PCN
Internet-Draft
Intended status: Informational
Expires: April 25, 2007

K. Chan
Nortel Networks
A. Charny
Cisco Systems
P. Eardley
BT Research

October 22, 2006

Pre-Congestion Notification Problem Statement
draft-chan-pcn-problem-statement-01

Abstract

DiffServ mechanisms have been developed to support Quality of Service (QoS). However, the level of assurance that can be provided with DiffServ without substantial over-provisioning is limited. Pre-Congestion Notification (PCN) investigates the use of per-flow admission control to provide the required service guarantees for the

admitted traffic. While admission control will protect the QoS under normal operating conditions, an additional flow pre-emption mechanism is necessary in the times of heavy congestion (e.g. caused by route changes due to link or node failure).

This document provides a problem statement on the addition of flow admission control and flow pre-emption functionality to a DiffServ network, in particular for the support of real time services such as voice and video.

Table of Contents

1. Introduction

1.1. Motivation

IP networks were initially designed to perform per IP packet forwarding treatment without discrimination. With the increased use of the IP network by applications with different transport functional requirement, the notion of Quality of Service (QoS) was introduced [19].

DiffServ [10] introduced differentiated per packet forwarding treatment to provide QoS: some packets are served at a higher scheduling priority than others. Diffserv Service Classes [18] categorises various DiffServ traffic and recommends how they can be used for packets from applications with different transport requirements. For instance there are Telephony and Real-time Interactive service classes. Applications like these need low loss, low delay and low jitter. A suitable Per Hop Behavior (PHB) is Expedited Forwarding (EF) [16], which works by assuring that packets (usually) encounter very short or empty queues. Each router is allocated a certain amount of bandwidth for the EF PHB, for instance. Excess packets are dropped and delayed, thus leading to poorer QoS for an end user running an application like voice-over-IP. Even if average traffic levels are known, due to traffic variations the level of assurance that can be provided with DiffServ without substantial over-provisioning is limited.

To help ensure that the average traffic loads remain within the allocated bandwidth limits, the DiffServ Architecture [10] introduces the idea of policing the amount of traffic in a class as it enters the network. The acceptable traffic level is described by a traffic conditioning agreement (TCA). However, in practice, TCAs police the aggregate traffic in a class at the network ingress, and for scalability reasons typically includes traffic to different destinations. As a result, TCA's do not guarantee that EF aggregate at any given node in the network does not exceed the allocated capacity [21], and so don't ensure that a particular end user's QoS is guaranteed. Also, in practice TCAs are static and so require accurate and/or conservative prediction of the traffic matrix.

To cope with the issue of exceeding bandwidth allocation to EF on some links, in practice a policer or shaper is assumed to be installed at the interior nodes as well. However, shaping or policing traffic causes excess packets be dropped and delayed, thus leading to poorer QoS for an end user running an application like voice-over-IP. Even if average traffic levels remain within the allocated bandwidth limits, traffic variations may limit the level of assurance that can be provided with DiffServ without substantial

over-provisioning.

These factors motivate us to work on per flow admission control for a DiffServ network, and in particular on measurement-based admission control, ie new flow requests are blocked dynamically in response to actual (incipient) congestion on a router within the DiffServ network.

However, despite flow admission control, sometimes there can be heavy congestion - for example caused by link or node failure that effectively reduces the network's capacity. The default option is that the QoS of all flows is degraded. However, by pre-empting some flows the QoS of the remaining flows can be protected. The work reported in [7] indicates that in the context where calls have different recongizable precedence levels (e.g. in the context of military/emergency calls [20]), this problem can be partially addressed by dropping lower-precedence calls preferentially while protecting higher precedence calls. However, as it was shown in [6], the need to pre-empt some flows of a given precedence level, while protecting the QoS of the rest of the flows of this precedence level remains.

This motivates us to work on per flow pre-emption for a DiffServ network, and in particular on measurement-based pre-emption, ie existing flows are dropped dynamically in response to actual congestion on a router within the DiffServ network.

Explicit Congestion Notification (ECN) [15] introduced the idea of a router indicating that it is congested by changing the header of packets ("marking" them). However, ECN in RFC3168 [15] is designed for TCP applications. This motivates us to develop the concept for real-time applications. A router "PCN-marks" packets as an early warning of its incipient congestion ("pre-congestion"). These markings are then used by the admission control and pre-emption mechanisms.

