TSVWG
Internet-Draft
Expires: January 12, 2006

J. Babiarz
S. Dudley
K. Chan
Nortel Networks

July 11, 2005

Discussion of Congestion Marking with RT-ECN
draft-babiarz-rtecn-marking-00

Status of this Memo

Copyright Notice

Abstract

At the 62nd IETF meeting, it was requested that the authors of Congestion Notification Process for Real-Time Traffic (RT-ECN) draft look at rate proportional marking as an method of indicating that traffic has exceeded a configured rate. In version 03 of RT-ECN draft (draft-babiarz-tsvwg-rtecn-03) we stated, when the rate exceeds the engineered traffic level, all packets as indicated by a DS codepoint from ECN-capable end-systems are marked to indicate

congestion for the duration of the experienced congestion. In this memo, we looked at the two approaches, provide analysis as well our conclusions.

Table of Contents

1. Introduction

At the 62nd IETF meeting, it was requested that the authors of Congestion Notification Process for Real-Time Traffic (RT-ECN) draft look at rate proportional marking as an method of indicating that traffic has exceeded a configured rate. In version 03 of [RT-ECN] draft we stated, when the rate exceeds the engineered traffic level, all packets as indicated by a DS codepoint coming from ECN-capable end-systems are marked to indicate congestion for the duration of the experienced congestion. We will refer to it as "threshold based" marking.

Our understanding of rate proportional marking is that if the measured traffic rate as indicated by a specific DS codepoint is exceeded by h%, that h% of traffic as a rate needs to be ECN marked. Our definition of threshold based marking is that when a rate is exceeded, all packets that are marked with the specific DS codepoint are ECN marked until the traffic rate drops below the measured threshold. The duration of ECN marking can be control by the fill level of the token bucket above empty. Both of these metering and marking approaches can be done using token bucket or other methods.

RT-ECN draft proposes a new set of ECN semantics to provide two levels of congestion as well metering and marking behavior. It is applied to real-time inelastic flows such as VoIP and video conferencing for control of admission and preemption of real-time flows.

As the analyses are some what lengthy, we will present our conclusions first followed by the detailed analysis of threshold based and rate proportional marking for flow admission control and preemption.

1.1 Requirements notation

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 [RFC2119].

2. Conclusions

Both rate proportional marking and threshold based marking approaches were compared for two different uses in the network. The first use is for admission of new flows into the network. The second is for preemption of existing flows.

As a result of our analysis, we believe that rate proportional marking of real-time traffic is not appropriate for admission control of new flows in to the network. Threshold based marking provides a much faster and more deterministic indication that traffic on the path is congested (exceeds configured level). Rate proportional marking approach would be inappropriate for use in situations where Service Level Agreements for bandwidth management are required. Since this is the most likely scenario for the use of admission control, this shortcoming severely handicaps its applicability.

For flow preemption, both the rate proportional and threshold based marking methods can work for a network where all flows have a single or no precedence (flow importance). For networks that need to support one or more level of precedence, the threshold based approach should be used. The threshold based approach works under all network conditions, traffic flow scenarios and with multi precedence levels for traffic within a service class. Rate proportional marking, if implemented as a strictly random marking process, could lead to situations where the percentage of marking does not represent the rate of congestion experienced on the end-to-end path. The second issue arises when there are two or more precedence (flow importance) levels of traffic being managed. A strictly random marking process is not flow precedence aware, and it may result in higher precedence traffic being targeted for preemption when lower precedence traffic is still present on the link. Again, from the perspective of the most likely use cases, this is highly undesirable. A rate proportional based marking approach does however provide additional information to end systems that may be used to assess the relative severity of congestion on the network or for comparison of which flow has encountered more congestion. A rate proportional marking approach could be used in networks were there are no flow precedence levels.

Further discussion is needed on whether threshold based and rate proportional marking should both be allowed for flow preemption with RT-ECN. Note, it is felt that both approaches could be used as long the marking method matches the end systems expectation.

