Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 14, 2008

A. Charny
Cisco Systems, Inc.
J. Babiarz
Nortel
M. Menth
University of Wuerzburg
J. Zhang
Cisco Systems, Inc & Cornell University

November 11, 2007

Comparison of Proposed PCN Approaches
draft-charny-pcn-comparison-00.txt

Status of this Memo

Copyright Notice

Abstract

Several, sometimes conflicting proposals have been offered for the consideration of the PCN WG regarding PCN internal node and PCN edge

node behaviors. Based on the WG charter, the WG needs to make a decision on which of the proposed PCN-interior-node and PCN-boundary-nodes behaviors to endorse. The primary goal of this draft is twofold. First, we attempt to summarize the functional differences between the proposed alternatives. Second, we provide a brief summary of performance evaluation results. Finally we propose a view on how a (parameterized) specification of the PCN-interior-node metering and marking function can be described to enable several of the proposed behaviors. We argue that if this parameterized specification is used for specifying the PCN-interior-node behavior, then it can support a range of behaviors at the PCN-boundary-node. The decision on which of the PCN-boundary-node behaviors to choose can then be considered separately. We also discuss complexities associated with choosing such uniform approach.

Requirements Language

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

Table of Contents

1. Introduction

1.1. Terminology

This draft uses the terminology defined in [I-D.ietf-pcn-architecture]

1.2. Introduction

A number of seemingly diverse approaches have been presented in the context of the work in the PCN WG, encompassing both the PCN-interior-node and PCN-boundary-node node behaviors. The goal of this informational draft is to provide a functional comparison of several candidate approaches for PCN behaviors. We refer the reader to [I-D.ietf-pcn-architecture] for an architectural overview of PCN.

The first draft of this document concentrates on the CL approach proposed in [I-D.briscoe-tsvwg-cl-architecture] , the 3SM approach described in [I-D.babiarz-pcn-3sm] , and the Single-Marking (SM) approach proposed in [I-D.charny-pcn-single-marking]. The approach proposed in [I-D.westberg-pcn-load-control] is not covered in the initial version of this draft, pending clarifications on the open questions regarding the details of the algorithms described in that draft.

At the time of writing of this draft, several performance studies have been undertaken and are on-going. The studies performed in [I-D.zhang-pcn-performance-evaluation] and [I-D.charny-pcn-single-marking] provide a side-by-side comparison of performance of the CL and SM proposals. A performance evaluation of the 3SM approach was reported in [I-D.babiarz-pcn-explicit-marking] and [TR437]. However, due to a number of differences in the experimental setups in different simulation studies, a side-by-side performance comparison between 3SM and the other two approaches is not fully possible at this point. Therefore, we present only a short overview of some of the performance evaluation results where possible.

We then argue that a unified (parameterized) formulation of the metering and marking behavior at the PCN-interior-nodes can be defined. Thus defined unified PCN-interior-node behavior may support multiple PCN-boundary-node behaviors, and hence in principle can be used in a variety of environments. We also discuss some of the tradeoffs and additional complexities associated with such unified PCN-interior-node behavior definition.

Sections 2-8 provide an overview of the three proposals, followed by a summary of functional differences between the proposals in Section

9. Section 10 briefly covers other comparison criteria not covered in this draft, including a brief summary of performance evaluation efforts as of the time of writing of this document. Section 11 provides a unified description of admission and termination functions of the PCN-interior-node covering all three proposals of CL, SM and 3SM, followed by a brief discussion of the tradeoffs associated with such unified definition in Section 12.

2. PCN-Interior-Node Metering and Marking Functions

This section provides a high level functional comparison of the metering and marking functions at the PCN-interior-node. For more detail, please see [I-D.briscoe-tsvwg-cl-phb], [I-D.charny-pcn-single-marking], and [I-D.babiarz-pcn-3sm].

2.1. Metering and Marking Types

Metering functions are defined in different proposals via the notions of Token Bucket (TB) or Virtual Queue (VQ). These two formulations are equivalent in the sense that each one can be implemented via the other with appropriate settings. They may count packet or bytes. The marking functions differ with respect to how the queue length of the Q or the fill state of the TB is used which has a direct influence on the marking result. The following are the marking behaviours used in [I-D.briscoe-tsvwg-cl-phb], [I-D.charny-pcn-single-marking] and [I-D.babiarz-pcn-3sm].

- o Excess-rate-marking marks packets which exceed the configured rate. In the TB formulation, the packets are marked when the arriving packet does not find enough tokens in the token bucket (and the marked packet does not consume tokens from the TB). In the VQ formulation, the packets are marked when the arriving packet would exceed the configured maximum size of the VQ (and the marked packet is not added to the Q). These two formulations are equivalent with the same rates of the TB and VQ, and the depth of the TB numerically equal to the maximum size of the Q. This is the marking used for termination in CL, and for both admission and termination in SM.

- o Excess-rate-marking-with-marking-frequency-reduction is similar to the Excess-rate-marking in the sense that it also marks packets which do not find tokens in the TB (in the TB formulation), or would exceed the maximum size of the VQ (in the VQ formulation). However, to reduce the number of marked packets, whenever a packet is marked a certain amount of tokens are added to the TB (in the TB formulation) or the same number of bytes is removed from the queue of the VQ (in the VQ formulation). This is the marking used

for termination in 3SM.