The rest of this document discusses our proposed goals, assumptions and some functional architecture directions.

1.2. Goals

From the functional standpoint, the goal of the proposed PCN approach is twofold:

- Flow Admission Control: block admission of new flows as soon as signs of incipient congestion are detected, to prevent congestion / overload.

- Flow Pre-emption: If traffic exceeds the desired/allocated capacity (e.g. due to a failure), pre-empt sufficient flows so that the QoS of the remaining flows is protected.

The following are proposed as design goals:

- The PCN-enabled packet forwarding network should be simple, scalable and robust

- Compatibility with other traffic (i.e. a proposed solution should work well when non-PCN traffic is also present in the network)

- Support of different types of real-time traffic (eg should work well with CBR and VBR voice and video sources)

- Reaction time of the mechanisms should be commensurate with the desired application-level requirements (e.g. a pre-emption mechanism needs to pre-empt flows before significant QoS issues are experienced by all real-time traffic, and before a user hangs up)

- Compatibility with different precedence levels of real-time applications (e.g. preferential treatment of higher precedence calls over lower precedence calls, MLPP [20].

2. Architecture and Deployment Scenarios

The above goals point to a high-level approach where functionality is split between:

- Nodes in the PCN-enabled network, which monitor their own state of (pre) congestion and mark packets if appropriate

- Nodes at the edge of the PCN-enabled network, which control admission of new flows and pre-emption of existing flows, based on information from nodes in the network. This information is in the form of the marked packets and not explicit signalling messages.

The aim of this split is to keep the bulk of the network simple, scalable and robust, whilst confining policy, application-level and security interactions to the edge of the PCN network.

Section 2.1 provides a high-level description of the functional architecture, and Section 2.2 considers some possible deployment scenarios.

2.1. Functional Architecture

Figure 1 shows a schematic diagram of the high-level functional architecture:

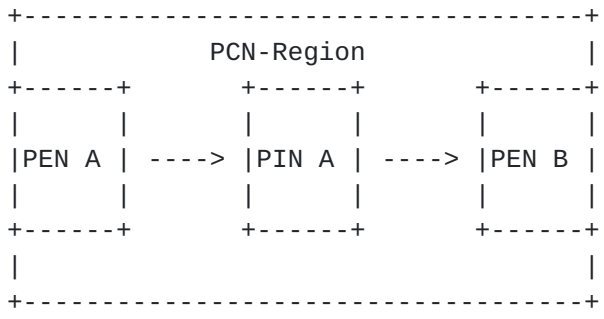

Figure 1: PCN-based Functional Architecture

The terms are defined as follows:

PCN-Region:

A DiffServ region of the Internet running PCN, that is the PCN-based mechanisms are used to decide whether to admit a new flow to the DiffServ region and whether to pre-empt an existing flow. All traffic enters/leaves the PCN-Region through a PCN End Node. Please note that the PCN-Region is also defined by the Diffserv Service Class [18] that is subject to the PCN mechanisms.

PCN Interior Node (PIN) (function):

The PCN Interior Node is an "on-path" function. It performs traffic metering and PCN-marking: the function that enables a network element to give an early warning of its own incipient congestion ("pre-congestion") on one of its interfaces, ie traffic is above a certain level, by marking, e.g. changing the header of packet(s).

PCN End Node (PEN) (function):

The PCN End Node is an "on-path" function. The PCN End Node is where the PCN Region ends. It indicates the significance of the PCN packet marking which terminates at this functional node. This functional description does not imply which physical device will implement this function (e.g., edge router, media gateway or end-host). This "on-path" function performs the detection of PCN-marks: the function that monitors

PCN-marking to obtain on-path congestion information as signaled through PCN-marking by PCN-enabled Interior Nodes. For PCN's purpose, these actions may include but not be limited to:

+ Make Flow Admission Control and/or Flow Pre-emption decisions.

+ Signalling the PCN information to others for making the Flow Admission Control and/or Flow Pre-emption decisions.

+ Perform measurement of marked packets across multiple IP packets of a flow to derive network information for a flow that a single packet can not provide.

