Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2007

A. Charny
F. Le Faucheur
V. Liatsos
Cisco Systems, Inc.
J. Zhang
Cisco Systems, Inc. and Cornell University

October 15, 2006

Pre-Congestion Notification Using Single Marking for Admission and Pre-emption

draft-charny-pcn-single-marking-00.txt

Status of this Memo

Copyright Notice

Abstract

Pre-Congestion Notification [I-D.briscoe-tsvwg-cl-architecture] approach proposes the use of an Admission Control mechanism to limit the amount of real-time PCN traffic to a configured level during the

normal operating conditions, and the use of a Pre-emption mechanism to tear-down some of the flows to bring the PCN traffic level down to a desirable amount during unexpected events such as network failures, with the goal of maintaining the QoS assurances to the remaining flows. In [I-D.briscoe-tsvwg-cl-architecture], Admission and Pre-emption use two different markings and two different metering mechanisms in the internal nodes of the PCN region. This draft proposes a mechanism using a single marking and metering for both Admission and Pre-emption, and presents a preliminary analysis of the tradeoffs. A side-effect of this proposal that a different marking and metering Admission mechanism than that proposed [I-D.briscoe-tsvwg-cl-architecture] may be also feasible.

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. Background and Motivation

Pre-Congestion Notification [I-D.briscoe-tsvwg-cl-architecture] approach proposes to use an Admission Control mechanism to limit the amount of real-time PCN traffic to a configured level during the normal operating conditions, and to use a Pre-emption mechanism to tear-down some of the flows to bring the PCN traffic level down to a desirable amount during unexpected events such as network failures, with the goal of maintaining the QoS assurances to the remaining flows. In [I-D.briscoe-tsvwg-cl-architecture], Admission and Pre-emption use different two different markings and two different metering mechanisms in the internal nodes of the PCN region. Admission Control algorithms for variable-rate real-time traffic such as video have traditionally been based on the observation of the queue length, and hence re-using these technques and ideas in the context of pre-congestion notification is highly attractive, and motivated the virtual-queue-based marking and metering approach specified in [I-D.briscoe-tsvwg-cl-architecture] for Admission. On the other hand, for Pre-emption, it is desirable to know how much traffic needs to be pre-empted, and that in turn motivates rate-based Pre-emption metering. This provides some motivation for employing different metering algorithm for Admission and for Preemption.

Furthermore, it is frequently desirable to trigger Pre-emption at a substantially higher traffic level than the level at which no new flows are to be admitted. There are multiple reasons for the requirement to enforce a different Admission Threshold and Preemption Threshold. These include, for example:

- End-users are typically more annoyed by their established call dying than by getting a busy tone at call establishment.

- There are often very tight (possibly legal) obligations on network operators to not drop established calls.

- Voice Call Routing often has the ability to route/establish the call on another network (e.g., PSTN) if it is determined at call establishment that one network (e.g., packet network) can not accept the call. Therefore, not admitting a call on the packet network at initial establishment may not impact the end-user. In contrast, it is usually not possible to reroute an established call onto another network mid-call. This means that call Preemption can not be hidden to the end-user.

- Preemption is typically useful in failure situations where some loads get rerouted thereby increasing the load on remaining links.

Because the failure may only be temporary, the operator may be ready to tolerate a small degradation during the interim failure period.

- o A congestion notification based Admission scheme has some inherent inaccuracies because of its reactive nature and thus may potentially over admit in some situations (such as burst of calls arrival). If the Preemption scheme reacted at the same Threshold as the Admission Threshold, calls may get routinely dropped after establishment because of over admission, even under steady state conditions.

These considerations argue for meetering for Admission and Pre-emption at different traffic levels and hence, implicitly, for different markings and metering schemes.

Different marking schemes require different codepoints. Thus, such separate markings consume valuable real-estate in the packet header, especially scarce in the case of MPLS Pre-Congestion Notification [I-D.davie-ecn-mpls] . Furthermore, two different measurement/ metering techniques involve additional complexity in the datapath of the internal routers of the PCN domain.

To this end, [I-D.briscoe-tsvwg-cl-architecture] proposes an approach, referred to as implicit pre-emption marking, that does not require separate pre-emption marking. However, it does require two separate measurement schemes: one measurement for Admission and another measurement for Preemption/Dropping. Furthermore, this approach mandates that the configured preemption rate be set to a drop rate. This approach effectively uses dropping as the way to convey information about how much traffic can fit under the preemption limit, instead of using a separate preemption marking. This is a significant restriction in that it results in preemption only taking effect once packets actually get dropped.

