PCN
Internet-Draft
Intended status: Standards Track
Expires: May 7, 2009

L. Westberg
Ericsson
A. Bhargava

A. Bader
Ericsson
G. Karagiannis
University of Twente
H. Mekkes
Researcher

November 3, 2008

LC-PCN: The Load Control PCN Solution
draft-westberg-pcn-load-control-05

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 can be performed using two methods:

- 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.

- 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

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. LC-PCN can support the ingress-egress-aggregate (i.e., trunk/pipe) bandwidth management model as well as the HOSE bandwidth management model.

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 in the core 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.

We believe that the LC-PCN features supported by, at least, PCN-interior-nodes can be combined with features supported by the above listed concepts.

This document aims to develop a common framework that could be used with external protocols. LC-PCN can support the ingress-egress-aggregate (i.e., trunk/pipe) bandwidth management model as well as the HOSE bandwidth management model [DuGo99]. In this document the term HOSE is referring to the aggregation of incoming traffic from all ingress edges, which is associated with one traffic class, i.e., PHB, towards one egress edge. This type of HOSE model is equivalent to the Multiple to Point (MP2P) type of aggregation.

The HOSE model ensures bandwidth limits without the need of maintaining per each ingress and egress pair ingress-egress-aggregated states. In this case all edges maintain one aggregated state per each traffic class, i.e., PHB, used in the PCN domain. This version of the draft focuses on how LC-PCN can support the ingress-egress-aggregate (i.e., trunk/pipe) bandwdith management model. Furthermore, it emphasizes which modifications have to be realized in order to also support the HOSE bandwdith management model.

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 [Eard08] are used.

3. LC-PCN Overview

Load Control PCN (LC-PCN) is achieved by two actions: Admission Control 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

Admission control can be accomplished in LC-PCN in two ways:

- o Admission control based on data marking: whereby in congestion situations, the admission control is accomplished using excess rate marking and metering to detect and to decide either a new flow request should be accepted or denied.

- o Admission control based on probing: where probing is required to accomplish the admission control procedure.

Note that the two admission control features can be used either independently or combined. In the ingress-egress-aggregate model the Admission Control features can be applied to flows that are aggregated between PCN-ingress-nodes and PCN-egress-nodes and use the same traffic class, i.e., use the same PHB. In this way edge-to-edge (i.e., ingress-egress) pair PCN aggregates can be maintained by PCN-ingress-nodes and PCN-egress-nodes. In the HOSE model the Admission Control features can be applied to flows that are belonging to the same traffic classs, i.e., use the same PHB. Note that these flows can start from different PCN-ingress-nodes and use different PCN-egress-nodes. Two PCN-domain-wide constraints are used. One of them is denoted as "N", used to indicate the proportionality between the measured out of profile packets (or bytes) and the remarked packets (or bytes). If "N" is used in the algorithm, then it must have the same value in all Diffserv nodes that use this mechanism. The parameter N is higher or equal to 1 (N >= 1).

Another PCN-domain-wide constraint, see [Char07], has to be used on the ratio U between the configured-admissible-rate on a link and the

level of PCN load on the link that should trigger the Flow Termination. This level represents the configured-termination-rate, which is not explicitly configured on the PCN_interior node. The value is typically set to U = 1,2, see [Char07].

Furthermore, it is important to note that in this draft we denote the not congested PCN packets (or bytes) as PCN unmarked packets (or bytes).

3.1.1. Admission control based on data marking

The admission control based on data marking is using features located at the PCN_ingress_edge, PCN-interior-node and PCN-egress-node. This type of admission control can only be used when the ingress-egress-aggregate (trunk/pipe) model is used.

3.1.1.1. PCN-interior-node features

The PCN-interior-node performs measurements on the PHB aggregated PCN traffic. When the PCN-interior-node detects that the measured PHB aggregated PCN traffic is higher than a preconfigured threshold, say configured-admissible-rate, then it is considered that the PCN-interior-node changes operational state from Normal state to Admission Control state, see Section 3.3.1. Furthermore, the measured PHB aggregated PCN traffic rate that is above the configured-admissible-rate is considered to be excess rate, which is marked using PCN_marking.

This can be accomplished using different metering and marking features. It is important to note that the excess rate measurements SHOULD be done before a queuing mechanism used by a PCN-interior-node, drop packets before/during buffer overflow. The constant N should be used such that the marked excess rate can represent also high levels of excess rate. This means that before marking the excess rate, the measured excess rate should be divided by N (when N >= 1). This can be e.g., implemented by marking every N-th packet (or byte) instead of marking each packet (or byte).

The PCN_marking SHOULD be done after the queuing mechanism drops the packets before/during buffer overflow. Several implementation alternatives of this algorithm are possible. One implementation alternative can be based on the algorithms discussed in [Char07]. In particular, a token bucket can be used with the rate configured with the rate equal to configured-admissible-rate. However, the token bucket specification should satisfy the functionality used for rate measurements and marking, which is described below as another implementation alternative. In particular, during admission control (and flow termination) the token bucket must mark every N packets

instead of marking each packet. Furthermore, the PCN_marking encoded packets must not be preferentially dropped. Instead, the typical random dropping of packets should be applied. Furthermore, when operating in optimisation mode, the token bucket must use an additional threshold, i.e., (U * configured_admissible_rate). When above this threshold all packets that are not being PCN_marking encoded must be marked as PCN_Affected_Marking encoded.

Other implementation alternatives can e.g., be based on rate measurements and marking. In particular, the PCN-interior-nodes packets are using the PCN_marking, whenever the measured PHB aggregated PCN traffic rate exceeds a pre-configured rate threshold denoted as configurable-admissible-rate.

It is important to note that the PCN_marking encoded packets SHOULD NOT be preferentially dropped by queuing mechanisms in PCN-interior-nodes. This can be accomplished using the following alternative. All packets, PCN marked and PCN unmarked (and PCN_Affected_Marking encoded, when the affected marking solution is supported) use one queue and in case of overload the packets are dropped randomly independently of either they are PCN_marking or PCN unmarked encoded.

3.1.1.2. PCN-egress-node features

The PCN-egress-node measures the rate of the received PHB aggregated PCN unmarked and PCN_marking encoded packets. Based on these measurements, the PCN-egress-node can use a similar functionality as the one specified in [Char07] and [CL-ARCH] to calculate the Congestion Level Estimate (CLE), which is the fraction of the marked traffic received from one PCN-ingress-node.

Note that the marked traffic used in the calculation of CLE is equal to the product of N and the measured marked traffic received by the PCN-egress-node. In pseudo code notation the value of the CLE can be calculated as follows:

```
  CLE = N * PCN_marking_rate/Total_received_rate,
    where: Total_received_rate = PCN_marking_rate + PCN_unmarked_rate

  IF (PCN_Affected_Marking encoding is used) THEN
      PCN_unmarked_rate = PCN_Affected_Marking_rate,
  ELSE
      PCN unmarked_rate = the rate of the not congested PCN packets
                          (or bytes).
```

If the value of CLE is higher than a certain value, e.g., 1%, then the PCN-egress-node is changing its operational state from Normal

state to Admission Control state. By using an external to PCN, signaling protocol the admission control procedure is accomplished by using a combination of the PCN operational state of the PCN-egress-node and an admission control request provided by the external to PCN, signaling protocol. When the admission control 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.

3.1.1.3. PCN-ingress-node features

If the external to PCN signaling protocol is also used by the PCN-ingress-node, then the PCN-ingress-node SHOULD be informed that an admission control request has been admitted or rejected by the PCN-egress-node. If the PCN-ingress-node is notified that the admission request is rejected, then the PCN-ingress-node rejects the admission control request. Otherwise it is accepted.

3.1.2. 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. The main reason of why this admission control feature should be used is to solve the possible ECMP (Equal Cost Multi-Path) issue. Furthermore, this feature can provide admission control support even when the edge-to-edge pair PCN aggregate is not yet initiated at one of the edges. This admission control type can be used to support both bandwidth management models, ingress-egress-aggregate and HOSE models.