3. Analysis for Admission Control

For admission control of new flows into the network, it is desirable that feedback about traffic level (congestion level) on the flow's path is fast, as call setup delay is a critical parameter that users of the VoIP service look at. For RT-ECN, during call setup, a signaling protocol such as SIP is used to trigger the sending of RTP probe packets in both directions along the path that voice or video will take to test the current traffic level. The RTP probe flow uses the same source/destination IP address and port number as the media to guarantee that the path is identical. The data rate of the probe flow is deliberately held to the minimum to avoid affecting the link more than necessary.

Taking the case of the RTP probe stream for a rate proportional marking approach, consider the case where 100 flows have been admitted and a RTP probe is used to test for admission of the 101st flow. The threshold for admission is at a data rate equivalent to 100 flows. Under this condition, there is a high probability that the ECN marking of the RTP probe packets would indicate that the path is not congested and the new flow would be admitted. The point at which the system could be guaranteed to deny admission would be higher than the threshold point. If the reason to set the threshold is to meet the needs of a Service Level Agreement, this make the process of choosing threshold for implementing the system somewhat problematic. The choice of threshold is affected by the expected traffic level as is illustrated in the following analysis. The analysis is a first level approximation of feedback mechanisms in the rate proportional and threshold based marking systems.

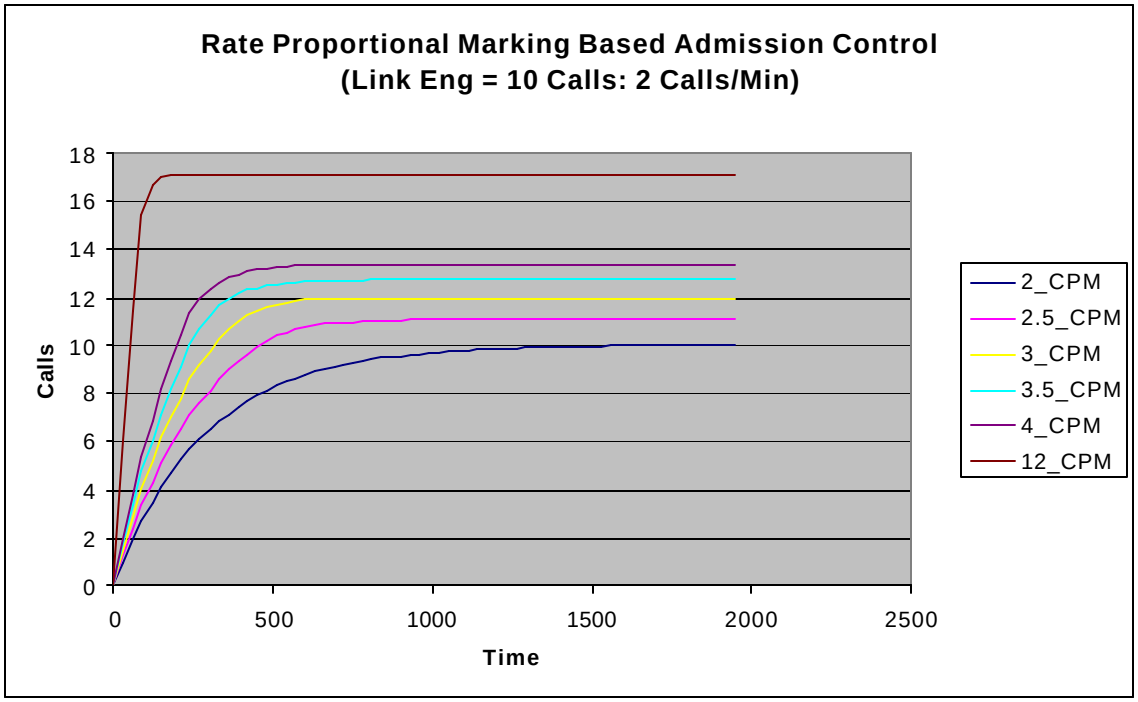
3.1 Rate Proportional versus Threshold Based Marking

The high level performance analysis of rate proportional marking versus threshold based marking for admission control is provided in this section. The charts shown here are representative only of some features of the two systems. They were generated using a spreadsheet and calculating using call arrival rates. To simplify the calculations, no attempt was made to accommodate randomness in arrivals, enforcement of whole calls being admitted within each calculation interval, or show the effects of latency in the network. All of these would affect both schemes and make them much less smooth than these charts show but the long term average number of calls admitted would be the same.