- Ramp-marking is defined as follows. In the VQ formulation, two VQ thresholds are defined (below the maximum VQ size). Packets are marked with a certain probability depending on the VQ size at the time of the packet arrival. This probability is 0 for VQ queue length from 0 to a lower VQ threshold, it rises linearly from 0 to 1 between the lower and a upper VQ threshold, and it is 1 above the upper VQ threshold. As a result, no packets are marked when the queue length is below the lower VQ threshold, a few packets are marked when the queue length is between the lower and the upper VQ thresholds, and all packets are marked when the queue length is above the upper VQ threshold. In the equivalent TB formulation, two additional TB fill thresholds (called the lower and upper TB thresholds, both not exceeding the TB depth) are defined . The packers are marked with the probability 1 for a TB fill state from 0 to a lower TB threshold. The probability decreases linearly from 1 to 0 between the lower and upper TB thresholds, and it is 0 above the upper TB threshold. As a result, all packets are marked when the fill state is below the lower threshold, a few packets are marked when the fill state is between the lower and the upper threshold, and no packets are marked above the upper threshold. This is the marking described in [I-D.briscoe-tsvwg-cl-phb] (in the VQ formulation) where it is used for admission.

- Threshold-marking marks packets that make the queue length of the VQ exceed a certain threshold which is lower than the queue size. As a result, all packets are marked when the metered traffic exceeds the VQ rate. In the equivalent TB formulation, Threshold-marking marks packets that make the number of tokens in the TB fall below a certain threshold which is larger than zero. As a result, all packets are marked when the metered traffic exceeds the TB rate. The VQ and TB formulations are equivalent with the VQ of rate R, maximum size S and VQ threshold T, and TB with rate R, depth B and threshold T, and TB threshold B-T. Threshold-marking s a special case of Ramp marking when the lower and the upper (TB or Q) thresholds are identical. It is used for admission in CL and 3SM.

2.2. Metering and Re-Marking of Previously Marked Packets

When packets travel over several links within a PCN domain, they are possibly marked. This section clarifies the question concerning packets of which markings should be taken into account for the metering process on subsequent links and if so which markings are re-marked.

2.2.1. Admission Marking

3SM and CL consider packets of all markings for metering, but TM-marked packets must not be re-marked to AM. SM requires that previously marked packets are excluded from metering, as not doing so would result in underestimation of sustainable admission rate in the multiple bottleneck scenarios, and consequently will result in the under-estimation of the sustainable termination rate at the PCN-ingress-node, in turn causing over-termination in the multiple bottlenecks scenarios.

2.2.2. Termination Marking

CL needs to exclude all previously termination-marked packets from metering in order to prevent underestimation of sustainable termination rate. If previously termination-marked packets are not excluded from metering, substantial over-termination in the multiple bottleneck scenarios might occur. 3SM can accommodate previously termination- marked packets being included for termination metering (although the exact impact of doing so needs to be further evaluated). Both in CL and 3SM, AM-marked packets may be remarked to TM. Note that SM does not employ termination marking at the PCN-ingress-node, an hence does not have any termination-marked packets at all.

2.3. Number of Metering Functions and Marking Codepoints

Both CL and 3SM require 2 different metering functions - one (pcn-lower-threshold) for admission and one (pcn-upper-threshold) for termination. Three marking codepoints are needed for both: unmarked, admission-marked (first PCN encoding) for admission, and termination-marked (second PCN encoding) to terminate flows.

In contrast, SM requires a single metering threshold and two different marking codepoints: marked and unmarked. (Note: There is a choice to be made whether the "marked" codepoint should use the first or the second PCN encoding state, if both are defined.)

3. Dropping Policies

A subtle difference in existing proposals for the PCN-interior-node behavior is related to how already marked packets need to be treated in the presence of loss. It turns out that all three approaches assume different drop preferences.

CL needs preferential dropping of termination-marked packets. If such preferential dropping is not implemented, then possible over-

termination may occur. It can be argued that the admission function of CL is not sensitive to whether or not admission-marked packets are preferentially dropped or not.

SM relies on preferential drop of marked packets. While admission function of this approach appears to be insensitive to the drop preference just as CL admission function, the termination function of SM will result in over-termination if preferential dropping of already marked packets is not implemented. While, insensitivity of CL admission function to marked packet drop remains to be studied, especially in the presence of large differences in packet sizes.

In contrast, the proposal in 3SM can benefit from preferential dropping of unmarked flow-termination packets, but it can function without at the expense of longer termination time. It can be argued that the admission function of 3SM is not sensitive to whether or not admission-marked packets are preferentially dropped or not.

4. PCN-Boundary-Node Behaviors

4.1. CL Boundary Behavior

In the CL approach, the PCN-egress-node measures the rate of admission-marked, termination-marked, and unmarked PCN-traffic per ingress-egress-aggregate. In addition, the PCN-ingress-node measures the rate of sent PCN-traffic per ingress-egress-aggregate. To support admission control, the PCN-egress-node calculates the Congestion Level Estimate (CLE) defined as a fraction of (admission- or termination-) marked traffic and the overall traffic, and signals the CLE to the PCN-ingress-node. The PCN-ingress-node accepts or rejects flows based on whether this CLE value for a particular ingress-egress aggregate exceeds a pre-defined threshold.

To support flow termination, the PCN-egress-node calculates (on a per-ingress-egress basis) the Sustainable (Termination) Rate, defined as the combined rate of unmarked and admission-marked packets, and signals the Sustainable Rate to the PCN-ingress-node. The PCN-ingress-node measures the ingress sending rate (again on a per-ingress-egress basis) and calculates the difference to the sustainable termination rate. It chooses an appropriate set of flows whose combined rate corresponds to that difference and terminates these flows. (Note: this choice is done without any knowledge of which flows had termination-marked packets.)

4.2. SM Boundary Behavior

In the SM approach, the PCN-egress-node measures the rate of marked and unmarked traffic per ingress-egress-aggregate. In addition, the PCN-ingress-node measures the rate of sent PCN-traffic per ingress-egress-aggregate.