3.1.2.1. PCN-interior-node features

The PCN-interior-node features that are used to detect the PCN operational states are the same as the ones described in Section 3.1.1.1. In this scenario an IP packet is used as a probe packet, meaning that the DSCP (and/or ECN) field, see Section 3.4, in the header of the IP packet is re-marked when the measured PHB aggregated PCN traffic rate exceeds a predefined congestion threshold, i.e, configured-admissible-rate. Note that a message used by an external, to PCN, on path signaling protocol, e.g., RSVP, can be used as a probe packet.

The PCN-interior-nodes SHOULD detect a probe packet by observing specific IP header information. Note that defining the IP header information that can be used for this purpose is out of the scope of this document. An example of such information could be the Router Alert option, which is carried by the IP packet data packet. Note that a PCN-ingress-node sets the Router Alert option of all packets

that are used as probe packets. This also holds for signaling protocol messages (e.g. RSVP PATH message) that are used by LC-PCN as 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 than the user packets.

An alternative solution that can be used to mark the probes is to apply an additional encoding/marking state and use the CL (Controlled Load) based admission marking [CL-PHB], where all packets are marked using the additional encoding/marking state when the PCN-interior-node operates in admission control state.

The PCN-interior-node has to PCN_marking encode the probe packet if it is operating in Admission Control state (or Flow Termination state). Otherwise the probe packet does not change its encoding state.

3.1.2.2. PCN-egress-node features

The PCN-egress-node measures the rate of the received PHB aggregated PCN unmarked and PCN_marking encoded packets. When the probe packet arrives at the PCN-egress-node that is belonging to a certain edge-to-edge pair PCN aggregate, and it is PCN_marking encoded then the request is rejected. Otherwise it is accepted.

Note that if an edge-to-edge 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 edge-to-edge pair PCN 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 should be rejected if the probe packet is PCN_marking encoded. Otherwise, i.e., if the probe packet is not PCN_marking encoded, it should be accepted.

3.1.2.3. PCN-ingress-node features

The PCN-ingress-node if needed is modifying specific IP header information In probe packets to give the possibility to the PCN-interior-nodes to make a distinction between probe packets and normal packets. Note that defining the IP header information that can be used for this purpose is out of the scope of this document. An example of such information could be the Router Alert option.

Furthermore, if an external to PCN signaling protocol is also used by the PCN-ingress-node, then the PCN-ingress-node SHOULD be informed that an admission control request has been admitted or rejected by the PCN-egress-node. If the PCN-ingress-node is notified that the admission request is rejected, then the PCN-ingress-node rejects the

admission control request. Otherwise it is accepted.

3.1.3. ECMP solution

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. This is because a probe packet can be PCN_marking encoded only by congested PCN-interior-nodes. 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 ingress-egress aggregate that detects a congestion.

Note that the ECMP problem can also occur when the HOSE model is used. In this case the ECMP problem is caused by flows that are belonging to the same traffic class aggregate that detects congestion but they are not passing through a congested PCN-interior-node.

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 remarking approach where the PCN_marking is proportional to the excess rate. The Flow Termination features, similar to the Admission Control features, can be applied to flows that are aggregated between PCN-ingress-nodes and PCN-egress-nodes and use the same traffic class, i.e., use the same PHB. In this way edge-to-edge aggregates can be maintained by PCN-ingress-nodes and PCN-egress-nodes. In the HOSE model the flow termination features can be applied to flows that are belonging to the same traffic class, i.e., use the same PHB. Note that these flows can start from different PCN-ingress-nodes and use different PCN-egress-nodes.

Furthermore, the "N" and "U" PCN-domain-wide constraints, specified in 3.1 are also used during Flow Termination.

Moreover, it is important to note that in this draft we denote the not congested PCN packets (or bytes) as PCN unmarked packets (or bytes).

3.2.1. PCN-interior-node

The PCN-interior-nodes can support two types of Flow Termination modes, a base mode and an optimization mode. The Flow Termination base mode that is supported by the PCN-interior-nodes can be accomplished using the admission control features described in Section 3.1.1. The optimisation mode is used to support the ECMP solution and the HOSE model.

The main addition that this optimization mode requires is that an additional operational state has to be maintained by the PCN-interior-node, i.e., a Flow Termination state, see Section 3.3.1. In particular, when the measured PHB aggregated PCN traffic is higher than the threshold equal to (U * configured-admissible-rate), then the PCN-interior-node changes from the Admission Control state to the Flow Termination state.

3.2.2. PCN-egress-node

The PCN-egress-node measures the rate of the received PHB aggregated PCN unmarked and PCN_marking encoded packets (or bytes).

However, inaccuracies in excess rate measurements might occur due to the delay between the metering and marking events that occur at the PCN-interior-nodes, the decisions that are made at PCN-egress-nodes, and the termination of flows that are performed by PCN-ingress-nodes, see Section 6 of [CsTa05].

In order to reduce these excess rate inaccuracies a sliding window method is used to keep track of the bandwidth to be terminated, calculated in a number of previous measurement intervals.

Depending on whether the PCN_Affected_Marking encoding is used in the PCN domain, the Flow Termination can be activated/triggered using two alternatives. When the PCN_Affected_Marking encoding is used then the Flow Termination state is activated/triggered when either at least one PCN_Affected_Marking packet is received by the PCN-egress-node OR when the ratio value of the N* PCN_marking encoded and the sum of the PCN_Affected_Marking and PCN_marking encoded packets (or bytes) is higher than the value of (U - 1), see [Char07]. In pseudo code form notation this can be written as:

```
IF (((N*PCN_marking_rate /(PCN_marking_rate+PCN_Affected_Marking_rate)
     > (U - 1)))
     OR (At least one PCN_Affected_Marking encoded packet arrived))
THEN
    PCN_egress_node Go TO flow termination state.
```

If the PCN_Affected_Marking encoding is not used within the PCN domain then the PCN-egress-node uses a similar functionality as discussed in [Char07] to activate/trigger the Flow Termination. This trigger is computed from the ratio of the N* PCN_marking encoded and the sum of the PCN_unmarked and PCN_marking encoded packets (or bytes).

The trigger is detected when the above given ratio is higher than the value of (U - 1), see [Char07]. In pseudo code form notation this can be written as:

```
   IF ((N*PCN_marking_rate / (PCN_marking_rate + PCN_unmarked_rate)
        > (U - 1))
   THEN
    PCN_egress_node Go TO flow termination state.
```

When this trigger is detected then the PCN-egress-node has to calculate the value of the configured-termination-rate-egress, which depends among others on the value of the N*PCN_marking_rate. The calculation of this value is described below using pseudo code notation:

```
IF (N*PCN_marking_rate > maximum supported bandwidth)
THEN
    configured-termination-rate-egress =
            (U - 1)*Total_received_rate
ELSE
    configured-termination-rate-egress = (U-1) *
            (PCN_unmarked_rate - ((N-1)*PCN_marking_rate))

where: Total_received_rate =
            PCN_marking_rate + PCN_unmarked_rate

IF (PCN_Affected_Marking encoding is used)
THEN
    PCN_unmarked_rate = PCN_Affected_Marking_rate
ELSE
    PCN_unmarked_rate = the rate of the not congested PCN packets
                        (or bytes).
```

The N * (measured excess rate) that is above this threshold, is used to calculate the number of flows to be terminated, such that the excess rate is severely reduced until it drops below the Flow Termination trigger.

Note that in the ingress-egress-aggregate model the excess rate and the flows to be terminated are associated with the same edge-to-edge (i.e., ingress-egress) pair PCN aggregate and with the same traffic class, i.e., PHB. In the HOSE model the excess rate and the flows to be terminated are associated with the same traffic class, i.e., PHB. The PCN-egress-node needs to store for each flow the address, e.g., IP address and port number, of the PCN-ingress-node from where the particular flow passed before arriving to the PCN-egress-node. This can be for example done by using information that is carried by the external protocols used in combination with the LC-PCN solution. For the flows that should be terminated, the PCN-egress-node informs the associated PCN-ingress-node to terminate them. If the PCN domain uses thet ingress-egress-aggregate model and if the PCN-egress-node receives any admission flow request, belonging to a ingress-egress-aggregate state operating in flow termination state then the request must be rejected.