The first chart illustrates the rate proportional based admission control scheme. The charts show a link engineered for 10 calls. The call arrival rate (how many new calls arrive per minute) is varied. Call hold time (how long the call remains active) is fixed at 300

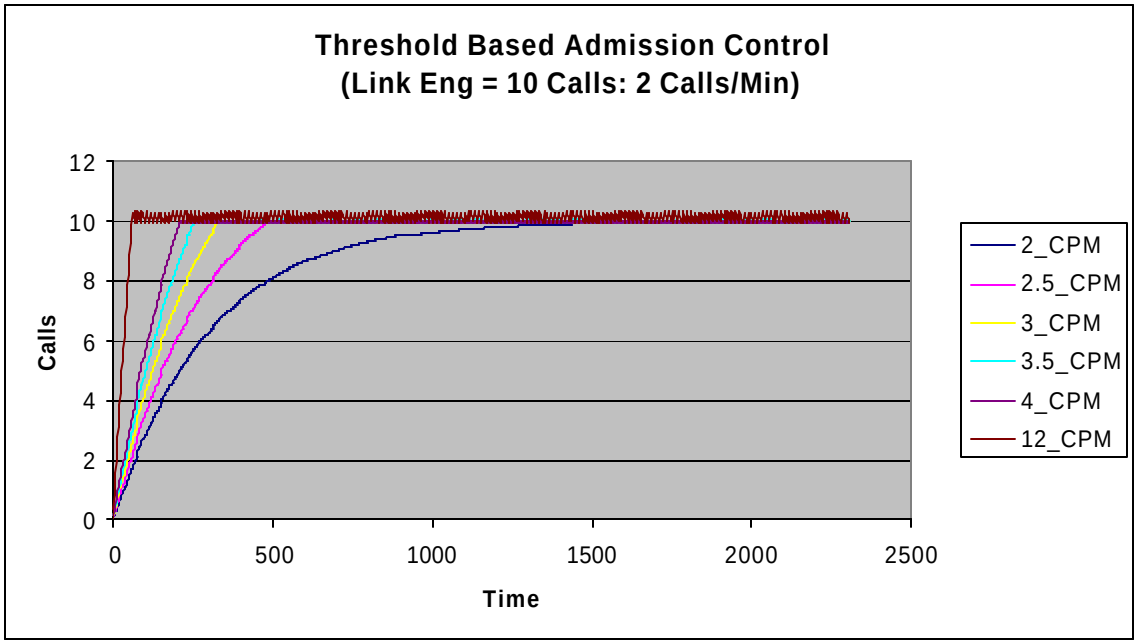
seconds (5 Minutes).

As the 2 CPM (Calls Per Minute) line illustrates, as we follow the time line from left to right, we start from no calls and the arrival rate exceeds the departure rate (2 CPM arrival rate vs 0 CPM departure rate) so the level rises. As the number of calls on the link increases, the departure rate starts to rise as well, until at 10 calls per minute, the arrival rate equals the departure rate and the system stabilizes at 10 calls.



Keeping the threshold for the rate proportional marking at 10 Calls per minute and looking at other arrival rates, as the arrival rate increases, the curve becomes steeper at the beginning but always eventually levels off where the departure rate equals the arrival rate. As can be seen from the graph, at an arrival rate of 4 Calls per minute (i.e. twice the arrival rate for which the link was engineered) the equilibrium point is about 13 calls.

The behavior of the threshold based admission control scheme is illustrated in the graph below. That behavior is: 100% admission when the number of calls is just below the threshold, and 100% denial when the number of calls just exceeds the threshold. Note that the curves show artifacts of the simplifications that we used in creating these charts in that they don't provide unit increases. Note also that no preemption process is at work here, only the natural departure rate of the system.



In comparing the two strategies, the general observation that can be made is that the rate proportional based admission control scheme does not achieve the objective of limiting admission at an engineered level of 10 calls. In comparison, the threshold based admission control scheme provides a hard engineered limit of 10 calls.