To support admission control, the PCN-egress-node calculates the congestion level estimate (CLE) as the fraction of marked traffic and the overall traffic, and signals the CLE to the PCN-ingress-node. The PCN-ingress-node accepts or rejects flows based on whether this CLE value exceeds a pre-defined threshold.

Although the boundary functions necessary to support admission control are similar for CL and SM, an important difference between the two algorithms stems from the fact that CL is using threshold (or ramp) marking, while SM uses excess-rate-marking. As a result, the meaningful value of CLE is also different. For example, in case of small overloads, SM will have only a small fraction of packets marked (and hence the appropriate CLE value needed to detect the overload without over admission is small), while a small (but consistent) overload with CL results in the majority of packets being marked, resulting in CL being quite robust for a range of CLE values.

To support flow termination, the PCN-egress-node signals the rate of unmarked packets as the so-called Sustainable (Admission) Rate to the PCN-ingress-node. The PCN-ingress-node multiplies it by a system-wide constant to get the Sustainable Termination Rate. The rest of the termination is done like in the CL approach. The PCN-ingress-node measures the ingress rate and calculates the difference to the sustainable termination rate. It chooses an appropriate set of flows whose overall rate corresponds to that difference and terminates theses flows. (NOTE: This choice is done without any knowledge of which flows had marked packets.)

4.3. 3SM Boundary Behavior

To support flow admission, a PCN-egress-node analyzes the packet markings per ingress-egress-aggregate. Depending on the markings it sends from time to time "admission-stop" or "admission-continue" messages to the corresponding PCN-ingress-nodes to control their admission of new flows. The PCN-ingress-node admits new flows when the last control message was "admission-continue" and it rejects them when it was "admission-stop". 3SM leaves deliberately open the way how the PCN-egress-node decides when to send a specific control message. A very simple option for the PCN-egress-node is to send an "admission-stop" message when a single packet with admission- or termination-marking is observed and to send an "admission-continue"

message some time after the last marked packet has been observed. This is the method used for performance evaluation of 3SM reported in [I-D.babiarz-pcn-explicit-marking]. A more sophisticated option is to calculate a CLE based on an exponentially weighted moving average (EWMA) counting marked messages as 1 and umarked messages as 0. When the CLE exceeds an upper threshold, an "admission-stop" message is sent, and when the CLE falls below a lower threshold, an "admission-continue" message is sent. Other implementations are possible since the decision logic is local to the PCN-egress-node.

To support flow termination, the PCN-egress-node in the 3SM approach again monitors the packet markings and signals the flow ID of termination-marked packets to the PCN-ingress-node whereby several flow IDs may be sent in a single message. The PCN-ingress-node terminates these flows. Note that neither the PCN-ingress-node nor the PCN-egress-node are required to perform rate measurements in 3SM.

4.4. Notes on Using Different Boundary Behaviors with Different Marking/Metering Behaviors

The previous three Subsections describe the PCN-boundary-node behaviors in conjunction with specific proposals where these behaviors were defined. It should be noted, however, that various features of specific boundary behaviors described in the previous sections may be used with different marking/metering strategies. For example, as indicated in Section 4.3, the boundary behavior that CL uses for admission can also be used with 3SM. Likewise, the Termination function of CL may choose to only terminate those flows whose packets are TM-marked - see Section 6 for discussion on how this additional functionality can be used to address ECMP issues.

In general, new boundary behaviors may be designed to work with proposed metering and marking mechanism. Nevertheless, in the remaining portion of this document we will assume specific boundary mechanisms as described in sections 4.1 - 4.3 unless stated otherwise.

5. Informationed Signaled Between the Boundary Nodes

In CL, the PCN-egress-node must send to the PCN-ingress-node two values: a CLE for Admission and Sustainable (Termination) rate for Termination

In SM, the PCN-egress-node sends to the PCN-ingress-node the two values as in CL: the CLE and Sustainable (Admission) Rate. (Note that while the format of the signaling information is the same for CL and SM, the meaning of Sustainable Rate is different in the two

cases).

In 3SM, the PCN-egress-node signals to the PCN-ingress-node using control messages for the Admission process (e.g. admission-stop or admission-continue messages). For Termination, the PCN-egress-node signals to the ingress the flow ID of each flow to be terminated. Note that several IDs can be communicated in a single signalling (i.e. RSVP) message.

6. Dealing with ECMP

For admission control, neither of the three algorithms considered in this draft address ECMP issue in the absence of probing of some sort. The probing discussion is deferred to the next section.

Without probing, in the case when congestion state of different paths in the network differ, flows may be admitted on a congested path while the other path can remain uncongested, or, conversely, admission may stop on all paths, when only one path becomes pre-congested.

For termination, 3SM will correctly identify for termination only those flows which pass congested paths even in the presence of ECMP. In contrast, both CL and SM may erroneously terminate flows that do not traverse congested paths. For CL, an option is available to choose for termination only those flows that are termination-marked. However, doing so then requires that the egress needs to additionally signal to the ingress which flows have been marked, on top of the other signaling information described in Section 5.

Single-marking does not explicitly mark traffic by termination-marking and so the above option to identify flows of congested path does not work. SM does have an option to terminate only those flows that are marked for admission. However, this may result in erroneous termination of flows on paths where traffic is above the Admission threshold but below the level that should cause Termination. Just as in the case of CL, doing so also involves the necessity to signal additional information (flow IDs) between the PCN-egress-node and PCN-ingress node.

7. Suitability for Probing for Admission Control

Although probing is currently out-of-scope of the PCN WG charter, it may be useful in a number of situations (in particular, dealing with ECMP for admission, or addressing flash crowd situations in the presence of many small ingress-egress aggregates). We do not attempt

here to address the issue of whether or not and exactly how well probing might work, nor do we discuss any protocol issues and complexities associated with probing. Instead we touch upon only one aspect of it - how well the PCN admission marking of by the three proposals in question might work with probing.