3.2.3. PCN-ingress-node

The flows that are Flow Termination notified by the PCN_egress-node have to be terminated by the PCN_ingress-node. Furthermore, depending on the used policy, the packets related to the flows that have to be terminated are either blocked or shifted to an alternative LC-PCN traffic class, i.e., PHB. Moreover, depending on the used policy, the PCN-ingress-node could reject all new flow admission requests that are associated with the same edge-to-edge pair PCN aggregate until no other requests to terminate flows are received from PCN-egress-nodes. In addition to the above, the PCN_ingress-node informs the associated flow sender about the occurred exceptional/severe congestion. The same features are used when the HOSE model is applied in the PCN domain. The only difference is that a policy that rejects all new flow admission requests cannot be used.

3.2.4. ECMP solution

In order to solve the ECMP issue that may occur during Flow Termination operational state, the LC-PCN solution could use an additional PCN marking encoding approach, denoted as: PCN_Affected_Marking.

This means that the descriptions of Section 3.2.1 and 3.2.2 have to be slightly modified.

Regarding the description provided in section 3.2.1, the PCN_Affected_Marking is used in the PCN-interior-node in the following way. When the measured PHB aggregated PCN traffic is higher than the threshold equal to (U*configured-admissible-rate), then the PCN-interior-node changes from the Admission Control state

to Flow Termination state, see Section 3.3.2. In Flow Termination state, the PCN-interior-node encodes all PCN unmarked (i.e., not congested PCN encoded) packets that are passing through the PCN-interior-node by using the PCN_Affected_Marking.

Regarding the PCN-egress-node description provided in section 3.2.2, the Flow Termination is triggered/activated at the moment that either at least one PCN_Affected_Marking packet is received by the PCN-egress-node OR when the ratio value of the N* PCN_marking encoded and the sum of the PCN_Affected_Marking and PCN_marking encoded packets (or bytes) is higher than the value of (U - 1), see [Char07]. In pseudo code form notation this can be written as:

```
IF       N * PCN_marking_rate
     ----------------------------------------     > (U -1) OR
 (PCN marking_rate + PCN_Affected_Marking_rate)

     (at least one PCN_Affected_Marking encoded packet arrived)
THEN
    PCN_egress_node Go TO flow termination state.
```

Furthermore, the PCN-egress-node uses the PCN_Affected_Marking to identify which flows were affected by the exceptional/severe congestion. In this way the PCN-egress-node, when operating in Flow Termination state, is able to terminate only the flows that received one or more PCN_Affected_Marking packets. The same features are used when the HOSE model is applied in the PCN domain. The main difference is related to the fact that the solution takes into account the fact that the excess rate and the flows to be terminated are associated with the same traffic class, i.e., PHB, but they are not required to belong to the same ingress-egress-aggregate, see Section 3.2.2.

3.3. Operational states in LC-PCN

This section describes the LC-PCN operational states that are used to identify when and how a PCN node is triggered to either remain or change into an operational state, i.e., Normal, Admission Control and Flow Termination.

3.3.1. Operational states in PCN-interior-nodes

Per each PHB supported with the PCN domain, the PCN-interior-node supports the operational states diagram depicted in Figure 2.

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

Figure 2: States of Operation

The terms used in Figure 2 and applied for PCN-interior-nodes are:

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

* Flow Termination state: this state is applied when the optimization mode solution is applied, when the ECMP solution described in Section 3.2.4 is used and when the HOSE model is used instead of the ingress-egress-aggregate model. This state represents the state related to a certain PHB when the PCN-interior-node is severely congested.

* 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 measured PHB aggregated PCN traffic is higher than the configured-admissible-rate. The measured PHB aggregated PCN traffic rate that is above the configured-admissible-rate is considered to be excess rate, which is encoded using PCN_marking.

* event B: this event is applied when the optimization solution is applied, when the ECMP solution is used and when the HOSE model is used instead of the ingress-egress-aggregate model. This event occurs when the measured PHB aggregated PCN traffic is higher than the threshold equal to (U * configured-admissible-rate).

* event C: this event occurs when the measured PHB aggregated PCN traffic is equal or lower than the configured-admissible-rate.

* event D: this event is only applied either when the optimization solution is applied, when the ECMP solution is used and when the HOSE model is used instead of the ingress-egress-aggregate model. This

event occurs when the measured PHB aggregated PCN traffic is equal or lower than the threshold equal to (U * configured-admissible-rate).

3.3.2. Operational states in PCN-egress-nodes

Per each PHB supported with the PCN domain, the PCN-egress-node supports the operational states diagram depicted in Figure 2. In case that the PCN domain supports the ingress-egress-aggregate model then the operational states are related to one ingress - egress pair of nodes. In case the HOSE model is used, then the operational states are related to one traficc class, i.e., PHB.

The terms used in Figure 2 and applied for PCN-egress-nodes 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 edge-to-edge (ingress-egress) pair PCN aggregate to identify the situation that a severe/exceptional event occurred and ongoing flows need to be terminated in order to solve this severe congestion.

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

* event A: this event is activated when the Congestion Level Estimate (CLE) is higher than a predefined value, e.g., 1%, see [Char07]. CLE is the ratio of the N* PCN_marked traffic rate, which is calculated as an EWMA and the total received rate, which is also calculated as an EWMA. In pseudo code notation the value of the CLE can be calculated as follows:

```
CLE = N * PCN_marking_rate/Total_received_rate,

where: Total_received_rate = PCN_marking_rate + PCN_unmarked_rate

IF (PCN_Affected_Marking encoding is used)
THEN
    PCN_unmarked_rate = PCN_Affected_Marking_rate,
ELSE
    PCN unmarked_rate = the rate of the not congested PCN packets
                        (or bytes)
```

* event B: this event can be activated using two alternatives, depending on whether the PCN_Affected_Marking encoding is used. When the PCN_Affected_Marking encoding is used then the Flow Termination state is activated/triggered when either at least one

PCN_Affected_Marking packet is received by the PCN-egress-node OR when the ratio value of the N* PCN_marking encoded and the sum of the PCN_Affected_Marking and PCN_marking encoded packets (or bytes) is higher than the value of (U - 1), see [Char07]. In pseudo code form notation this can be written as: In pseudo code form notation this can be written as:

```
IF          N * PCN_marking_rate
    -------------------------------   > (U -1)  OR
(PCN marking_rate + PCN_Affected_Marking_rate)

      (at least one PCN_Affected_Marking encoded packet arrived)
THEN
    activate event B
```

If the PCN_Affected_Marking encoding is not used within the PCN domain then the PCN-egress-node uses a similar functionality as discussed in [Char07] to activate/trigger the Flow Termination. The trigger is detected when the ratio of the N* PCN_marking encoded and the sum of the PCN_unmarked and PCN_marking encoded packets (or bytes) is higher than the value of (U - 1), see [Char07]. In pseudo code form notation this can be written as:

```
IF ((N * PCN_marking_rate / (PCN_marking_rate + PCN_unmarked_rate)
      > (U - 1))
THEN
    activate event B

IF (PCN_Affected_Marking encoding is not used)
THEN
    PCN unmarked_rate = the rate of the not congested PCN packets
                        (or bytes)
```

When this trigger is detected then the PCN-egress-node has to calculate the value of the configured-termination-rate-egress, which depends among others on the value of the N * PCN_marking_rate. The calculation of this value is described below using pseudo code notation:

```
IF (N * PCN_marking_rate > maximum supported bandwidth)
THEN
    configured-termination-rate-egress =
                      (U-1) * Total_received_rate
ELSE
    configured-termination-rate-egress =
           (U-1)* (PCN_unmarked_rate - ((N-1) *
                                 PCN_marking_rate))

where: Total_received_rate =
              PCN_marking_rate + PCN_unmarked_rate

IF (PCN_Affected_Marking encoding is used)
THEN
    PCN_unmarked_rate = PCN_Affected_Marking_rate
ELSE
    PCN unmarked_rate = the rate of the not congested PCN
                        packets (or bytes)
```

* event C: this event occurs when the CLE is lower or equal than the predefined value used to trigger event A.