This document presents an approach that allows the use of a single PCN marking and a single metering technique at the internal devices without requiring that the dropping and pre-emption thresholds be the same. This document also investigates some of the tradeoffs associated with this approach.

1.2. Terminology

- o Pre-Congestion Notification (PCN): two algorithms that determine when a PCN-enabled router Admission Marks and Pre- emption Marks a packet, depending on the traffic level.

- Admission Marking condition- the traffic level is such that the router Admission Marks packets. The router provides an "early warning" that the load is nearing the engineered admission control capacity, before there is any significant build-up in the queue of packets belonging to the specified real-time service class.

- Pre-emption Marking condition- the traffic level is such that the router Pre-emption Marks packets. The router warns explicitly that pre-emption may be needed.

- Configured-admission-rate - the reference rate used by the admission marking algorithm in a PCN-enabled router.

- Configured-pre-emption-rate - the reference rate used by the pre-emption marking algorithm in a PCN-enabled router.

- CLE - congestion level estimate computed by the egress node by estimating as the fraction of admission-marked packets it receives.

- PIN - PCN internal node - an internal node in the PCN region.

- PEN - PCN edge node - an ingree or egress edge node of the PCN region.

2. The Single Marking Approach

2.1. High Level description

The proposed approach is based on several simple ideas:

- Replace virtual-queue-based marking for Admission Control by excess rate marking:

  * meter traffic exceeding the Admission Threshold and mark excess traffic (e.g. using a token bucket with the rate configured to Admission Rate Threshold)

  * at the edges, stop admitting traffic when the fraction of marked traffic for a given edge-to-edge aggregate exceeds a configured threshold (e.g. stop admitting when 3% of all traffic in the edge-to-edge aggregate received at the ingress is marked)

- Impose a PCN-region-wide constraint on the ratio between the Admission threshold on a link and Pre-emption threshold on that link (e.g. pre-emption threshold is 20% higher than Admission

threshold on all links in the PCN region)

- The edge PCN device determines whether Pre-emption level is reached anywhere in the network by measuring the amount of unmarked traffic (assuming the marked traffic actually is above the threshold triggering blocking admission), i.e. the traffic that did not get admission marked. This is analogous to the notion of sustainable pre-emption rate in [I-D.briscoe-tsvwg-cl-architecture] .

The remaining part of this section gives more detailed of a possible operation of the system.

2.2. Operation at the PIN

The PCN Internal Node (PIN) meters the aggregate PCN traffic and marks the excess rate. A number of implementations are possible to achieve that. A token bucket implementation is particularly attractive because of its relative simplicity, and even more so because a token bucket implementation is readily available in the vast majority of existing equipment. The rate of the token bucket is configured to correspond to the target Admission rate, and the depth of the token bucket can be configured by an operator based on the desired tolerance to PCN traffic burstiness.

Note that no preemption threshold is explicitly configured at the PIN, and the PIN does nothing at all to enforce it or mark traffic based on Pre-emption threshold.

2.3. Operation at the Egress PEN

The PCN Egress Node (PEN) measures the rate of both marked and unmarked traffic on a per-ingress PEN basis, and reports to the ingress PEN two values: the rate of unmarked traffic from this ingress PEN, which we deem Sustainable Admission Rate and the Congestion Level Estimate (CLE), which is the fraction of the marked traffic received from this ingress PEN. Note that Sustainable Admission Rate is analogous to the sustainable pre-emption rate of [I-D.briscoe-tsvwg-cl-architecture], except in this case it is based on the admission threshold rather than pre-emption threshold, while the CLE is exactly the same as that of [I-D.briscoe-tsvwg-cl-architecture]. The details of the rate measurement are outside the scope of this draft.