Threshold-marking employed by both CL and 3SM results in all packets being marked once the traffic rate exceeds PCN-lower-threshold (i.e the admissible rate threshold). Therefore, a single probe packet is guaranteed to be marked if the PCN-traffic rate exceeds the PCN-lower-threshold threshold. Therefore, the admission decision can be reliably made based on a single probe. One possibility may be to use signaling (e.g. RSVP) messages as probes in this case.

In SM, admission metering and marking is based on Excess-rate-marking. In this case, only a fraction of traffic is marked when traffic exceeds the configured (admissible rate) threshold. Therefore, when the PCN-traffic rate exceeds this threshold, a single probe will only be marked with a certain probability, and so a series of probes need to be sent to detect congestion with high probability.

Thus, Threshold-marking of CL and 3SM allows faster and simpler admission decisions than Excess-rate-marking of SM.

In addition it should be noted that the necessity to send multiple probes for SM adds a potential problem with using RSVP messages as probes. Extensions necessary to do so have not been considered at this time.

8. Configuration Complexity and Configuratio Restrictions

The approach in SM requires a global configuration parameter at the edges reflecting the assumed ratio between the (implicit) termination threshold and (explicit) admission threshold, which is assumed to be global on all links in the PCN domain. This necessitates the use of a global parameter that needs to be configured to the same value on all PCN-boundary-nodes in the network. This clearly leads to an additional configuration complexity of SM compared to CL and 3SM.

An additional issue caused by the assumption of the global ratio between (implicit) termination threshold and (explicit) admission threshold of SM is related to the flexibility of resiliency planning. In this context, a natural approach is to use PCN-lower-threshold to represent the expected utilisation on different links for expected traffic matrix under normal, non-failure, conditions, and to view PCN-upper-threshold to represent expected worst case utilization under any of the "planned" failures.

As discussed in [Menth] and [I-D.charny-pcn-single-marking] , the global ratio between the termination and admission utilisation levels assumed in SM limits the flexibility of traffic engineering for resiliency to a certain extent. Specifically, While all three approaches can be configured to ensure that the planned matrix can be protected against all the planned failure conditions, the nature of the guarantee for the admitted traffic is different. Both CL and 3SM can be configured to protect all admitted traffic (but would not admit more than the planned traffic matrix), while SM can be configured to admit more traffic than planned, but will not guarantee protection against planned failures for traffic exceeding planned utilization for the planned traffic matrix.

It should be noted that in principle, all three algorithms can be configured to protect all admitted traffic (whether or not this admitted traffic exceeds the planned traffic matrix or not). However, as argued in [Menth], in this case SM can generally admit less traffic than CL or 3SM.

We refer the reader to [Menth] and [I-D.charny-pcn-single-marking] for a more detailed discussion on this issue.

9. Functional Comparison Summary

The following Table summarizes the differences between the three approaches discussed so far.

(preamble)

Comparison criteria	SM	3SM	CL
# of encoding encoding states	2	3	3
# of metering mechanisms in forwarding path of interior nodes	1	2	2
Type of metering & marking for admission	excess-rate marking	threshold marking	threshold or ramp marking

Type of metering and marking for termination	N/A	excess-rate with marking frequency reduction	excess-rate marking
Metering and remarking of previously marked packets	do not meter AM-marked packets	meter all pkts for admission and termination; do not re-mark TM-marked pkts	do not meter TM-marked pkts for termination meter all pkts; for admission, do not re-mark TM-marked pkts
Packet Drop preference	preferentially drop AM-marked packets; over-termination otherwise	no sensitivity to moderate drop of AM-marked pkts preferentially drop non-TM-marked packets-over-termination otherwise, especially under high loss	no sensitivity to moderate drop of AM-marked packets; preferentially drop TM-marked packets - over-termination otherwise
Egress Function	measure rates of marked and unmarked pkts, compute CLE, send both to ingress	observe packet markings, send control messages to control admission at ingress; send IDs of TM-marked packets to ingress	measure rates of AM/TM marked and unmarked packets compute CLE,send CLE and the rate of unmarked pkts to ingress
Ingress Termination function	compute sustainable rate; measure sending rate; compute rate to terminate, select flows to terminate	terminate those flows whose ids were signalled by the egress	measure sending rate; compute rate to terminate, select flows to terminate
Ingress Admission Function	stop admitting when CLE exceeds configured value; restart admitting when CLE	stop admitting when notified by egress; resume after a timeout or when notified by ingress	stop admitting when CLE exceeds configured value; restart admitting when CLE reduces below configured

	falls below configured value		value
rate measurement at boundary nodes	required 1 at ingress 2 at egress	optional 1 at egress	required 1 at ingress 2 at egress
Information signalled from egress to ingress	CLE, sustaina- ble(admission) rate	control messages to stop & possibly restart admission; ids of flows with TM-marked packets	CLE, sustainable (termination) rate
ECMP support for Termination	no; only partial support with additional complexity at the edge + signaling flow flow IDs from egress to ingress	yes	no; but full support with additional complexity at the edge + plus signalling flow IDs from egress to ingress
ECMP support for Admission	no w/out probes yes with probes but needs many probes; use of RSVP as probes not understood	no w/out probing yes with probing (needs one probe, can use RSVP as probe)	no w/out probing yes with probing (needs one probe, can use RSVP as probe)
Network-wide parameter configuration coordination	yes (one global parameter at the edges)	no	no
Support for resiliency planning	yes but weaker guarantees under planned failures for traffic excee- ding planned traffic matrix; if all admitted traffic is to be protected,	yes	yes

	SM can admit less traffic than CL or 3SM		

(Table 8.1. Functional Comparison of CL, SM and 3SM)

Finally, a few words are due on relative complexities of the three schemes. It should be clear from the above table that relative complexity has a number of dimensions. Specifically:

- o Complexity of the forwarding path. From that standpoint SM appears to be the simplest, as it requires only a single metering mechanism, CL appears to come next, closely followed by 3SM.

