is arriving at the PCN-ingress-node. The PCN-ingress-node will be able to calculate the number of flows that have to be terminated or forwarded in a lower priority queue.

5. Security Considerations

The security considerations associated with this document are similar to the one described in [Eard07].

6. IANA Considerations

To be Added

7. Acknowledgements

To be Added

8. Informative References

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

[AnHa06] Lachlan, A. and S. Hanly, "The Estimation Error of Adaptive Deterministic Packet Marking", 44th Annual Allerton Conference on Communication, Control and Computing, , 2006.

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

[Babi07] Babiarz, J. and et. al., "Three State PCN Marking", draft-babiarz-pcn-3sm-00 (work in progress), , June 2007.

[Bernet99] Bernett, Y., Yavatkar, R., Ford, P., Baker, F., Zhang, L., Speer, M., and R. Braden, "Interoperation of RSVP/Intserv and Diffserv Networks", Work in Progress , March 1999.

[Berson97] Berson, S. and R. Vincent, "Aggregation of Internet Integrated Services State", Work in Progress, , December 1997.

[CL-ARCH] Briscoe, B. and et. al., "An edge-to-edge Deployment model for pre-congestion notification: Admission control over a Diffserv region", , October 2006.

[CL-PHB] Briscoe, B. and et. al., "Pre-congestion notification marking", , October 2006.

[Char07] Charny, A. and et. al., "Pre-Congestion Notification Using Single Marking for Admission and Termination", draft-charny-pcn-single-marking-02 (work in progress), , July 2007.

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

[Eard07] Eardley, P., "Pre-Congestion Notification Architecture", draft-ietf-pcn-architecture-01 (work in progress), , October 2007.

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

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

[RMD] Bader, A., "RMD-QOSM: The resource management in Diffserv QoS Model", draft-ietf-nsis-rmd-11.txt (work in progress), , March 2007.

[Stoica99] Stoica, I. and et. al., "Per Hop Behaviors Based on Dynamic Packet States", Work in Progress , February 1999.

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

[Westberg00] Westberg, L. and et. al., "Load Control of Real-Time Traffic", IETF Work in Progress , April 2000.

Authors' Addresses

Lars Westberg
Ericsson
Torshamnsgatan 23
SE-164 80 Stockholm
Sweden

Email: Lars.westberg@ericsson.com

Anurag Bhargava
Ericsson
920 Main Campus Dr., Suite 500
Raleigh, NC 27606
USA

Phone: +1 919 472 9964
Email: anurag.bhargava@ericsson.com

Attila Bader
Ericsson
Laborc 1
Budapest
Hungary

Email: Attila.Bader@ericsson.com

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enscede
Netherlands

Email: g.karagiannis@ewi.utwente.nl

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).