2.4. Operation at the Ingress PEN

2.4.1. Admission Decision

Just as in [I-D.briscoe-tsvwg-cl-architecture], the admission decision is based on the CLE. The ingress PEN stops admission of new flows if the CLE is above a pre-defined threshold (e.g. 3%). Note that although the logic of the decision is exactly the same as in the case of [I-D.briscoe-tsvwg-cl-architecture], the detailed semantics of the marking is different. This is because the marking used for admission in this proposal reflects the excess rate over the admission threshold, while in the marking is based on exceeding a virtual queue threshold. Notably, in the current proposal, if the average sustained rate of admitted traffic is 5% over the admission threshold, then 5% of the traffic is expected to be marked, whereas in the context of [I-D.briscoe-tsvwg-cl-architecture] a steady 5% overload should eventually result in 100% of all traffic being admission marked. A consequence of this is that for smooth traffic, the approach presented here will not mark any traffic at all until the rate of the traffic exceeds the configured admission threshold by the amount corresponding to the chosen CLE threshold. At first glance this may seem to result in a violation of the pre-congestion notification premise that attempts to stop admission before the desired traffic level is reached. However, in reality one can simply embed the CLE level into the desired configuration of the admission threshold. That is, if a certain rate X is the actual target admission threshold, then one should configure the rate of the metering device (e.g. the rate of the token bucket) to X-y where y corresponds to the level of CLE that would trigger admission blocking decision. A more important distinction is that virtual-queue based marking reacts to short-term busrtiness of traffic, while the excess-rate based marking is only capable of reacting to rate violations at the timescale chosen for rate measument. Whether this distinction is sufficiently important for the case when no actual queuing is expected even if the virtual queue is full is an open question, which we attempt to start answering in the performance evaluation presented at the end of this draft.

2.4.2. Pre-emption Decision

When the ingress observes a non-zero CLE and Sustainable Admission Rate Ra, it first computes the Sustainable Pre-Emption rate Rp by simply multiplying Ra by the system-wide constant u, where u is the system-wide ratio between pre-emption and admission thresholds on all links in the PCN domain: Rp = Ra*u. The ingress PEN then performs exactly the same operation as is proposed in [I-D.briscoe-tsvwg-cl-architecture] with respect to Rp, namely, it pre-empts the appropriate number of flows to ensure that the rate of traffic it sends to the corresponding egress PEN does not exceed the sustainable pre-emption rate Rp. Just as in the case of

[I-D.briscoe-tsvwg-cl-architecture], an implementation may decide to slow down the pre-emption process by preempting fewer flows than is necessary to cap its traffic to Rp by employing a variety of techniques such as safety factors or hysteresis. In summary, the operation of pre-emption at the ingress PEN is identical to that of [I-D.briscoe-tsvwg-cl-architecture], with the only exception that the sustainable pre-emption rate is computed from the sustainable admission rate rather than derived from a separate marking. This is enabled by imposing a system-wide restriction on the pre-emption-to-admission thresholds ratio and changing the semantics of the admission marking.

3. Discussion

3.1. Benefits

The key benefits of using a single metering/marking scheme for both Admission and Preemption presented in this document are summarized below:

- Reduced implementation requirements on core routers due to a single metering implementation instead of two different ones.

- Ease of use on existing hardware: given that the proposed approach is particularly amenable to a token bucket implementation, the availability of token buckets on virtually all commercially available routers makes this approach especially attractive.

- Reduced number of codepoints which need to be conveyed in the packet header. If the PCN-bits used in the packets header to convey the congestion notification information are the ECN-bits in an IP core and the EXP-bits in an MPLS core, as is currently proposed in [put marking draft reference here] and [I-D.davie-ecn-mpls], those are very expensive real-estate. The current proposals need 5 codepoints, which is especially important in the context of MPLS where there is only a total of 8 EXP codepoints which must also be shared with Diffserv. Eliminating one codepoint considerably helps.

- A possibility of using a token-bucket-, excess-rate- based implementation for admission provides extra flexibility for the choice of an admission mechanism, even if two separate markings and thresholds are used.

3.2. Tradeoffs and Issues

While the benefits of the proposed approach are attractive, there are several issues and tradeoffs that need to be carefully considered.

3.2.1. Restrictions on Pre-emption-to-admission Thresholds

An obvious restriction necessary for the single-marking approach is that the ratio of 9implicit) pre-emption and admission thresholds remains the same on all links in the PCN region. While clearly a limitation, this does not appear to be particularly crippling, and does not appear to outweigh the benefits of reducing the overhead in the router implementation and savings in codepoints.