- o Complexity of conditional metering (i.e checking whether a particularly marked packet needs to be metered). From that standpoint, all three approaches appear to be comparable. (Note that while SM admission function is more complex from this point of view than admission functions of CL and 3SM, the additional complexity of not metering AM-marked packets has to be balanced against similar complexity of the termination functions of CL and 3SM).

- o Complexity of implementing preferential packet loss. From that standpoint all three approaches appear comparable, when both admission and termination functions are considered.

- o Complexity of boundary behaviors. From that standpoint 3SM appears to be the least complex, CL coming next and SM following closely.

- o Complexity of support for ECMP. From that standpoint 3SM requires no additional functionality, while CL and SM require extra complexity at the boundary nodes and extra signaling information exchange between the egress and the ingress (and in addition SM has only partial support - see section 6 and table 8.1 for its limitations).

- o Global configuration complexity. From that standpoint Cl and 3SM come first, with SM being the more complex as the only one requiring global parameter setting.

These comparisons are very crude and are only intended to roughly summarize the detailed differences described in Table 8.1.

10. Other Comparison Criteria

This section discusses a number of other comparison criteria that have not been studied in detail or are currently under consideration

1. Co-existence with RFC3168 (work in progress);

2. Multicast support (TBD, NOTE: this is out-of-scope for the current charter)

3. Future extensions to multiple domains ([I-D.briscoe-tsvwg-re-ecn-border-cheat] for CL, not clarity on other approaches; NOTE: this is out-of-scope of the current charter)

4. Extensibility to host-initiated PCN (3SM designed to support host-initiated PCN but performance analysis is ongoing; NOTE: this is out-of-scope for the current charter.

5. Extensibility to rate-adaptive traffic (TBD; NOTE: this is out-of-scope of the current charter)

6. Support of multiple precedence levels (TBD; NOTE: this is out-of-scope of the current charter)

7. Performance comparison (ongoing, see next Subsection).

10.1. Performance Comparison

As mentioned in the Introduction, at the time of writing this document several performance studies have been reported in [I-D.zhang-pcn-performance-evaluation], [I-D.charny-pcn-single-marking][I-D.babiarz-pcn-explicit-marking], and [TR437]. While the first two studies have attempted a careful side-by-side comparisons of CL and SM, the set of experiments reported is less extensive, and a number of differences existing in the simulation models and experiment setup make a side-by-side apples-to-apples comparison of the 3 schemes difficult. In this we will attempt to summarize performance evaluation criteria that could be used for comparison of the different approaches, as well as provide high level conclusions on some of them, where available studies allow those conclusions.

We refer the reader to [I-D.zhang-pcn-performance-evaluation], [I-D.charny-pcn-single-marking], [I-D.babiarz-pcn-explicit-marking], and [TR437] for details that could not be captured within the format of the table below.

DISCLAIMER: statements in the table below should be understood only in terms relative to the three considered algorithms and with respect to specific experiments performed; they and are intended for crude qualitative comparison only based on (ongoing) simulation studies as of the time of writing of this document . Quantification of these statements can be found in the corresponding performance studies referenced earlier in this section.

(preamble)

Comparison Criteria	SM	3SM	CL
Sensitivity to low bottleneck aggregation	poor perfor- mance at low bottleneck aggregation	good performance at low aggrega- tion reported (evaluation ongoing)	poor perfor- mance at low bottleneck aggregation
Sensitivity to low ingress- egress aggregation	performance degradation for both termination and admission under some traffic types	good in reported experiments; traffic and topology sensitivity study needed	admission insensitive; termination sensitive for some traffic types
Sensitivity to marking & measurement parameters	a range of params exist with consis- tent perfor- mance across a range of traffic types & topologies	evaluation ongoing slow-down param. S has the biggest impact on flow termination speed (its choice de- pends on topology and traffic rate)	relatively insensitive to marking and measurement params across a range of traffic types and topologies
Accuracy of admission across traf- fic types and topologies	good for large ingress-egress aggregation	good (but sensitivity to parameters TBD)	good
Accuracy of termination (single bottleneck)	over- termination compared to CL if termination trigger is not smoothed;	good (but sensitivity to params TBD)	good

	otherwise slower reaction than CL		
Accuracy of termination wet bottleneck utile. (multi-bottneck case)	more over-termination than CL	not reported	good
Reaction time time to termination	slower than CL with smoothing of termination trigger, fast otherwise to 3SM	slower than CL, comparison with SM TBD; parameter sensitivity TBD	fast
Sensitivity to large and small flows (termination)	not reported; flow selection at ingress more complicated (or less accurate with widely differrent rates; study ongoing	preferentially selects large flows; slow down parameter hard to select for mix of traffic rates (over-termination or slower reaction if slowdown parameter does not reflect average rate; evaluation ongoing	not reported ; flow selection at ingress more complicated (or less accurate) with widely differrent rates; (study ongoing)
Beat-down effect multi-bottleneck case)	more beatdown of long-haul aggregates than CL	not reported	beat-down of flows traversing more bottlenecks

(Table 9.1. Summary of (up-to-date) performance comparison.)

11. Unified Descriprion of Marking and Metering Functions

In this section we address the question of how the metering and marking functions of the three approaches can be described in a unified way so that different marking behaviors considered in this draft can be obtained by different parametrization choices. A potential benefit of such unified description is that a single PCN-

internal-node behavior could support a wider range of different PCN-boundary-node behaviors, and hence, potentially, can be of use under different deployment scenarios. We discuss the difficulties associated with such unified description in the following section, where we also raise the question of whether the benefits of such unified behavior outweigh a number of drawbacks discussed in that section.