* event D: this event can be activated using two alternatives, depending on whether the PCN_Affected_Marking encoding is used. When the PCN_Affected_Marking is used, then the psuedo code that describes the detection/activation of the trigger is given below:

```
IF    N * PCN_marking_rate
   -----------------------------------------------  <= (U -1)
    (PCN_marking_rate + PCN_Affected_Marking_rate)

       AND (NO PCN_Affected_Marking encoded packet(s) arrived))
THEN
     activate event D
```

When the PCN_Affected_Marking is not used, then the psuedo code that describes the detection/activation of the trigger is given below:

```
   IF         N * PCN_marking_rate
         ------------------------------      <= (U - 1)
     (PCN_marking_rate + PCN_unmarked_rate)

   THEN
          activate event D

    IF (PCN_Affected_Marking encoding is used)
    THEN
        PCN_unmarked_rate = PCN_Affected_marking_rate,
     ELSE
        PCN unmarked_rate = the rate of the not congested PCN
                              packets (or bytes)
```

3.3.3. Operational states in PCN-ingress-nodes

Per each edge-to-edge pair of PCN aggregates the PCN-ingress-nodes support the same operational states diagram as depicted in Figure 2. In case that the PCN domain supports the ingress-egress-aggregate model then the operational states are related to one ingress - egress pair of nodes. In case the HOSE model is used, then the operational states are related to one traficc class, i.e., PHB.

The terms used in Figure 2 and applied for PCN-ingress-nodes are:

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

* Flow Termination state: it represents the state used to identify the situation that a severe/exceptional event occurred and ongoing flows need to be terminated in order to solve this severe congestion. In Flow Termination, the PCN-ingress-node MAY block all new admission flow requests that are associated with the same edge-to-edge pair of PCN aggregates. This depends on the policy used by the PCN-ingress-node.

* Admission Control state: state where the load is relatively high, close to the level when pre-congestion can occur. The PCN-ingress-node rejects a flow that is requesting admission to the PCN domain.

* event A: this event occurs when the PCN-ingress-node receives a response from the PCN-ingress-node that a flow that is requesting admission to the PCN domain is rejected.

* event B: this event occurs when the PCN-ingress-node receives one response from the PCN-ingress-node that a flow has to be terminated due to the fact that the PCN-ingress-node operates in the Flow Termination operational state.

* event C: this event occurs after the PCN-ingress-node rejected the flow that was requesting admission and informed the flow sender about it.

* event D: this event is activated either after the moment that the notified flows to be terminated are terminated or when the PCN-ingress-node does not receive anymore responses from the PCN-egress-node that flows have to be terminated. A policy that is available at the PCN-ingress-node SHOULD select one of the ways described above to activate event D.

3.4. Encoding of PCN traffic

The encoding that can be used for LC-PCN can be based either on DSCP or on a combination between ECN and DSCP IP fields. In the current version of the draft it is assumed that the encoding is based on only the DSCP IP field. In particular, the encoding can be accomplished in the following way. The PCN traffic can be distinguished from the non PCN traffic by using a first additional DSCP, say not_congested_PCN_DSCP, to identify the not congested PCN traffic.

The single marking state used during PCN_marking encoding can use a second additional DSCP, say PCN_marking_DSCP. When the PCN_Affected_Marking is used then an additional third DSCP is needed, say PCN_Affected_Marking_DSCP.

The first, second and third additional DSCP values are representing DSCP values that are assigned by IANA as DSCP experimental values.

It is important to note that 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 not_congested_PCN_DSCP, PCN_marking DSCP and PCN_Affected_Marking

DSCP to other DSCPs, say not_congested_PCN_new_DSCP, PCN_marking_new_DSCP and PCN_Affected_Marking_new_DSCP, respectively, might be necessary. This is because the neighbor PCN operator may use different Diffserv mapping schemes.

When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-egress-nodes and PCN-ingress-nodes SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur. 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. When no trust relationship exists between multiple neighboring PCN domains then the PCN-ingress-nodes SHOULD PCN encode the incoming traffic that is used as incoming PCN traffic using the not congested PCN DSCP.

4. LC-PCN detailed description

This section describes the details of the used LC-PCN algorithms. Section 4.1, 4.2 and 4.3 describe the "Admission control based on data marking", "Admission control based on probing" and "Flow Termination" scenario, respectively, for the situation that the end-to-end sessions are using unidirectional reservations. Section 4.4 describes the two admission control procedures and Section 4.5 describes the flow termination scenario for the situation that the end-to-end sessions are using bi-directional reservations.

4.1. Admission control based on data marking for unidirectional flows

This type of admission control uses excess rate marking and metering to provide admission control for unidirectional flows. In pre-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. This type of admission control can be used only when the ingress-egress-aggregate model is applied within the PCN domain.

4.1.1. Operation in PCN-interior-nodes

The PCN-interior-node performs measurements on the PHB aggregated PCN traffic, see Figure 3, and changes operational state from Normal to Admission Control state when the event A trigger occurs, see Section 3.3.1.

As mentioned in Section 3.1.1.1, the measured aggregated PCN traffic rate that is above the configured-admissible-rate is considered to be

excess rate, which is marked using PCN_marking. When the PCN-interior-node operates in Admission Control state and the configured-admissible-rate is exceeded then PCN unmarked packets are proportionally to the excess rate re-marked, using the PCN_marking encoding, see event A, in Section 3.3.1.

The above described functionalities can be accomplished using different metering and marking features. Several implementation alternatives of this algorithms are possible. One implementation alternative can be based on the algorithms discussed in [Char07]. In particular, a token bucket can be used with the rate configured with the rate equal to configured-admissible-rate. However, the token bucket specification SHOULD satisfy the functionality used for rate measurements and marking, which is described below as another implementation alternative. In particular, during admission control (and flow termination) the token bucket must mark every N packets instead of marking each packet. Furthermore, the PCN_marking encoded packets SHOULD NOT be preferentially dropped. This can be accomplished using the following alternative. All packets, PCN marked and PCN unmarked (and PCN_Affected_Marking encoded, when the affected marking solution is supported) use one queue and in case of overload the packets are dropped randomly independently of either they are PCN_marking or PCN unmarked encoded. Furthermore, when operating in optimisation mode, the token bucket must use an additional threshold, i.e., (U * configured_admissible_rate). When above this threshold all packets that are not being PCN_marking encoded must be marked as PCN_Affected_Marking encoded.

Another implementation alternative can for example be based on rate measurements and marking. In particular, the PCN-interior-nodes packets using the PCN_marking, whenever the measured PHB aggregated PCN traffic rate exceeds a pre-configured rate threshold denoted as configurable-admissible-rate. An example of the detailed operation of this later procedure is described below. The predefined configured-admissible-rate, see Section 3.1.1.1 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 configured-admissible-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.

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

* 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 PCN unmarked, PCN_marking (and PCN_Affected_Marking bytes, when the ECMP solution is used, see Section 3.2.4) 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 (and PCN_Affected_Marking, when the ECMP solution is used, see Section 3.2.4) bytes.

Then the PCN-interior-node calculates the current estimated excess rate (i.e., overloaded rate), say signaled_overload_rate, by using the following equation:

```
signaled_overload_rate =
   ((total_received_bytes) / T) - configured-admissible-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 > configured-admissible-rate
THEN
 {
   IF (incoming_PCN_marking_rate <> 0)
   THEN
    { signaled_remarked_bytes =
         ((signaled_overload_rate -
          incoming_PCN_marking_rate) * T) / N
    }
   ELSE signaled_remarked_bytes =
          signaled_overload_rate * T / N
  }
```

Where the "incoming_PCN_marking_rate" is calculated as follows:

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

where input_PCN_marking_bytes represents the measured number of bytes carried by PCN_marking encoded packets.

When incoming PCN_marking encoded packets (or 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 PCN_marking encoded packets (or bytes) are not dropped. This could for example be accomplished by using different dropping rate thresholds for PCN_marking encoded and PCN unmarked (and PCN_Affected_Marking encoded, when ECMP solution is used) bytes, see Section 3.1.1.1.

4.1.2. Operation in PCN-egress-nodes

The PCN-egress-node measures the rate of the received PHB aggregated PCN_unmarked and PCN_marking marked packets. The measurements on the PCN unmarked and unmarked traffic can be implemented using a similar functionality as the one specified in [Char07] and [CL-ARCH] to calculate the Congestion Level Estimate (CLE), which is the ratio of the N*PCN_marked_traffic received from one PCN-ingress-node, which is calculated as an EWMA and the total rate (PCN_marking_rate and PCN_unmarked_rate) received, which is also calculated as an EWMA.