3.2.2. Performance Implications and Tradeoffs

Replacement of a relatively well-studied queue-based measurement-based admission control approach by a cruder excess-rate measurement technique raises a number of algorithmic and performance concerns that need to be carefully evaluated. For example, a token-bucket excess rate measurement is expected to be substantially more sensitive to traffic burstiness and parameter setting, which may have a significant effect in the case of lower levels of traffic aggregation, especially for variable-rate traffic such as video. In addition, the appropriate timescale of rate measurement needs to be carefully evaluated, and in general it depends on the degree of expected traffic variability which is frequently unknown.

In view of that, an initial performance comparison of the token-bucket based measurement is presented in the following section. Within the constraints of this preliminary study, the performance tradeoffs observed between the queue-based technique suggested in [I-D.briscoe-tsvwg-cl-architecture] and a simpler token-bucket-based excess rate measurement do not appear to be a cause of substantial concern for cases when traffic aggregation is reasonably high at the bottleneck links as well as on a per ingress-egress pair basis. Details of the simulation study, as well as additional discussion of its implications are presented in section 4.

Also, one mitigating consideration in favor of the simpler mechanism is that in a typical DiffServ environment, the real-time traffic is expected to be served at a higher priority and/or the target admission rate is expected to be substantially below the speed at which the real-time queue is actually served. If these assumptions hold, then there is some margin of safety for an admission control algorithm, making the requirements for admission control more forgiving to bounded errors - see additional discussion in section 4.

Note that an implication of the above that even if two markings and two metering mechanisms are used, these consideration may imply that an excess-rate token bucket implementation of admission metering and marking may be feasible, which could be a benefit for existing equipment routinely supporting a token-bucket implementation.

3.2.3. Effect on Proposed Anti-cheating Mechanisms

Replacement of the queue-based admission control mechanism of [I-D.briscoe-tsvwg-cl-architecture] by an excess-rate based admission marking changing the semantics of the pre-congestion marking, and consequently interfereres with mechanisms for cheating detection discussed in [I-D.briscoe-tsvwg-re-ecn-border-cheat]. Implications of excess-rate based marking on the anti-cheating mechanisms need to be considered.

3.2.4. Standards Implications

The change of the meaning of admission marking for pre-congestion notification from the queue-based to excess-rate marking poses a question of coexistence of devices having different interpretation of admissions marking (and hence different metering and marking mechanisms in the core. The question of how and if the two mechanisms can co-exist in one PCN region has obvious impact on standardization efforts, and needs to be carefully considered.

4. Performance Evaluation Comparison

4.1. Relationship to other drafts

Initial simulation results of admission and pre-emption mechanisms of [I-D.briscoe-tsvwg-cl-architecture] were reported in [I-D.briscoe-tsvwg-cl-phb]. A follow-up study of these mechanisms is presented in a companion draft draft-zhang-cl-performance-evaluation-00.txt. The current draft concentrates on a preliminary performance comparison of the admission control mechanism of [I-D.briscoe-tsvwg-cl-phb] and the token-bucket-based admission control described in section 2 of this draft.

4.2. Limitations, Conclusions and Direction for Future Work

4.2.1. Limitations

Due to time constraints, the study performed so far was limited to a single bottlneck case. The key questions that have been investigated are the comparative sensitivity of the two schemes to parameter settings and the effect of traffic burstiness and of the degree of

aggregation on a per ingress-egress pair on the performance of the admission control algorithms under study. The study is limited to the case where there is no packet loss. While this is a reasonable initial assumption for an admission control algorithm that is supposed to maintain the traffic level significantly below the service capacity of the corresponding queue, nevertheless future study is necessary to evaluate the effect of packet loss.

This draft does not discuss performance of the pre-emption algorithm, as it does not differ between the approach described in this draft and that of draft [I-D.briscoe-tsvwg-cl-architecture].

4.2.2. High Level Conclusions

The results of this (preliminary) study indicate that there is some potential that a reasonable complexity/performance trafdeoff may be viable for the choice of admission control algorithm. In turn, this suggests that using a single codepoint and metering technique for admission and pre-emption may be a viable option warranting further investigation.

The key high-level conclusions of the simulation study comparing the performance of queue-based and token-based admission control algorithms are summarized below:

1. At reasonable level of aggregation at the bottleneck and per ingress-egress pair traffic, both algorithms perform reasonably well for the range of traffic models considered (see section 4.3. for detail).