In Figure 10.1, we show a relatively compact version of such unified algorithm using the notion of Token Bucket (TB) . This pseudocode does not support ramp-marking, as the current performance evaluation studies indicate little additional benefit of ramp-marking compared to threshold-marking. However, in the Appendix we present a (more complex) version of the marking algorithm that does support ramp-marking as well.

We note that the algorithm below (as well as a more complex complete version in the Appendix) can be further optimized. We make no optimization attempt in the interest of clarity.

The TB has a rate of TB.rate and a depth of TB.size. It has one marking thresholds TB.threshold to support threshold marking. If TB.threshold is configured to be greater than zero, then packets are marked if the TB fill state is below TB.threshold after their arrival and removal of tokens from the queue; otherwise packets are not marked. The "classic" token bucket used by SM and the termination function of CL is obtained by setting TB.theshold = 0.

In addition, the slowdown factor TB.slowdown is used to implement marking frequency reduction: TB.slowdown tokens are added to the token bucket when a packet is marked. The metering is applied only to packets whose marking is part of a specific subset that we call TB.meteredMarking. The TB.markingType indicates the type of codepoint that is used for marking. In addition, the TB has a variable TB.fill that records the number of tokens in the bucket and the variable TB.lastUpdate records the last update instant of the bucket. The global variable "now" indicates the current time. Packets have size packet.size (in bytes) and marking packet.mark. The algorithm is followed by a table describing the parameterization necessary to implement Admission and Termination Functions for different proposals.

```
   (preamble)
Parameters:
TB.rate: token rate of TB in bytes/s
TB.size: depth of TB in tokens (bytes)
TB.threshold: marking threshold of TB in bytes
TB.slowdown: slowdown factor for marking frequency reduction of TB in bytes
TB.markingType: PCN-first-encoding ("admission")  or
                PCN-second-encoding ("termination").
TB.meteredMarkings: set of packet markings that are eligible for metering by
TB,
                    it is a subset of ("unmarked", "admission", "termination").
                    NOTE: this set depends on whether it is admission or
                    termination that the function below is used for.
NOTE: settings of these parameters for different approaches are shown
in Tables 10.2 and 10.3

Input: packet
    // take passed time since last update into account
    TB.fill = min(TB.size, TB.fill+(now-TB.lastUpdate) * TB.rate);
    TB.lastUpdate = now;

    // meter and mark
    If (packet.mark in TB.meteredMarkings)
        if (TB.fill < packet.size)
            // re-marking of TM-marked packets to AM not allowed
            if (!(packet.mark == "termination" and TB.markingType ==
"admission"))
                packet.mark = TB.markingType;
            endif
        else
            TB.fill = TB.fill - packet.size;
            if (TB.fill < TB.threshold)
                // re-marking of TM-marked packets to AM not allowed
                if (!(packet.mark == "termination" and TB.markingType ==
"admission"))
                    packet.mark = TB.markingType;
                endif
            endif
        endif
    endif

    // marking frequency reduction
    if (packet.mark == "termination")
        TB.fill = min(TB.size, TB.fill + TB.slowdown);
    endif

Output: void
```

(Figure 10.1. Simple generalized metering and marking algorithm
based on TB-formulation)

(preamble)

	CL Admission	SM	3SM Admission
TB.rate	PCN-lower-threshold	PCN-lower-threshold	PCN-lower-threshold
TB.size	configured	configured	configured
TB.threshold	configured	0	configured
TB.slowdown	0	0	0
TB.metered-Markings	"unmarked" "admission" "termination"	"unmarked"	"unmarked" "admission" "termination"
TB.markingType	"admission"	"admission"	"admission"

(Figure 10.2 Admission Settings for the Three Algorithms)

(preamble)

	CL Termination	SM	3SM Termination
TB.rate	PCN-upper-threshold	N/A	PCN-upper-threshold
TB.size	configured	N/A	configured
TB.threshold	0	N/A	0
TB.slowdown	0	N/A	configured
TB.metered-Markings	"unmarked" "admission"	N/A	"unmarked" "admission" "termination"
TB.markingType	"termination"	N/A	"termination"

(Figure 10.3 Termination Settings for the Three Algorithms)

Figures 10.2 and 10.3 provide parameter settings corresponding to different metering and marking functions.

It should be noted, that when the algorithm described by the pseudocode in Figure 10.1 is used to perform threshold-marking, it has a slightly different behavior than the algorithm described in CL

and 3SM. The packet marking algorithm of 3SM bases its marking decision on TB.fill before tokens are removed while our algorithm makes the decision afterwards. In addition, the admission marking algorithms of both 3SM and CL set TB.fill=0 when TB.fill<packet.size while our algorithm leaves TB.fill unchanged in this case. This is done in the interest of simplification, as we believe that the impact of this change on performance will be minimal in practice. The pseudocode we provide in the Appendix is a complete (and more complex) version of the unified algorithm to address this issue. As mentioned earlier, this complete algorithm also supports the ramp marking. In our judgement, however, a simpler version of algorithm 10.1 would suffice in practice.

We also present an equivalent VQ formulation of the same algorithm in the Appendix.

12. Difficulties with Allowing Multiple Marking Behaviors

There are a number of difficulties associated with allowing multiple edge behaviors in the PCN framework. Below is a list of some of these difficulties.

- o Additional implementation complexity in the core devices needed to support multiple options.

- o Additional configuration complexity needed to support these different options (especially important in the case when different PCN domains configured for different options merge)

- o Differences in packet dropping preferences represent additional complexity if different policies need to be used with different options (although the exact impact of not implementing any dropping preferences at all for different algorithms is under study).