+ Perform measurement of marked packets across multiple IP flows to derive additional network information.

2.2. Notion of Trust

With the above Functional Architecture, there exists a notion of trust between the different functional elements:

o PENs trusting PINs to provide the correct network information.

o PINs trusting PENs that admission control and pre-emption will be administered correctly so the PINs will not be over-whelmed with traffic.

o Users/Applications trusting the PENs that they will provide dependable informations for taking application actions.

o PENs trusting other PENs that when one PEN performs its task of flow admission control, other PENs will also perform their flow admission control actions. So that the "good citizens" does not get penelized.

We discuss some notion of trust further in section 4.1 Assumption 1: Controlled Environment and in section 6 Security Implications.

2.3. Deployment Scenarios

The previous section describes the functional architecture. The association of these functions to physical devices may depend on the deployment scenario. We make some general comments about the physical devices where the functions above will typically reside:

- The PCN Interior Node function typically resides in a network element like a router or a switch where packet forwarding is handled.

- The PCN End Node function typically resides in a router, but may also be on a host or a proxy. It is typically the closest PCN-enabled device to the user.

Operators of networks will want to use the PCN functions (and standards) in various arrangements, for instance depending on how they are performing admission control outside the PCN-region, their goals beyond those in Section 1.2, and assumptions in addition to those in Section 4.

Hence we shall work on several deployment scenarios. Initially we have the following possibilities in mind:

- IntServ over DiffServ [14], the DiffServ region is PCN-enabled. This is described in CL Architecture [2].

- SIP-controlled Admission and Pre-emption: trusted SIP endpoints (gateway or host) perform application flow admission and flow pre-emption, based on network information provided by PCN marking. This is described in SIP Controlled Admission and Preemption [4].

- Pseudowire: PCN may be used as a congestion avoidance mechanism for end-user deployed pseudowires (collaborate with the PWE3 WG in investigation of this possibility).

3. Standards

To solve the PCN functionality described above, we will work on developing a standard for each of the following problems:

- How should the measurement of pre-congestion be done at the PINs? For determining when an interior node should mark a packet in order to give early warning of its own congestion? Should there be a standardized algorithm? Or just the required behavior should be standardized?

- How should such a mark be encoded by the PINs in a packet (in the ECN and/or DSCP fields)?

- How should these markings (at packet granularity) be interpreted by the PENs for making flow admission control and flow pre-emption decisions (at flow granularity)?

Initial work addressing these questions has been reported to the IETF in CL Architecture [2], RT ECN [1], NSIS RMD [5]. Note that other options are possible.

One of the key questions that need to be answered in the context of standardisation is, what level of detail of standardisation is appropriate for the first bullet? For example, should PCN be specified as an algorithm relating the probability of PCN-marking a packet to (some specific description of the) traffic level? Or something more detailed (e.g. implementation specifics) or less detailed (describe the behaviour in more general terms than an algorithm). We want flexibility, but also want to be sure that different standards-compliant implementations will work together.

A similar issue arises for the third bullet. Additionally, it might be possible to specify more than one way of reacting to the PCN-markings. On the plus side, different reaction behaviours may be more suited to different deployment scenarios. But this could require coordination of the PCN End Nodes for a particular PCN-region, so they agreed to use the same reaction behaviour.

On the second bullet, CL PHB [3] has some options for how to do the encoding, focussed on use of the ECN field, and an initial analysis of their pros and cons. Another possibility is to use the DSCP field, as in NSIS RMD [5], or a combination of the two. The WG will study the trade-offs between different encoding options.

4. Assumptions and Constraints on Problem Scope

In order to make rapid progress, initially we will restrict the problem space in several ways. NOTE: Subsequent re-chartering may investigate solutions for when some of these restrictions are not in place. The working assumption is that the standards developed in the initial phase should not need to be modified to satisfy the solutions for when these restrictions are removed.

4.1. Assumption 1: Controlled Environment

We assume that the PCN-enabled Internet Region is a controlled environment, i.e. all the interior and end nodes of the region run PCN and trust each other.

There are several reasons for proposing this assumption:

- The PCN-Region has to be fully encircled by a ring of PCN End Nodes, otherwise packets could enter the PCN-Region without being subject to admission control, which would potentially destroy the

QoS of existing flows.