2. Both schemes are stressed for small levels of ingress-egress pair aggregation levels (e.g. a single video-like bursty VBR flow per ingress-egress pair). However, while the queue-based scheme results in tolerable performance even at low levels of per ingress-egress aggregation, the token-bucket-based scheme is substantially more sensitive to parameter setting than the queue-based scheme, and its performance for the high rate bursty "video-like" traffic with low levels of ingress-egress aggregation is quite poor unless parameters are chosen carefully to curb the error.

3. Even for small per ingress-egress pair aggregation, reasonable performance across a range of traffic models can be obtained for both algorithms (with a narrower range of parameter setting for the token-bucket based approach) .

4. The absolute value of round-trip time (RTT) or the RTT difference between different ingress-egress pair within the range of

continental propagation delays does not appear to have a visible effect on the performance of both algorithms.

4.2.3. Future work

This study is but the first step in performance evaluation of the token-bucket based admission control. Further evaluation should include a range of investigation, including the following:

- a study in the multiple bottleneck case
- a wider range of topologies and traffic matrices
- fairness issues (how different ingress-egress pairs get access to bottleneck bandwidth)
- interactions between admission control and preemption
- effect of loss of marked packets
- much more

5. Appendix A: Simulation Details

5.1. Network and Signaling Model

Simulations presented in this document are limited to a single bottleneck case. The network is modeled as either Single Link or Multi Link Network shown in the figures below. (We termed the latter "RTT").

Figure A.1: Simulated Single Link Network.

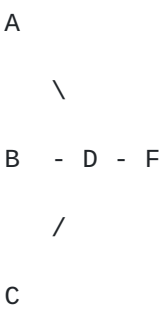

Figure A.2: Simulated Multi Link Network.

Figure A.1 shows a single link between an ingress and an egress node, all flows enter at node A and depart at node B. In Figure A.2, A set of ingresses (A,B,C) connected to an interior node in the network (D). The number of ingresses varied in different simulation experiments in the range of 2-100. All links have generally different propagation delays, in the range 1ms - 100 ms. This node D in turn is connected to the egress (F). In this topology, different sets of flows between each ingress and the egress converge on the single link D-F, where pre-congestion notification algorithm is enabled. The capacities of the ingress links are not limiting, and hence no PCN is enable on those. The bottleneck link D-F is modeled with a 10ms propagation delay in all simulations. Therefore the range of round-trip delays in the experiments is from 22ms to 220ms. Our simulations concentrated primarily on capacities of 'bottleneck' links with sufficient aggregation - OC3 for voice and for "video-like" traffic, up to 1 Gbps. In the simulation model, a call requests arrive at the ingress and immediately sends a message to the egress. The message arrives at the egress after the propagation time plus link processing time (but no queuing delay). When the egress receives this message, it immediately responds to the ingress with the current Congestion-Level-Estimate. If the Congestion-Level-Estimate is below the specified CLE-threshold, the call is admitted, otherwise it is rejected. An admitted call sends packets according to one of the chosen traffic models for the duration of the call (see next section). Propagation delay from source to the ingress and from destination to the egress is assumed negligible and is not modeled.

5.2. Traffic Models

Three types of traffic were simulated (CBR voice, on-off traffic approximating voice with silence compression, and on-off traffic with higher peak and mean rates (we termed the latter "video-like" as the chosen peak and mean rate was similar to that of an mpeg video stream, although no attempt was made to match any other parameters of this traffic to those of a video stream). The distribution of flow

duration was chosen to be exponentially distributed with mean 2min, regardless of the traffic type. In most of the experiments flows arrived according to a Poisson distribution with mean arrival rate chosen to achieve a desired amount of overload over the configured-admission-limit in each experiment. Overloads in the range 2x to 5x and underload with 0.95x have been investigated. For on-off traffic, on and off periods were exponentially distributed with the specified mean. Traffic parameters for each flow are summarized below:

5.2.1. CBR Voice (CBR)

- Average rate 64 Kbps
- Packet length 160 bytes
- packet inter-arrival time 20ms

5.2.2. VBR Voice (VBR)

- Packet length 160 bytes
- Long-term average rate 21.76 Kbps
- On Period mean duration 340ms; during the on period traffic is sent with the CBR voice parameters described above
- Off Period mean duration 660ms; no traffic is sent during the off period.