- o Difference in the PCN information signalled between the egress and ingress require definition and implementation of different signalling options

- o Additional complexity of standardization process

- o The unified specification is limited to the three approaches considered in this draft and does not consider any other possible approaches including that suggested in[I-D.westberg-pcn-load-control]

A question then arises whether the benefits of specifying a more

flexible unified core behavior outweigh the above drawbacks. We do not attempt to answer this question in this draft, but rather pose it for a general discussion of the WG.

Finally, it should be noted that not all of the above difficulties pertain to different combination of the approaches in the same degree. For example, the differences between SM and CL with respect to the PCN-boundary-node functions and the information signalled between the PCN-egress-node and PCN-ingress-node are relatively small compared to the substantial differences between the boundary behavior and information transport of 3SM compared to both CL and SM. Therefore, as described in [I-D.charny-pcn-single-marking], SM could be defined as a "stepping stone" to CL with relatively small additional complexity in such a way that the information transport is identical, and the boundary behavior is almost identical (and can be switched between the two schemes by a toggle of a configuration parameter). Thus, transition from SM to CL essentially amounts to upgrading the core nodes only. In contrast, transition from SM or CL to 3SM (or vice a versa) requires not only a change in the core behavior, but a substantial change in the boundary behavior and information transport.

13. Security Considerations

TBD

14. IANA Considerations

TBD

15. Appendix

15.1. Formulation of the Simple Generalized Metering and Marking Algorithm

The simple generalized metering and marking algorithm can also be described based on a virtual queue (VQ). The transformation is straightforward and the result is given in the algorithm shown in Figure 14.1.

(preamble)

```
Parameters:
VQ.rate: service rate of VQ in bytes/s
VQ.size: queue size of VQ in bytes
VQ.threshold: marking threshold of VQ in bytes
VQ.slowdown: slowdown factor for marking frequency reduction of VQ in bytes
VQ.markingType: PCN-first-encoding ("admission")  or  PCN-second-encoding
("termination").
VQ.meteredMarkings: set of packet markings that are eligible for metering by
VQ,
                    it is a subset of ("unmarked", "admission", "termination").

VQ.length: number of bytes currently in the queue of the VQ

Input: packet
    // take passed time since last update into account
    VQ.length = max(0, VQ.length-(now-VQ.lastUpdate) * VQ.rate);
    VQ.lastUpdate = now;

    // meter and mark
    If (packet.mark in VQ.meteredMarkings)
        if (VQ.length+packet.size > VQ.size)
            if (!(packet.mark == "termination" and VQ.markingType ==
"admission"))
                packet.mark = VQ.markingType;
            endif
        else
            VQ.length = VQ.length + packet.size;
            if (VQ.length > VQ.threshold)
                // re-marking of TM-marked packets to AM not allowed
                if (!(packet.mark == "termination" and VQ.markingType ==
"admission"))
                    packet.mark = VQ.markingType;
                endif
            endif
        endif
    endif

    // marking frequency reduction
    if (packet.mark == "termination")
        VQ.length = max(0, VQ.length - VQ.slowdown);
    endif

Output: void
```

(Figure 14.1)

The tables in Figs 14.3 and 14.4 at the end of this Section give the corresponding parameter setting for the three proposals.

NOTE: There are two further optimization options that are possible for this (and the TB-based) metering and marking algorithm:

1. If TM-packets are not metered for admission-marking, the inner if-clause to avoid the re-marking of TM-marked packets to AM can be removed. Current performance results suggest that this could be done, but for the sake of extensibility towards future rate adaptation it is better to keep open the option of metering all packets also for admission marking.

2. The marking frequency reduction may be applied when packets are re-marked to TM. This is at the expense of increased unfairness and over termination in the presence of several simultaneously overloaded links. The advantage is an improved runtime of the algorithm and a possibly simpler implementation.

15.2. VQ Formulation of the Complex Generalized Metering and Marking Algorithm

The simple generalized metering and marking algorithm in 14.1 has the following shortcomings:

1. It does not support ramp marking which may be desirable for CL;

2. If the packet does not fit into the queue, the queue cannot be filled up to its size. This is a property for admission marking in CL and 3SM that is not implemented by a simplified algorithm (although the impact of this change appears to be minimal);

3. If the packet fits into the queue, the marking decision is always based on the queue size including the size of the new packet. This leads to a nicer formulation of threshold or ramp marking (however, the impact of this change seems minimal) If the generalized algorithm takes also these 3 issues into account, it becomes significantly more complex.

To improve shortcoming 1, the algorithm has now two marking thresholds to support ramp and threshold marking instead of a single one: VQ.lowerThreshold and VQ.upperThreshold. Packets are not marked if the queue length is below VQ.lowerThreshold, the marking probability linearly increases from 0 to 1 between VQ.lowerThreshold and VQ.upperThreshold, and packets are definitely marked if the queue length is VQ.upperThreshold and above. If both thresholds are set to the same positive value, threshold marking is performed: packets are not marked if the queue length is below or equal to that threshold, otherwise they are marked.