The PCN_marking_rate can be then calculated as follows:

```
            PCN_marking_rate =
                     input_PCN_marking_bytes / T
```

where input_PCN_marking_bytes represents the measured number of
     bytes carried by the PCN_marking encoded packets

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

If the value of CLE is higher than a certain value, e.g., 1%, then the PCN-egress-node is changing its operational state from Normal state to Admission Control state, see Section 3.3.2.

```
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: 3 Admission control based on data marking and probing

The admission control procedure is accomplished by using a combination of the PCN operational state of the PCN-egress-node and an admission control request provided by an external to PCN, signaling protocol. When the admission control request arrives at a PCN-egress-node that operates in admission control state then the request SHOULD be rejected. If it operates in Normal state it SHOULD be accepted. When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-egress-node SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur, see Section 3.4.

4.1.3. Operation in PCN-ingress-nodes

The PCN-ingress-node can receive a reservation request message belonging to an external to PCN, signaling protocol, e.g., RSVP. This reservation request message can be used during the admission control process. If the PCN-ingress-node receives a response, from the PCN-egress-node, which notifies that the reservation request message belonging to the external signaling protocol was successfully

processed, then the reservation request SHOULD be admitted. Otherwise it SHOULD be rejected, see Section 3.3.3. Both situations SHOULD be notified to the sender of the flow.

When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-ingress-node SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur, see Section 3.4. Furthermore, when the DSCP encoding is used to encode the not congested PCN state, see Section 3.4, then the PCN- ingress-node SHOULD remark to not congested PCN encoding state, all incoming to PCN domain, packets associated to flows that need to use the LC-PCN features.

4.2. Admission control based on probing for unidirectional flows

This type of admission control uses probing, whereby a probe packet is sent along the forwarding path in a network to determine whether a unidirectional flow can be admitted based upon the current congestion state of the network. In pre-congestion situations the probe packets are PCN_marking encoded to notify the PCN-egress-node that a congestion occurred on a particular PCN-ingress-node to PCN-egress-node path. The Admission control based on probing feature is used to solve the ECMP issue that might occur during the process of admission control, see Section 3.1.3.

This admission control procedure can be used for both bandwidth management models, ingress-egress-aggregate model and the HOSE model. The main difference between the admission control features used in these models is that the ingress-egress-aggregate model maintains and uses aggregated states per each ingress pair and per each traffic class. The HOSE model maintains and uses per traffic class, i.e., PHB, states, but it does not use aggregates per each ingress and egress pair.

4.2.1. Operation in PCN-interior-nodes

The PCN-interior-node features that are used to detect the PCN operational states, are the same as the ones described in section 4.1.1. In this scenario an IP packet is used as a probe packet, see Figure 3. A probe packet that passes through a PCN-interior-node that operates in Admission Control state (or in Flow Termination state, when either the Flow Termination optimization mode or the ECMP solution described in Section 3.2.4 are used) MUST remark the PCN unmarked encoded probe packet to PCN_marking encoded probe packet.

The PCN-interior-nodes SHOULD detect a probe packet by observing the Router Alert option, which is carried by the probe packet. Note that

a PCN-ingress-node sets the Router Alert option of all packets that are used as probe packets. This also holds for signaling protocol messages that are used by LC-PCN as 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. If the PCN-interior-node operates in Admission Control state (or in Flow Termination state, when either the Flow Termination optimization mode or the ECMP solution described in Section 3.2.4 are supported) then PCN-interior-node SHOULD PCN_marking encode the probe packet. Otherwise, the encoding state of the probe packet SHOULD NOT change.

4.2.2. Operation in PCN-egress-nodes

The PCN-egress-node measures the rate of the received aggregated PCN unmarked and PCN_marking encoded packets. When the probe packet arrives at the PCN-egress-node that is belonging to a certain ingress-egress PCN aggregate, and it is PCN_marking encoded then the request SHOULD be rejected. In this way it is ensured that the probe packet passed through the node that it is congested and therefore, it can be used to solve the associated ECMP issue, see Section 3.4.

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 even when no edge-to-edge pair PCN aggregate state is available at the PCN-egress-node and when a probe packet arrives at the PCN-egress-node, then this request SHOULD be rejected if the probe packet is PCN_marking encoded. Otherwise, i.e., if the probe packet is not PCN_marking encoded, it SHOULD be accepted. When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-egress-node SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur, see Section 3.4.

4.2.3. Operation in PCN-ingress-nodes

Similar, to Section 4.1.3, the PCN-ingress-node can receive a reservation request message belonging to an external to PCN, signaling protocol, e.g., RSVP PATH message. Subsequently, the PCN-ingress-node sends a probe packet, see Figure 3, towards the PCN-egress-node. When RSVP is used, the RSVP PATH message is the probe packet. Note that the probe packet should use the same flow ID information and encoding state (ECN and/or DSCP) as the data packets associated with the received reservation request message. The PCN-ingress- node if needed is modifying specific IP header information In probe packets to give the possibility to the PCN-interior-nodes to make a distinction between probe packets and normal packets. Note that defining the IP header information that can be used for this

purpose is out of the scope of this document. An example of such information could be the Router Alert option. In this case the PCN-ingress-node sets the Router Alert option carried by the probe packet.

Note that probe packets can be either user data packets or messages used by an external, to PCN, on path signaling protocol, e.g., RSVP PATH. If the PCN-ingress-node receives a response that notifies that the probe was successfully processed, then the reservation request is admitted. In case of RSVP, the response is RSVP RESV message. Otherwise it is rejected, see Section 3.3.3. Both situations have to be notified to the sender of the flow.

When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-ingress-node SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur, see Section 3.4. Furthermore, when the DSCP encoding is used to encode the not congested PCN state, see Section 3.4, then the PCN- ingress-node SHOULD remark to not congested PCN encoding state, all incoming to PCN domain, packets associated to flows that need to use the LC-PCN features.

4.3. Flow Termination for unidirectional flows

The Flow Termination handling method requires the following functionalities. This flow termination handling procedure can be used for both bandwidth management models, ingress-egress-aggregate model and the HOSE model. The main differences between the flow termination features used in these models are the following. The ingress-egress-aggregate model maintains and uses aggregated states per each ingress pair and per each traffic class. The HOSE model maintains and uses per traffic class, i.e., PHB, states, but it does not use aggregates per each ingress and egress pair.

The ingress-egress-aggregate model MUST use the flow termination base mode and it MAY use the flow termination optimisation mode and the ECMP solution that applies for flow termination support. The HOSE model MUST use the flow termination base mode, the optimization mode and the ECMP solution that applies for flow termination support.

For both models the PCN-egress-node needs to store for each flow, per flow reservation information and the address, e.g., IP address and port number, of the PCN-ingress-node from where the particular flow passed before arriving to the PCN-egress-node. This can be for example done by using information that is carried by the external protocol used in combination with the LC-PCN solution.

4.3.1. Operation in the PCN-interior-node

The PCN-interior-nodes can measure the PHB aggregated PCN traffic that exceeds a configured-admissible-rate and mark this excess PCN traffic, see Figure 4. This can be accomplished using different metering and marking features, see Section 4.1.1.

The admission control features described in Section 4.1.1 can be applied also for the situation that the PCN-interior-node operates in the base mode of the Flow Termination state. The optimisation mode is used to support the ECMP solution and the HOSE model.

```
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: 4 LC-PCN Flow Termination handling

4.3.1.1. Optimization mode features for Flow termination

In order to solve the ECMP issue described in Section 4.3.1.2 an additional optimization mode feature can be used. The main addition that this optimization mode solution requires is that an additional operational state has to be maintained by the PCN-interior-node, i.e., a Flow Termination state, see Section 3.4.1. In particular, when the measured PHB aggregated PCN traffic is higher than the threshold equal to (U * configured-admissible-rate), then the PCN-interior-node changes from the Admission Control state to the Flow Termination state. When a token bucket implementation is used and when operating in optimisation mode, the token bucket must use an additional threshold, i.e., U*configured_admissible_rate. When above this threshold all packets that are not being PCN_marking encoded must be marked as PCN_Affected_Marking encoded.