5.2.3. High Rate ON-OFF traffic with Video-like Mean and Peak Rates ("Video")

- Long term average rate 4 Mbps
- On Period mean duration 340ms; during the on-period the packets are sent at 12 Mbps (1500 byte packets, packet inter-arrival: 1ms)
- Off Period mean duration 660ms

5.3. Parameter Settings

5.3.1. Queue-based settings

All the queue-based simulations were run with the following Virtual Queue thresholds:

- virtual-queue-rate: configured admission rate, 1/2 link speed
- min-marking-threshold: 5ms at virtual-queue-rate
- max-marking-threshold: 15ms at virtual-queue-rate
- virtual-queue-upper-limit: 20ms at virtual-queue-rate

At the egress, the CLE is computed as an exponential weighted moving average (EWMA) on an interval basis, with 100ms measurement interval chosen in all simulations. We simulated the weight ranging 0.1 to 0.9. The CLE threshold is chosen to be 0.05, 0.15, 0.25, and 0.5.

5.3.2. Token Bucket Settings

The token bucket rate is set to the configured admission rate, which is half of the link speed in all experiments. Token bucket depth ranges from 64 to 512 packets. Our simulation results indicate that depth of token bucket has no significant impact on the performance of the algorithms and hence, in the rest of the section, we only present the result with 64 bucket depth.

The CLE is calculated in the same way as in queue-based approach with weights from 0.1 to 0.9. The CLE thresholds are chosen to be 0.0001, 0.001, 0.01, 0.05. Note that the meaning of tthe CLE is different for the Token bucket and queue-based algorithms, so there is no direct correspondence between the choice of the CLE thresholds in the two cases.

5.4. Simulation Details

To evaluate the performance of the algorithms, we recorded the actual admitted load at a granularity of 100ms, from which the mean admitted load over the duration of the simulation run can be computed. We verified that the actual admitted load at any time does not deviate much from the mean admitted load in each experiment by computing the coefficient of variation (CV is consistently 0.06 for CBR, 0.13 for VBR and 0.6 for Video for all experiments). Finally, the performance of the algorithms is evaluated using a metric called over-admission-percentage, which is calculated as a percentage difference between the mean admitted load and the configured admission rate. Given reasonably small deviation of the admitted rate from the mean admitted in the experiments, this seems reasonable.

5.4.1. Queue-based Results

We found that virtual-queue admission control algorithm works reliably with the range of parameters we simulated, for all three

types of traffic. In addition, for both CBR and VBR traffic, the performance is insensitive to the parameters change. Table A.1 summarized the over-admission-percentage values from 32 experiments with different [weight, CLE threshold] settings. The overload column represents the ratio of the demand on the bottleneck link to the configured admission threshold. While in our simulations we tested the range of overload from 0.95 to 5, we present here only the results of the endpoints of this overload interval. For the intermediate values of overload the results are even closer to the expected than at the two boundary loads, The statistics show that for CBR and VBR traffic these over-admission-percentage values are rather similar, with the admitted load staying within -2%+2% range of the desired admission threshold, with quite limited variability across the experiments (see the Standard Deviation column)

Over Admission Perc Stats					Over Load	Topo	Type
Min	Median	Mean	Max	SD			
0.007	0.007	0.007	0.007	0	0.95	S.Link	CBR
0.224	0.792	0.849	1.905	0.275	5		
0.008	0.008	0.008	0.008	0	0.95	RTT	
0.200	0.857	0.899	1.956	0.279	5		
-1.45	-0.96	-0.98	-0.86	0.117	0.95	S.Link	VBR
-0.07	1.507	1.405	1.948	0.421	5		
-1.56	-0.75	-0.80	-0.69	0.16	0.95	RTT	
-0.11	1.577	1.463	2.199	0.462	5		

Table A.1 Summarized performance for CBR and VBR across different settings.

For Video-like high-rate VBR traffic, the algorithms does show certain sensitivity to the tested parameters. Table A.2 recorded the over-admission-percentage for each combination of weights and CLE threshold.