To improve shortcoming 2 and 3, the Boolean variable VQ.alwaysUpdateState is introduced. If it is set to true, the queue length should always be increased by the packet size; this is the desired behaviour for admission marking when all packets are expected

to be marked when the admissible rate is exceeded by the PCN rate. If it is set to false, the queue length is only increased if the packet is not marked; this is the desired behaviour for excess-rate-marking.

```
   (preamble)
Additional parameters:
VQ.lowerThreshold: lower marking threshold of VQ in bytes
VQ.upperThreshold: upper marking threshold of VQ in bytes
VQ.alwaysUpdateState: VQ length state always increased or not

Input: packet

    // take passed time since last update into account
    VQ.length = max(0, VQ.length-(now-VQ.lastUpdate)*VQ.rate);
    VQ.lastUpdate = now;

    // meter and mark
    if (packet.mark in VQ.meteredMarkings)
        if (VQ.alwaysUpdateState == true)
            // threshold or ramp-marking
            if (VQ.length > VQ.upperThreshold)
                // re-marking of TM-marked packets to AM not allowed
                if (!(packet.mark == "termination" and VQ.markingType ==
"admission"))
                    packet.mark = VQ.markingType;
                endif
            elseif (VQ.length > VQ.lowerThreshold)
                choose random number u (0 < u < 1);
                if (u < (VQ.length-VQ.lowerThreshold)/(VQ.upperThreshold-
VQ.lowerThreshold))
                    // re-marking of TM-marked packets to AM not allowed
                    if (!(packet.mark == "termination" and VQ.markingType ==
"admission"))
                        packet.mark = VQ.markingType;
                    endif
                endif
            endif
            VQ.length = min(VQ.size, VQ.length+packet.size);
        else
            // excess-rate-marking
            if (VQ.length + packet.size > VQ.size)
                // re-marking of TM-marked packets to AM not allowed
                if (!(packet.mark == "termination" and VQ.markingType ==
"admission"))
                    packet.mark = VQ.markingType;
                endif
            else
```

```
            VQ.length = VQ.length + packet.size;
        endif
    endif
endif
```

```
    // marking frequency reduction
    if (packet.mark == "termination")
        VQ.length = max(0, VQ.length - VQ.slowdown);
    endif
```

Output: void

(Figure 14.2)

(preamble)

	CL Admission	SM	3SM Admission
VQ.rate	PCN-lower-threshold	PCN-lower-threshold	PCN-lower-threshold
VQ.size	configured	configured	configured
VQ.always-UpdateState	true	false	true
VQ.lowerThreshold	configured	0	configured
VQ.upperThreshold	configured	0	configured
VQ.slowdown	0	0	0
VQ.meteredMarkings	"unmarked" "admission" "termination"	"unmarked"	"unmarked" "admission" "termination"
VQ.markingType	"admission"	"admission"	"admission"

(Figure 14.3. Admission settings for the three algorithms (VQ formulation))

(preamble)

	CL Termination	SM	3SM Termination
VQ.rate	PCN-lower-threshold	N/A	PCN-upper-thrsehold
VQ.size	configured	N/A	configured
VQ.always-UpdateState	false	N/A	false
VQ.lowerThreshold	0	N/A	0
VQ.upperThreshold	0	N/A	0
VQ.slowdown	0	N/A	configured
VQ.meteredMarkings	"unmarked" "admission"	N/A	"unmarked" "admission" "termination"
VQ.markingType	"termination"	N/A	"termination"

(Figure 14.4. Termination settings for the three algorithms (VQ formulation))

16. References

16.1. Normative References

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

16.2. Informative References

[I-D.babiarz-pcn-3sm]
Babiarz, J., "Three State PCN Marking", draft-babiarz-pcn-3sm-00 (work in progress), July 2007.

[I-D.babiarz-pcn-explicit-marking]
Liu, X. and J. Babiarz, "Simulations Results for 3sM", draft-babiarz-pcn-explicit-marking-01 (work in progress), July 2007.

[I-D.briscoe-tsvwg-cl-architecture]
Briscoe, B., "An edge-to-edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region", draft-briscoe-tsvwg-cl-architecture-04 (work in progress), October 2006.

[I-D.briscoe-tsvwg-cl-phb]
Briscoe, B., "Pre-Congestion Notification marking", draft-briscoe-tsvwg-cl-phb-03 (work in progress), October 2006.

[I-D.briscoe-tsvwg-re-ecn-border-cheat]
Briscoe, B., "Emulating Border Flow Policing using Re-ECN on Bulk Data", draft-briscoe-tsvwg-re-ecn-border-cheat-01 (work in progress), June 2006.

[I-D.briscoe-tsvwg-re-ecn-tcp]
Briscoe, B., "Re-ECN: Adding Accountability for Causing Congestion to TCP/IP", draft-briscoe-tsvwg-re-ecn-tcp-04 (work in progress), July 2007.

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

[I-D.davie-ecn-mpls]
Davie, B., "Explicit Congestion Marking in MPLS", draft-davie-ecn-mpls-01 (work in progress), October 2006.

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

[I-D.lefaucheur-emergency-rsvp]
Faucheur, F., "RSVP Extensions for Emergency Services", draft-lefaucheur-emergency-rsvp-02 (work in progress), June 2006.

[I-D.westberg-pcn-load-control]
Westberg, L., "LC-PCN: The Load Control PCN Solution", draft-westberg-pcn-load-control-01 (work in progress), September 2007.

[I-D.zhang-pcn-performance-evaluation]
Zhang, X., "Performance Evaluation of CL-PHB Admission and

Termination Algorithms",
draft-zhang-pcn-performance-evaluation-02 (work in
progress), July 2007.

16.3. References

[Menth] "PCN-Based Resilient Network Admission Control: The Impact of a Single Bit", 2007.

[TR437] "Comparison of Marking Algorithms for PCN-Based Admission Control, Technical Report No. 437, University of Wuerzburg", October 2007.

Authors' Addresses

Anna Charny
Cisco Systems, Inc.
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: acharny@cisco.com

Joseph Babiarz
Nortel
3500 Carling Avenue
Ottawa, Ontario K2H 8E9
Canada

Email: babiarz@nortel.com

Michael Menth
University of Wuerzburg
Informatik III Am Hubland
Wuerzburg, 97074
Germany

Email: menth@menth@informatik.uni-wuerzburg.de

Joy Zhang
Cisco Systems, Inc & Cornell University
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: acharny@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgment

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