- Similarly, a PCN End Node has to trust that all the interior routers are doing PCN-marking. A non-PCN router won't be able to alert that it's suffering pre-congestion, which potentially would lead to too many calls being admitted (or too few being pre-empted). Worse, a rogue router could perform attacks such as marking all packets so that no flows were admitted.

One way of assuring the above two points is that the entire PCN-region is run by a single operator. Another possibility is that there are several operators but they trust each other to a sufficient level. Please note that this restriction applies to packets in the traffic class that is subject to the PCN mechanisms.

4.2. Assumption 2: Many Flows and Additional Load

We assume that there are many flows on any bottleneck link in the PCN-enabled region.

Measurement-based admission control assumes that the past is a reasonable reflection of the future: the network conditions are measured at the time of a new flow request, however the actual network performance must be OK during the call some time later.

One issue is that if there are only a few variable rate flows, then the aggregate traffic level may vary a lot, perhaps enough to cause some packets to get dropped. If there are many flows then the aggregate traffic level should be statistically smoothed. How many flows is enough depends on a number of things such as the variation in each flow's rate, the total PCN bandwidth, and the size of the "safety margin" between the traffic level at which we start PCN-marking and at which packets are dropped.

4.3. Assumption 3: Real-Time Applications

We assume that packets come from real time applications generating inelastic traffic like voice and video requiring low delay, jitter and packet loss, i.e. as defined by the Controlled Load Service [9].

This assumption is to help focus the effort where it looks like PCN would be most useful, ie the sorts of applications where per flow QoS is a known requirement. For instance, the impact of this assumption would be to guide simulations work. NOTE: PCN should be readily extendible to other applications like ones that typically use Assured Forwarding [12].

5. Open Design Issues

Whilst working on the general issues of flow admission control and flow pre-emption, we have found several issues that proved hard to solve. They are briefly documented here - further details are in [2]. In general they seem to be characteristics of most measurement-based admission control schemes, but some may not be relevant to particular deployment scenarios. From the perspective of this problem statement, besides just noting the issue the PCN WG could:

- Upgrade the issue, so it's added to the "Goals" section earlier, or to the "Assumptions" section as appropriate

- Downgrade the issue, either because it isn't that important or because it's better dealt with outside the PCN solution

- Wait and see, ie as the PCN solution is developed assess how much extra complexity solving the issue would add

The comments below are about admission control, but generally a similar issue arises for flow pre-emption.

ECMP (Equal Cost Multi-Path) Routing:

> In order to decide whether to admit a new flow, the CL Architecture [2] scheme determines what the ingress and egress PENs would be and measures the current level of PCN-marking between them (Congestion-Level-Estimate). If routers in the PCN-region run ECMP, then traffic between a particular pair of PENs may follow several different paths. The problem is that if just one of the paths is congested such that packets are being PCN-marked, then the Congestion-Level-Estimate measured by the egress PEN will be diluted by unmarked packets from other non-congested paths.

Bi-Directional Sessions:

> CL Architecture [2] describes a flow admission control mechanism. However, from the application perspective, for a bi-directional session the two flows should be admitted as a pair - for instance a bi-directional voice call only makes sense if flows in both directions are admitted.

Global Coordination:

> CL Architecture [2] makes its admission decision based on PCN-markings between a particular pair of PENs. Decisions about flows through a different pair of PENs are made independently.

However, one can imagine network topologies and traffic matrices where from a global perspective it would be better to make a coordinated decision across all the pairs of PENs for the whole PCN-region. For example, to block (or even pre-empt) flows on one PEN pair so that more important flows through a different pair could be admitted.

Aggregate Traffic Characteristics:

Even when the number of flows is stable, the traffic level through the PCN-region will vary because the sources vary their traffic rates. The CL Architecture [2] mechanism works best when there's "some" variability in the total traffic level at a router's interface (ie in the aggregate traffic from all sources). Too much variation means that a router may (at one moment) not be doing any PCN-marking and then (at another moment) be overloaded, ie drop packets. This makes it hard to tune the admission control scheme to stop admitting new flows at the right time. However, too little variation can also be a problem. For example, if all the sources are constant bit rate and are synchronised, then the total traffic level at a router's interface could be (almost) at its capacity and all packets could still be serviced instantly. However, admitting one more flow could tip the router over its capacity, so its queue grew indefinitely until it had to drop packets. "Some" traffic variation means that as the traffic level nears the capacity limit, some packets are PCN-marked but there's still enough capacity to cope with the traffic fluctuations. Hence new flows can be blocked and packets are never dropped.