		EWMA Weights					Over Load	Topo
		0.1	0.3	0.5	0.7	0.8		
C L E T H R	0.05	-4.87	-3.05	-2.92	-2.40	-2.40		Single Link
	0.15	-3.67	-2.99	-2.40	-2.40	-2.40	0.95	
	0.25	-2.67	-2.40	-2.40	-2.40	-2.40		
	0.5	-0.24	-1.60	-2.40	-2.40	-2.40		
E S H O L D	0.05	-4.03	2.52	3.45	5.70	5.17		RTT
	0.15	-0.81	3.29	6.35	6.80	8.13	5	
	0.25	2.15	5.83	6.81	8.62	7.95		
	0.5	6.55	9.35	9.38	8.96	8.41		
E S H O L D	0.05	-11.77	-8.35	-5.23	-2.64	-2.35		RTT
	0.15	-9.71	-7.14	-2.01	-2.21	-1.13	0.95	
	0.25	-5.54	-6.04	-3.28	-0.88	-0.27		
	0.5	-2.00	-2.56	-1.52	0.53	0.39		
	0.05	-5.04	-0.65	4.21	6.65	9.90		
	0.15	-1.02	1.58	7.21	8.24	10.07	5	
	0.25	-0.76	1.96	7.43	9.66	11.26		
	0.5	6.70	8.42	10.10	11.11	11.02		

Table A.2 Over-admission-percentage for "Video"

It follows from these results that while performance is tolerable across the entire range of parameters, choosing the CLE and EWMA weights in the middle of the tested range appear to be more beneficial for the overall performance across the chosen range of overload, assuming the chosen values for the remaining parameters. The high level conclusion that can be drawn from Table A.2. is that (predictably) high peak-to-mean ratio video-like traffic is substantially more stressful to the queue-based admission control algorithm, but a set of parameters exists that keeps the overadmission within about -3% - +10% of the expected load even for the bursty video-like traffic. Note that for vide-like traffic these results hold even though there is no aggregation at all on a per-ingress-egress pair in the chosen RTT topology there is only a single "video" flow per ingress.

5.4.2. Token Bucket-based Results

Compared to the virtual queue based algorithms, token bucket-based admission control algorithm shows substantially higher sensitivity to the parameter settings for the over-load conditions (overload greater than 1) Under the under-loaded conditions for voice-like CBR and VBR traffic the sensitivity to the tested parameters remains limited for

the token-bucket as well (the latter is summarized in Table A.3).

Over Admission Perc Stats					Load	Topo	Type
Min	Max	Median	Mean	SD			
0.007	0.007	0.007	0.007	0	0.95	S.Link	CBR
0.008	0.008	0.008	0.008	0	0.95	RTT	
-2.00	-0.95	-1.02	-0.78	0.268	0.95	S.Link	VBR
-2.83	-1.20	-1.31	-0.70	0.510	0.95	RTT	

Table A.3 Summarized performance for CBR and VBR across different settings for under-loaded conditions.

Table A.4 shows over-admission-percentage for different settings. It is important to note here that for the token bucket-based admission control no traffic will be marked until the rate of traffic exceeds the configured admission rate by the chosen CLE. As a consequence, even with the ideal performance of the algorithms, the over-admission-percentage will not be 0, rather it is expected to equal to CLE threshold if the algorithm performs as expected. Therefore, a more meaningful metric for the token-based results is actually the over-admission-percentage (listed below) minus the corresponding (CLE threshold * 100). For example, for CLE = 0.05, one would expect that 5% overadmission is inherently embedded in the algorithm, with the algorithm by design reacting to 5% overload (or more) only. Hence, with CLE = 0.05 a 10% over-admission in the token-bucket case should be compared to a 5% overadmission in the queue-based algorithm. When comparing the performance of token bucket (with the adjusted over-admission-percentage) to its corresponding virtual queue result, we found that token bucket performs only slightly worse for voice-like CBR and VBR traffic.

However the results for Video-like traffic require some additional commentary. Note from the results in Table A.4. that even for video-like traffic, in the Single Link topology the performance of the token-based solution is comparable to the performance of the queue-based scheme in table A.2, (adjusted by the CLE as discussed above). However, for the RTT topology, especially for the larger EWMA weights, the performance for "video" traffic becomes very bad, with up to 48% (adjusted by CLE) over-admission in a high overload situation (5x). We investigated two potential causes of this drastic degradation of performance by concentrating on two key differences between the Single Link and the RTT topologies: the difference in the

round-trip times and the degree of aggregation in a per ingress-egress pair aggregate.