The performance of the rate proportional scheme is highly dependent on the actual call arrival rates. In a situation where a Service Level Agreement requires choosing thresholds that must be hard limits, this complicates the engineering process so that the system must either accept the possibility of oversubscription, which could result in degraded service for ALL Real-Time flows on the link (not just the offending set of flows), or of setting overly conservative thresholds that guarantee that bandwidth would be unused. Either way, the choice of threshold is not intuitively apparent so the risk

of mis-understanding between parties about the purpose of the threshold is very high.

For this reason we believe that threshold based marking is the correct approach for admission control of new real-time flows. It provides much faster indication to the new flow that the path is below or above the configured traffic level prior to admitting the flow into the network, and it provides an intuitive understanding of the performance of the network so that working with the system is simpler.

4. Analysis for Preemption

Preemption of flows on a network is a behavior intended to protect the network from unusual situations. An example would be the case where a network failure results in re-routing of traffic on the network, delivering more traffic to a particular link than was originally admitted. It also includes the case where a network operates with multiple precedence levels (e.g. commercial E911 service, Government Emergency Telecommunication Service (GETS), Defense Switch Network (DSN), etc.) and where higher precedence calls are more important than routine calls. In this case, the admission control threshold described previously would apply to routine calls only. High precedence calls would presumably be admitted to some higher traffic level (other congestion level). Normally, the percentage of these calls is quite low compared with routine calls on the network. There are, however, times when a disaster event might cause a very large number of higher precedence calls to be initiated. At those times, there needs to be a mechanism for the network to protect it self and shed load, presumably by preferentially preempting routine or lower precedence sessions to permit the higher precedence flows to be admitted.

As an aside to this discussion, it should be noted that the ability to preferentially shed load in a "panic" situation is an existing capability of the Public Switched Telephone Network (PSTN). The mechanisms at work however are based on the physical connections of wires within the PSTN "switch". Critical services are physically wired to the low numbered positions on the line cards. If the switch needs to start shedding load, these are the last to go, and the first to come back into service. Since there is no real protocol involved in this mechanism, it can't be directly ported to the VoIP world. However, the migration from TDM based to VoIP based telephony currently involves the loss of this capability.

Returning now to our analysis of rate proportional and threshold based preemption schemes, we note that there must, unavoidably, be some difference between a preemption threshold and an admission control threshold. The admission control process could exist for purposes of meeting the needs of a Service Level Agreement to allocate bandwidth between applications on a converged link. However, the preemption process is a second level process that protects the network from over submission. It is used to protect a converged network from having all of its bandwidth consumed by real-time flows that have DiffServ markings that give them preferential access to network resources. The preemption threshold provides a way of implementing a protection scheme that is targeted at removing some of the lower precedence traffic from the network during high congestion periods.

For rate proportional based preemption schemes, some of the same issues which are faced in a rate proportional based admission control scheme will face the preemption scheme. That is to say that, if not properly controlled, it could result in preemption of higher precedence traffic when lower precedence traffic still exists on the link. Rate proportional based schemes do, however, bring something else to the table so they cannot be dismissed completely.

4.1 Where Rate Proportional Marking May be Useful

For rate proportional marking, a flow that passes through a higher level of congestion would have a higher number of packets ECN marked whereas a flow of the same rate that passes through lower level of congestion would have fewer packets ECN marked. A flow that passes through several congestion points would also have a higher number of packets ECN marked versus a flow of the same rate that passes only through a single congestion point. A flow of higher rate would get more packets ECN marked than a flow of lower rate flowing through the same congestion point.

The implications of these observations are as follows. The absence of an ECN marking on a single packet does not indicate the absence of congestion in the network. However, where markings are viewed over a sufficiently long period of time, and assuming that we have controlled the implementation so that the percentage of marking is truly representative of the overall percentage of congestion, rate proportional marking provides additional information to end systems that is not available in a threshold based preemption scheme. However, time is required to arrive at a good estimate of the actual network behavior. Lower rate systems may need to wait longer before they could make a determination of whether an event was persistent vs transitory, or significantly above vs slightly above.