Furthermore, when the PCN-interior-nodes calculates the overload rate that has to be signalled, in a similar way as described in Section 4.1.1. The optimisation mode must also be used when the PCN domain

supports the HOSE model.

4.3.1.2. ECMP solutions

As discussed in Section 3.2.4, the ECMP issue that may occur during Flow Termination operational state, could be solved by using an additional PCN marking encoding approach, denoted as: PCN_Affected_Marking.

In this case both the Flow Termination base and optimization modes have to be slightly modified, see Section 3.3.1.

Furthermore, in Flow Termination state, the PCN-interior-node marks all PCN unmarked (i.e., not congested PCN encoded) packets that are passing through the PCN-interior-node. The same ECMP features are used when the HOSE model is applied in the PCN domain.

4.3.2. Operation in PCN-egress-nodes

The PCN-egress-node measures the rate of the received PHB aggregated PCN unmarked and PCN_marking encoded packets, see Figure 4. The Flow Termination activation / triggering depends among others on whether the PCN domain supports PCN_Affected_Marking encoding, see Section 3.2.2 and Section 3.3.2.. The implementation of the Flow Termination algorithm can be 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 signaling protocol and that the PCN-edge-nodes can store and maintain the priority information related to each of the end-to-end flows. The terminated flows are selected from the flows belonging to the same edge-to-edge pair PCN aggregate and having the same PHB traffic class as the PHB of the PCN_marking encoded packets (and PCN_Affected_Marking encoded packets, when the ECMP solution is used).

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.

An example of the algorithm for the calculation of the number of

flows, belonging to the same edge-to-edge pair PCN aggregate and associated with each priority class that have to be terminated is described using the pseudocode below. First, when the PCN-egress-node operates in the Flow Termination state, see Section 3.4.2, then the total amount of PCN_marking_rate, per edge-to-edge pair PCN aggregate, associated with the PHB traffic class, say incoming_PCN_marking_rate, is calculated. This rate represents per edge-to-edge pair PCN aggregate, the flow termination bandwidth, that should be terminated. 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 measured number of bytes carried by PCN_marking encoded packets.

To provide reliable estimation of the encoded information several techniques can be used, see [AtLi01], [AdCa03], [ThCo04], [AnHa06]. The value of the incoming_PCN_marking_rate that has to be used to calulate the bandwidth that has to be terminated, needs to be adjusted, since the excess rate that was calulated during the admission control state must not be taken into acount. We denote this new value as adjusted_incoming_PCN_marking_rate and it is equal to:

```
adjusted_incoming_PCN_marking_rate =
        incoming_PCN_marking_rate  -
                          configured-termination-rate-egress
```

where configured-termination-rate-egress is defined in Section 3.2.2 and in the description of event B in Section 3.3.2.

In Flow termination, inaccuracies in excess rate measurements might occur due to the delay between the metering and marking event that occurs at the PCN-interior-nodes, the decisions that are made at PCN-egress-nodes, and the termination of flows that are performed by PCN-ingress-nodes, see section 6 in [CsTa05]. Furthermore, 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.

In order to solve these inaccuracies when operating in Flow Termination state, the PCN- egress-nodes use a sliding window memory to keep track of the measured adjusted_incoming_PCN_marking_rate_ in a couple of previous measurement intervals. At the end of a measurement intervals, T, before using the measured adjusted_incoming_PCN_marking_rate to calculate the bandwidth that needs to be terminated, the actual measured adjusted_incoming_PCN_marking_rate is decreased with the sum of already adjusted_incoming_PCN_marking_rate stored in the sliding window memory, since that bandwidth to be terminated 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]. However, based on several experiments that have been performed for the situation that the sliding window is applied at the PCN-egress-node instead the PCN-interior-node, it is recommended that the best value that can be used for the sliding window size at the egress is equal to 1.

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

If Mi is the adjusted_incoming_PCN_marking_rate stored in ith memory cell (i = [1..n]), then at the end of every measurement interval, the adjusted_incoming_PCN_marking_rate that is used to calculate the bandwidth that has to be terminated is calculated as follows:

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

termination_PCN_marking_rate  =
       adjusted_incoming_PCN_marking_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 < - termination_PCN_marking_rate
```

The term denoted as terminated_bandwidth in the below pseudocode 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). Note that this is somewhat over-conservative for situations that the number of flows that are included into the ingress-egress-aggregate is small.

```
terminated_bandwidth = 0;
priority_class = 0;
while terminated_bandwidth < termination_PCN_marking_rate
{
  terminate_bandwidth(priority_class) =
     termination_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 SHOULD generate and send 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 SHOULD send within a notify message the to be released bandwidth, associated with the edge-to-edge pair PCN aggregated state. When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-egress-node SHOULD restore the original DSCP values of the PCN remarked packets,

otherwise multiple actions for the same event might occur, see Section 3.4.

Note that in the ingress-egress-aggregate model the excess rate and the flows to be terminated are associated with the same edge-to-edge (i.e., ingress-egress) pair PCN aggregate and with the same traffic class, i.e., PHB.

In the HOSE model the excess rate and the flows to be terminated are associated with the same traffic class, i.e., PHB, see Section, see Section 3.2.2. Furthermore, the HOSE model always uses the PCN_Affected_marking encoding. For the flows that should be terminated the PCN-egress-node informs the associated PCN-ingress-node to terminate them. If the PCN domain does not support the ECMP solution, and if it uses the ingress-egress-aggregate model and if the PCN-egress-node receives any admission flow request, belonging to a ingress-egress-aggregate state operating in flow termination state then the request must be rejected. If the PCN domain is supporting the ECMP solution and/or is supporting the HOSE model then the PCN-egress-node rejects new flow admission requests if the flow admission request packet is either PCN_marked or PCN_Affected_Marking encoded. Otherwise it is admitted.

4.3.2.1. ECMP solutions

When the ECMP solution is used by the PCN-egress-node then the following modifications are required. The rate of the PCN unmarked (or bytes), used on the calculations of the event that triggers the Flow Termination state, see Section 3.3.2 has to be replaced, by the rate of PCN_Affected_Marking encoded packets (or bytes). Note that this is already explained in Section 3.3.2. Furthermore, the PCN-egress-node uses the PCN_Affected_Marking to identify which flows were affected by the exceptional/severe congestion. In this way the PCN-egress-node, when operating in Flow Termination state, see Section 4.3.2, is able to terminate only the flows that received one or more PCN_Affected_Marking packets. The same ECMP features are used when the HOSE model is applied in the PCN domain. The main difference is related to the fact that the solution takes into account the fact that the excess rate and the flows to be terminated are associated with the same traffic class, i.e., PHB, but they are not required to belong to the same ingress-egress-aggregate, see Section 3.2.2.

4.3.3. Operation in 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 4, 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 e.g., sending external to PCN, release signaling messages.

Furthermore, depending on the used policy, the packets related to the flows that have to be terminated are either blocked or shifted to an alternative LC-PCN traffic class, i.e., PHB. Moreover, depending on the used policy, the PCN-ingress-node could reject all new flow admission requests that are associated with the same edge-to-edge pair PCN aggregate until no other requests to terminate flows are received from PCN-egress-nodes.

The same features are used when the HOSE model is applied in the PCN domain. The only difference is that a policy that rejects all new flow admission requests cannot be used.

In the case that the PCN domain supports the ingress-egress-aggregate model and 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. The functionality required to resize the edge-to-edge pair PCN aggregated state is out of the scope of PCN.

When DSCP is used for PCN encoding and no trust relationships exist between the PCN-domains, then for packets that are forwarded outside the PCN-domain, the PCN-ingress-node SHOULD restore the original DSCP values of the PCN remarked packets, otherwise multiple actions for the same event might occur, see Section 3.4. Furthermore, when the DSCP encoding is used to encode the not congested PCN state, see Section 3.4, then the PCN- ingress-node SHOULD remark to not congested PCN encoding state, all incoming to PCN domain, packets associated to flows that need to use the LC-PCN features.