To investigate the effect of the difference in round-trip times,we also conducted a subset of the experiments described above using the RTT topology that has the same RTT across all ingress-egress pairs rather than the range of RTTs in one experiment. We found out that neither the absolute nor the relative difference in RTT between different ingress-egress pairs appear to have any visible effect on the over-load performance or the fairness of both algorithms (we do not present these results here as their are essentially identical to those in Table A.4). In view of that and noting that in the RTT topology we used for these experiments for the video-like traffic, there is only 1 highly bursty flow per ingress, we believe that the severe degradation of performance in this topology is directly attributable to the lack of traffic aggregation on the ingress-egress pair basis. We also note that even for this highly challenging scenario, it is possible to find a range of parameters that limit the overadmission case for video traffic to quite a reasonable range of -3% + 10% (adjusted by the CLE). Luckily, these are the same parameter settings that work quite well for the other types of traffic tested.

		EWMA Weights					Over Load	Topo	Type
		0.1	0.3	0.5	0.7	0.9			
C	0.0001	-0.99	0.09	0.24	0.41	0.43			
L	0.001	0.02	0.37	0.43	0.46	0.45	5	S.	
E	0.01	1.37	1.32	1.32	1.31	1.31		Link	
	0.05	5.61	5.58	5.60	5.58	5.57			
T									CBR
H	0.0001	6.50	7.96	8.37	8.42	8.84			
R	0.001	7.07	8.54	8.65	8.55	8.66	5	RTT	
S	0.01	7.93	9.08	8.71	8.63	9.40			
	0.05	11.01	10.59	10.86	10.39	10.51			
C	0.0001	-2.95	-1.39	-0.63	0.23	0.78			
L	0.001	-1.51	-0.23	0.44	0.63	1.39	5	S.	
E	0.01	1.37	2.01	2.29	2.60	2.76		Link	
	0.05	6.31	6.71	6.80	6.97	7.05			
T									VBR
H	0.0001	-1.93	-0.23	0.99	2.09	3.38			
R	0.001	-0.75	0.89	2.07	3.12	4.27	5	RTT	
S	0.01	1.91	3.42	4.35	5.36	6.38			
	0.05	7.69	9.22	10.22	11.27	12.06			
	0.0001	-10.67	-10.58	-7.95	-6.27	-4.99			
	0.001	-8.67	-8.04	-7.61	-4.37	-2.89	0.95		
	0.01	-4.28	-2.59	-4.44	-2.13	-2.20			
C	0.05	-0.24	-0.66	-1.08	-0.92	-0.23		S.	
L								Link	
E	0.0001	-16.36	-10.24	-6.50	-2.17	2.74			
	0.001	-10.54	-5.63	-2.70	0.94	3.54	5		
T	0.01	-4.11	1.26	5.38	5.75	8.82			
H	0.05	6.31	10.49	11.75	14.21	15.08			
R									Video
E	0.0001	-15.83	-10.35	-2.96	0.17	5.42			
S	0.001	-12.82	-7.62	-0.47	2.24	6.59	0.95		
H	0.01	-6.17	-0.11	2.16	5.28	10.34			
O	0.05	0.52	6.14	7.34	9.32	14.07			
L								RTT	
D	0.0001	-8.51	1.86	11.14	22.51	30.24			
	0.001	-4.80	1.49	15.35	24.56	33.96	5		
	0.01	0.56	8.26	25.71	35.63	42.72			
	0.05	14.08	19.69	32.50	39.55	52.28			

Table A.4. Token bucket admission control performance.

6. IANA Considerations

This document places no requests on IANA.

7. Security Considerations

TBD

8. References

8.1. Normative References

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

8.2. Informative References

[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-03 (work in progress), June 2006.

[I-D.briscoe-tsvwg-cl-phb]
Briscoe, B., "Pre-Congestion Notification marking", draft-briscoe-tsvwg-cl-phb-02 (work in progress), June 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-02 (work in progress), June 2006.

[I-D.davie-ecn-mpls]
Davie, B., "Explicit Congestion Marking in MPLS", draft-davie-ecn-mpls-00 (work in progress), June 2006.

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

Authors' Addresses

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

Email: acharny@cisco.com

Francois Le Faucheur
Cisco Systems, Inc.
Village d'Entreprise Green Side - Batiment T3 ,
400 Avenue de Roumanille, 06410 Biot Sophia-Antipolis,
France

Email: flefauch@cisco.com

Vassilis Liatsos
Cisco Systems, Inc.
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: vliatsos@cisco.com

Xinyang (Joy) Zhang
Cisco Systems, Inc. and Cornell University
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: joyzhang@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgment

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