A rate proportional marking tells the end system not only that the threshold has been exceeded, but also gives the end system a way of estimating whether the flow is experiencing small or large congestion. It does not, however, provide an indication of the number of flows that represent the non-conformant traffic as the measuring point normally is not flow aware. Also should the packet flows encounter a second or third congestion point on the path, additional marking will be performed distorting the rate of congestion that is reported to the end system.

Once a flow has been admitted, it may be useful in some application of RT-ECN to report the percentage of a flow's packets that exceed the preemption threshold. Assuming that the implementation effectively controlled the probability of marking on a flow by flow basis, this marking could be used in the selection of which flow

should be preempted first. The flow with the highest percentage marking passes through a higher congestion point or it passes through several points of moderate congestion and thus makes a better candidate for preemption than other flows reporting a smaller percentage of marked packets. It may still not be selected for preemption because of rules governing the behavior of precedence vs routine traffic but the ability to make these decisions may improve the overall performance of the system.

4.2 Limitations of Rate Proportional Marking

Real-time flows may be variable rate or constant rate, and may have fixed or variable packet sizes. Variable rate traffic may consist of variable size packets with fixed emitted intervals, fixed size packets with variable emitted intervals or variable size packets with variable emission intervals. In IP networks different flows from different end systems, although constant rate, may use different fixed size packet (60 versus 200 byte) as well as different packet emit intervals, therefore different constant rate flows as well variable rate flows may be flowing through the congestion point. Normally, a router measures aggregated traffic and is not flow aware. Marking is performed on the traffic aggregate and not per flow. If the aggregate traffic rate is exceeded by "k" bits per second, then the expectation for a rate proportional marker is to mark packets at "k" bits per second on packet boundaries. However, the ECN marker does not know to which flow the packet that is being marked belongs, therefore flows will have there packets marked randomly.

Below is an example illustrating where rate proportional marking by itself would not identify the number of flows that are non-conformant:

> All traffic is sourced from endpoints that send 200 bytes every 20ms (constant rate 80kbps) or 50 packets per second. A single rate control (congestion) point is configured on a router to support 10 independent flows of 50 packets per second for a total rate of 800kbps or 500 packets per second. Now one additional flow of 50 packets per second (200 bytes every 20ms) is added into the path for a total of 11 flows which is equivalent to 880kbps or 550 packets per second. Below is an example illustrating where rate proportional marking by itself would not identify the number of flows that are non-conformant:
>
> Using rate proportional marking for the above simple case on average 50 out of 550 packets every second would be marked as non-conformant until the load is reduced through the rate control (congestion) point in the network. The marking of 50 packets every second would be randomized with no association to flows.

Packets belonging to more than one flow would be marked as non-conformant. Depending on the measurement time interval in the endpoints and the traffic characteristics, many and possibly all 11 endpoints will see some packets marked as flowing through a congested point in the network. However, each endpoint does not have enough information to determine the rate of congestion in the network. We can't use a simple policy such as "preempt at level X of marking" in the endpoint to make the preemption decisions.

Rate proportional marking in routers does not identify the number of flows that needs to be preempted nor does congestion marking of packets of a single flow as observed at the endpoint provide enough information to determine the level of congestion is experienced.. Other mechanisms in the preemption system need to be in place.

A second example illustrates a potential for unfairness in marking between flows:

A link carries traffic from 8 fixed rate voice flows with G.729 codec at 10 ms framing intervals and 3 fixed rate voice flows with G.711 codec at 20 ms framing interval. The first 8 voice flows have 50 byte packets at 100 times per second. The other 3 flows have 200 byte packets at 50 times per second. If we assume a token bucket style of metering, the point in the cycle that is most likely to detect the threshold first is on the packet with the largest size. In fact, if we are just barely above the preemption threshold, we would expect that the empty token bucket event would occur every time on the large packet instead of being distributed evenly among all flows.