4.4. Admission control based on data marking and probing for bi-directional flows

This section describes the admission control scheme that uses the admission control function based on datamarking and 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 data marking and 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 and 4.2 for the situations that the admission control with data marking and admission control with probing are used, respectively. The main difference is related to the location of the PCN-interior-node 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 and 4.2, Figure 3. 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 probe packet and any packet of the "reverse" direction flows will be PCN_marking encoded. The PCN-ingress-node is in this way notified that a pre-congestion situation occurred in the network and therefore it will 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 data marking and probing for
            bi-directional admission control (pre-congestion on path
            PCN-egress-node towards PCN-ingress-node)

4.5. 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 4.3.

```
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 (reverse)    |              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 4.3. The main difference is related to the location of the the PCN-interior-node that operates in Flow Termination state, , i.e. "forward" or "reverse" path. 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 4.3. 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.

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 (see Figure 7), 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 (see Figure 9). 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 Section 3.3.3), 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)

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

   o 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.

   o 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 4.3.2, 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 bandwidth 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 "reverse" directions.

5. Performance Evaluation

The goal of this section is to describe and compare the LC-PCN performance with the SM solution described in [Char07]. Due to time constraints only a subset of the LC-PCN performance experiments will be discussed in this version of the draft. In particular, only flow termination experiments for unidirectional flows will be discussed. And from these flow termination experiments, only a small subset of the results will be shown. Regarding the admission control experiments, only the admission control based on data marking experiments are relevant for this comparison. However, if implemented well, then the admission control based on data marking LC-PCN and SM solutions should be identical. Therefore, it is expected that they should have the same performance and are therefore not discussed in this version of the draft.

5.1. Flow Termination Experiments

Three sets of flow termination experiments have been performed. The first set of experiments is used to test the sensitivity to low ingress-egress aggregation levels, see also Section 7.3.1 in [Char07]. The second set of experiments is used to test over-termination in the multi-bottleneck scenarios, see Section 7.3.2 in [Char07]. The third set of experiments is used to test the impact of the reaction time for situations that the overload is higher than 100% the maximum capacity of the link and when the value of the proportionality parameter N is varied. In order to compare the LC-PCN results with the SM results a subset of the SM results presented in Sections 7.3.1 and 7.3.2 in [Char07] are copied and used for comparison reasons in this draft.

5.2. Simulation Setup and Environment

5.2.1. Network Topology and Signaling Models

Both LC-PCN ingress-egress-aggregate (trunk/pipe) and the HOSE bandwidth management models have been used during these experiments. Both bandwidth management models support the flow termination optimisation mode and the use of the PCN_Affected_Marking encoding.

Both bandwidth management models consider that the flows that are Flow Termination notified by the PCN-egress-node have to be terminated by the PCN_ingress-node. Furthermore, the packets related

to the flows that have to be terminated are blocked. Moreover, the PCN-ingress-node does not reject new flow admission requests. When operating in admission control or flow termination state, the PCN-egress-node rejects new flow admission requests.

Furthermore, it is considered that during flow termination a PCN-ingress-node does not block any new incoming flow admission requests. These new flow admission requests are then rejected by the PCN-egress-node.

The used network topologies are identical to the ones described in Section 8.1.1 of [Char07]. In particular, Figure 10 and Figure 11 are identical to Figures A.2 and A.3, respectively, from [Char07].

```
     A
      \

   B  -- D -- F

      /
    C
```

Figure 10: Simulated Multi Link Network, same as in [Char07]

```
A--B--C     A--B--C--D      A--B--C--D--E--F
|  |  |     |  |  |  |      |  |  |  |  |  |
|  |  |     |  |  |  |      |  |  |  |  |  |
D  E  F     E  F  G  H      G  H  I  J  K  L

  (a)          (b)                (c)
```

Figure 11: Simulated Multiple-bottleneck (Parking Lot) Topologies, same as in [Char07]

The description of these network topologies is given in section 8.1.1 form [Char07]. In particular, Figure 10 describes a multi-link network, denoted also as RTT, which uses interconnects a subset of ingresses (A, B, C) to an interior node, i.e., D. The interior node D is connected to the egress node F, via a link that is considered in the simulations to be the bottleneck. Therefore, for this link the LC-PCN algorithm is enabled. The capacities of the ingress to interior links are not limiting, i.e., are not bottlenecks, and

therefore the LC-PCN algorithm is not enabled on those. It is important to note that all links are T3 links that are supporting a capacity of 45 Mbs.

This topology is used to study the Sensitivity to Low Ingress-Egress aggregation levels experiments. The number of ingresses varied, in the range 2 - 35. All links' RTT are set to 1ms to eliminate the potential RTT influence.

Figure 11 describes a multi-bottleneck network topology (or Parking Lot, PLT). In the studied experiments only the multi-bottleneck topology with 5 bottlenecks depicted in Figure 11.c is used. In particular this topology is used to study the over-termination in multi-bottleneck scenarios. In Figure 11.c there is one ingress-egress pair, ingress A to egress F, that carries the aggregate of long flows traversing all 5 bottlenecks. The other 5 ingress- egress pairs (G - H, H -I, I - J, J- K, K- L) that carry flows that are traversing a single bottleneck link and exiting at the next hop. For example, the ingress-egress pair G- H carries flows that pass the bottleneck A - B and are exiting at the egress H. In all cases it is considered that the vertical links are not limiting and that only the horizontal links are bottlenecks. The capacity of all links are considered to be T3 links, i.e., 45 Mbs. The propagation delays for all links in the topology are set to 1ms. In is considered that the propagation delays from source to ingress and from destination to the egress are negligible and are not modeled.

The flows are generated using an exponential distribution and their holding times it is assumed to have an exponential distribution with an average of 1 minute.

5.2.2. Traffic models

This section is based on Section 8.1.2 from [Char07]. The studied experiments are using the CBR voice codec and is described in Table 1. The CBR traffic is modeled as a source that generates packets that have a constant size of 160 bytes and are generated at a constant inter-arrival time. Next versions of this draft will also include other traffic models presented in [Char07].

Name/ Codecs	Packet Size (Bytes)	Inter-Arrival Time (ms)	On/Off Period Ratio	Average Rate (kbps)
"CBR"	160	20	1	64

Table 1 Simulated Voice Codec.

5.3. Parameter Settings

The packet size is 160 bytes, see Table 1. Furthermore, the weight used in the EWMA used to calculate the CLE is set to the value of 0.5. The CLE threshold is chosen to be 0.001. The capacities of all links are considered to be T3 links, i.e., 45 Mbs links. The length of the used queues in all nodes and for all experiments is set to 4994 packets. The flows use only one priority. The number of windows used in the sliding window algorithm is equal to 1. The congestion-admissible-rate is set to 0.5 of the link speed. The value of U is set equal to 1.2.

The simulation model used by the PCN interior nodes use the rate based measurement and marking algorithm. However, it is considered that the rate based measurement and marking algorithm can be fully specified using the token bucket specification described in [Char07], with the following modifications.

o During admission control (and flow termination) the token bucket must mark every N packets instead of marking each packet.

o The PCN_marking encoded packets must not be preferentially dropped. In particular, in situations of overload, the PCN_marking, PCN_Affected_Marking and PCN_unmarked encoded packets are dropped randomly.

o When operating in optimisation mode, the token bucket must use an additional threshold, i.e., U*configured_admissible_rate. When above this threshold all packets that are not being PCN_marking encoded must be marked as PCN_Affected_Marking encoded.

5.4. Performance Metrics

The used performance metrics are the over-termination and the reaction-time.

The over-termination performance metric is defined in [Char07], as the percent deviation of the measured mean rate of the load from the expected load level. The load mean rate is measured in the following way. The actual achieved throughput at 100 ms intervals is measured. Then the average of these 100 ms rate samples is computed over the duration of the experiment (where relevant, excluding warmup/startup conditions). The measured/actual average rate of the load is then compared to the desired/optimal traffic load. In pseudo code the over termination percentage can be described as follows:

Over termination =

$$\frac{(\text{actual termination} - \text{optimal termination})}{\text{optimal termination}} * 100$$

The reaction time is defined as the duration of time that a bottleneck node remains in flow termination state. The lower the duration that a bottleneck node remains in flow termination, the faster is the reaction time.

5.5. Ingress-Egress Aggregation Experiments

This section describes the results of the first set of experiments. The goal of this set of experiments is to study the over-termination and reaction time to the level of aggregation. The performance metrics used in these experiments is the over-termination and the reaction time. In this set of experiments the value of N = 1 was used. The over-termination results are shown in Table 2 while the reaction time results are shown in Table 3..

	No. Ingre	Flow per Ingre	Over-Termination % SM	LC-PCN trunk	LC-PCN HOSE
CBR	2	289	4.112	11.32	5.556
	10	57	6.710	9.232	9.097
	35	16	14.04	6.201	7.439
	70	8	16.39	6.136	7.149

Table 2: Over - termination comparison between SM, LC-PCN ingress-egress-aggregate and LC-PCN HOSE

Comparing the results presented in Table 2 the following observations and conclusions can be drawn. The overtermination is under 11% for all experiments. The LC-PCN trunk and HOSE models, including the model that applies the ECMP solution, are not sensitive to aggregation in terms of flows per PCN-ingress-node.

	No. Ingre	Flow per Ingre	Reaction time (ms) SM	LC-PCN trunk	LC-PCN HOSE
	2	289		200	200
CBR	10	57		200	200
	35	16		300	200
	70	8		600	300

Table 3: Reaction time comparison between SM, LC-PCN ingress-egress-aggregate and LC-PCN HOSE

Table 3 provides the reaction time results obtained for LC-PCN trunk and LC-PCN HOSE based experiments for different flow aggregation situations. The reaction times associated with SM are not known, and therefore, they are not shown in Table 3. the reaction times vary between 200 and 300 ms. It is important to observe that for the situation that the LC-PCN trunk model is used and the number of flows per ingress is very low, i.e., 8 flows per ingress then the reaction time is higher than average, i.e., 600 ms. This result show that the part of the flow termination algorithm described in Section 4.3.2, which calculates how many flows to be terminated, i.e., calculate_terminate_flows, is over-conservative. This means that when the bandwidth to be terminated is smaller than the rate of the flows that is requesting the lowest bandwidth then no flow will be selected for termination. In particular, see Section 4.3.2, 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). Due to this fact the reaction time is increased. This issue can be solved by allowing the calculate_terminate_flows procedure to select a flow for termination even if the bandwidth to be terminated is lower than the smallest bandwidth allocated to a flow. In pseudo-code this can be specified as adding the following step to the procedure denoted as calculate_terminate_follows:

```
IF (sum_bandwidth_terminate(priority_class) <
                smallest allocated bandwidth to a flow)
THEN
   select flow with smallest allocated bandwdith for
     termination
```

5.6. Multi Bottleneck Experiments

The goal of these experiments as also emphasized in Section 8.3.2 from [Char07] is to study the beat down effect of flows traversing multiple bottleneck links. In this set of experiments the value of N = 1 was used. The over-termination results of these experiments, see Table 4, are compared with some of the SM results that are presented in Section 8.3.2, table A.11, from [Char07]. Note that the bottleneck rows are ordered based on the flow traversal order (from upstream to downstream). In particular the CBR SM results obtained for the 5-PLT topology are used.

As explained in Section 8.3.2 from [Char07] at failure event time, all bottleneck links have a load of roughly 3/4 of its link size. In addition, the long IEA constitutes 2/3 of this load, while the short one is 1/3. The performance metrics used in these experiments are the over-termination and the reaction time.

The calculation of the LC-PCN over-termination percentages is done in the same way as described in [Char07]. "We take each link in the topology separately and compute the "rate-proportionally fair" rates that each IEA sharing this bottleneck will need to be reduced to (in proportion to their demands), so that the load on that bottleneck independently becomes equal to the termination threshold (this threshold being implicit for SM, explicit for CL), assuming the initial sum of rates exceeds this threshold. After this is done independently for each bottleneck, we assign each IEA the smallest of its scaled down rates across all bottlenecks. We then compute the "reference" utilization on each link by summing up the scaled down rates of each IEA sharing this link. Our over-termination is then reported in reference to this "reference" utilization. We note that this reference utilization may frequently be already below the termination threshold of a given link. This can happen easily in the case when a large number of flows sharing a given link is "bottlenecked" elsewhere.", from [Char07].

	Topo. 5 PLT	Over-termination % LC-PCN trunk	LC-PCN HOSE	SM
	BN1	30.03	23.58	35.04
CBR	BN2	27.38	25.29	23.54
5	BN3	21.69	19.22	23.36
	BN4	21.50	23.59	23.78
	BN5	24.53	26.24	24.08

Table 4: Over-termination comparison of SM, LC-PCN for 5 PLT topology

Comparing the results presented in Table 4 the following observations and conclusions can be drawn. The over-termination for LC-PCN trunk model varies between 21 and 30 %. The over-termination for LC-PCN HOSE model, including the model that supports the ECMP solution, varies between 19 and 26%.

	Topo. 5 PLT	Reaction time (ms) LC-PCN trunk	LC-PCN HOSE	SM
	BN1	200	200	
CBR	BN2	200	200	
5	BN3	200	200	
	BN4	200	200	
	BN5	200	200	

Table 5: Reaction time comparison between SM, LC-PCN ingress-egress-aggregate and LC-PCN HOSE

Table 5 provides the reaction time results obtained for LC-PCN trunk and LC-PCN HOSE based experiments for different flow aggregation situations. The reaction times associated with SM are not known, and therefore, they are not shown in Table 5. The reaction times are equal to 200 msec for all bottleneck links for both LC-PCN trunk and LC-PCN HOSE models.

5.7. Reaction times versus N Experiments

The goal of this set of experiments is to observe the impact of the value of N on the reaction times when the level of overload is much higher than 100% of the capacity of the link. In this set of experiments, the simulated multi link network topology depicted in

Figure 10 is used, where the number of ingresses is 10. Furthermore, in this set of experiments only the LC-PCN trunk model is used.

	No. Ingre	Flow per Ingre	Overload%	N	LC-PCN trunk reaction time (msec)
	10	57	180%	1	1400
CBR	10	57	180%	2	800
	10	57	250%	1	1400
	10	57	250%	3	700

Table 6: Reaction time versus N

The reaction time results depicted in Table 6, show that in situations of high overload the flow termination performance, from the point of view of reaction times, can be increased up to a factor of 2 when a higher value than 1 is chosen for parameter N.

5.8. Experiment Conclusions

Based on the results obtained from the experiments presented in Sections 5.5, 5.6 and 5.7 this document recommends the following:

- o Leave open the option to use PCN_Affected_Marking encoding since it can solve the ECMP problem and it can provide an efficient solution for the HOSE model. In this document the term HOSE is referring to the aggregation of incoming traffic from all ingress edges, which is associated with one traffic class, i.e., PHB, towards one egress edge. This type of HOSE model is equivalent to the Multiple to Point (MP2P) type of aggregation.

- o Leave open the option of using random dropping in PCN-interior-nodes for PCN_Marking, PCN_Affected_Marking and PCN_unmarked encoded packets.

- o Leave open the option of using the parameter N such that the marked excess rate can represent also high level of measured excess rate:

**** Implemented by marking every N-th packet (or byte) instead of marking each packet (or byte).

6. Security Considerations

The security considerations associated with this document are similar to the one described in [Eard08].

7. IANA Considerations

To be Added

8. Acknowledgements

The authors express their acknowledgement to the following colleagues: Andras Csaszar, Attila Takacs, David Partain, Zoltan Turanyi, Geert Heijenk, Anna Charny, Philip Eardley, Kwok Chan, Bob Briscoe, Joe Babiarz, Michael Menth.

9. 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-01 (work in progress), , November 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-03 (work in progress), , November 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.

[DuGo99] Duffield, N. and P. Goyal, "A Flexible Model for Resource Management in Virtual Private", Proc. of ACM/SIGCOMM pp. 95 - 108, December 1999.

[Eard08] Eardley, P., "Pre-Congestion Notification Architecture", draft-ietf-pcn-architecture-08 (work in progress), , October 2008.

[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-13.txt (work in progress), , July 2008.

[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
302 Barthel Drive
Cary, NC 27513
USA

Phone: +1 919 522 0964
Email: bhargava.anurag@gmail.com

Attila Bader
Ericsson
Laborc 1
Budapest
Hungary

Email: Attila.Bader@ericsson.com

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

Email: g.karagiannis@ewi.utwente.nl

Hein Mekkes
Researcher
Javastraat 125
7521 ZE Enschede
Netherlands

Email: heinmekkes@gmail.com