This example illustrates the fact that control over marking rates on a flow by flow basis is not generally provided by typical processes for random marking. Great care must be exercised to ensure that these issues are overcome. The example here used two voice flows. The problem becomes much more severe with a combination of variable rate video flows, since the video flows on an instantaneous basis could be as much as 20 times the throughput of a single flow, even if they are only 5 times as much on an average basis. They also send multiple packets in quick succession which makes one of those packets far more likely to be marked than the other voice flows. A simple rate proportional marking scheme discriminates by more aggressively marking flows with large packets or flows that are more bursty.

Example 3 exposes limitation of using rate proportional marking to determining how many or which flows to preempted:

Now let's consider a different scenario, in this scenario all flows monitored are transmitted from the same source. The receiving network edge device computes the non-conformant rate and signals the rate to the transmitter. The simple response is for the transmitter to reduce its rate by the signaled amount, terminating one or more flows. However, there are situations where the computed congestion rate by the receiving end system is not accurate.

The first situation is where multiple congestions points exist along the path. Both congestion points marking packets to indicate their level of congestion, resulting in packet flows being marked twice as well with some remarking by the second congestion point of packets marked by the fist congestion point. In that case, the measured congestion in the receiving end system as a representation rate proportional marking is not accurate. This could lead to more flows being preempted than may be necessary.

The second situation arises when flows take are monitored at the receiving end take different paths through the network and were levels of congestion may not be the same. The summing of marked packets at the monitoring end system does not provide what the congestion level is along a specific path. Rate proportional marking its self does not provide how many flows need to be preempted from each path.

Example 4 discusses rate proportional marking and different flow precedence level:

Another issue arises when there is traffic from two of more precedence levels in the network and where there are many higher precedence flows and only a few or one low precedence flow going through a congestion point the problem becomes more acute. The router that is performing rate proportional marking is not flow or precedence aware and, if using a simple token bucket approach and will mark packets as token are used up. Since there is a larger number of high precedence vs low precedence packets flowing through the congestion point, the probability is high that the higher precedence packets will be marked and the one low precedence flow will not be marked. Without a significantly longer latency and more complex decision making, this would result in preemption of the higher precedence flow even though there was a low precedence flow on the path that should be the flow selected for preemption. US DoD's DSN or DRSN networks are examples were 5 or 6 levels of flow precedence is used with the requirement that the lowest precedence flow(s) that pass through a congestion point(s) versus preempting a flow(s) that pass through the highest

congestion point or one that pass through several congestion points.

Example 5 discusses control over the randomness and farness of marking:

> Finally the issue of control over the randomness is an important issue in determining whether those benefits could be achieved. For example, the traditional rate base policing algorithms based on token buckets result in almost guaranteed marking of the last few packets of a "bunch" of packets that arrive very close together. Codecs that emit packets on a fixed interval have a high likelihood of creating the scenario where packets from two flows arrive in the same sequence and at about the same relative timing from frame to frame. Since nothing prevents the last packets in the sequence from having a higher precedence, it is highly likely that a lower precedence flow that falls at the beginning of that bunch will remain unmarked from frame to frame while the precedence flow at the tail of that bunch would always be marked.

5. Security Considerations

This document doesn't propose any new mechanisms for the Internet protocol, and therefore doesn't introduce any new security considerations.

6. Acknowledgements

The authors acknowledge a great many inputs, most notably from David Black, Attila Bader and Georgios Karagiannis.

7. Normative References

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

[RT-ECN] Babiarz, J., Chan, K., and V. Firoiu, "Congestion Notification Process for Real-Time Traffic", draft-babiarz-tsvwg-rtecn-03 (RT-ECN), March 2005.

Authors' Addresses

Jozef Babiarz
Nortel Networks
3500 Carling Avenue
Ottawa, Ont. K2H 8E9
Canada

Phone: 613-763-6098
Fax: 613-768-2231
Email: babiarz@nortel.com

Stephen Dudley
Nortel Networks
4001 E. Chapel Hill Nelson Highway
Research Triangle Park, NC 27709
U.S.A.

Email: SMDudley@nortel.com

Kwok Ho Chan
Nortel Networks
600 Technology Park Drive
Billerica, MA 01821
US

Phone: 978-288-8175
Fax: 978-288-4690
Email: khchan@nortel.com

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.