Speed of Reaction:

The CL Architecture [2] mechanism has a limited speed of reaction: if a big burst of admission requests occurs in a very short space of time (eg prompted by a televote), they could all get admitted before enough PCN-marks are seen to block new flows. In other words, any additional load offered within the reaction time of the mechanism mustn't move the CL-Region directly from no congestion to overload.

6. Security Implications

Packets from normal precedence and higher precedence sessions [20] aren't distinguishable by PCN Interior Nodes. This prevents an attacker specifically targeting, in the data plane, higher precedence packets (perhaps for DoS or for eavesdropping). However, PCN End Nodes can access this information to help decide whether to admit or

pre-empt a flow. The separation of network information provided by the Interior Nodes and the precedence information at the PCN End Nodes allows simpler, easier and better focused security enforcement.

PCN End Nodes police packets to ensure a flow sticks within its agreed limit. This is similar to the existing IntServ behaviour. Between them the PCN End Nodes must fully encircle the PCN-Region, otherwise packets could enter the PCN-Region without being subject to admission control, which would potentially destroy the QoS of existing flows.

It is assumed that all the Interior Nodes and PCN End Nodes run PCN and trust each other (ie the PCN-enabled Internet Region is a controlled environment). For instance a non-PCN router wouldn't be able to alert that it's suffering pre-congestion, which potentially would lead to too many calls being admitted (or too few being pre-empted). Worse, a rogue router could perform attacks such as marking all packets so that no flows were admitted.

So security requirements are focussed at specific parts of the PCN-Region:

The PCN End Nodes become the trust points. The degree of trust required depends on the kinds of decisions it has to make and the kinds of information it needs to make them. For example when the PCN End Node needs to know the contents of the sessions for making the decisions, when the contents are highly classified, the security requirements for the PCN End Nodes involved will also need to be high.

PCN-marking by the Interior Nodes along the packet forwarding path needs to be trusted, because the PCN End Nodes rely on this information.

7. IANA Considerations

To be completed.

8. Acknowledgements

To be completed.

9. Informative References

[1] Babiarz, J., "Congestion Notification Process for Real-Time

Traffic", draft-babiarz-tsvwg-rtecn-05 (work in progress), October 2005.

[2] 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.

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

[4] Babiarz, J., "SIP Controlled Admission and Preemption", draft-babiarz-pcn-sip-cap-00 (work in progress), October 2006.

[5] Bader, A., "RMD-QOSM - The Resource Management in Diffserv QOS Model", draft-ietf-nsis-rmd-08 (work in progress), October 2006.

[6] Baker, F. and J. Polk, "MLEF Without Capacity Admission Does Not Satisfy MLPP Requirements", draft-ietf-tsvwg-mlef-concerns-00 (work in progress), February 2005.

[7] Silverman, S., "Multi-Level Expedited Forwarding Per Hop Behavior (MLEF PHB)", draft-silverman-tsvwg-mlefphb-03 (work in progress), October 2005.

[8] Braden, B., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[9] Wroclawski, J., "Specification of the Controlled-Load Network Element Service", RFC 2211, September 1997.

[10] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

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

[12] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[13] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.

[14] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J., and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, November 2000.

[15] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[16] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[17] Charny, A., Bennet, J., Benson, K., Boudec, J., Chiu, A., Courtney, W., Davari, S., Firoiu, V., Kalmanek, C., and K. Ramakrishnan, "Supplemental Information for the New Definition of the EF PHB (Expedited Forwarding Per-Hop Behavior)", RFC 3247, March 2002.

[18] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

[19] "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanisms", Proceedings of SIGCOMM '92 at Baltimore MD, August 1992.

[20] "Multilevel Precedence and Pre-emption Service (MLPP)", ITU-T Recommendation I.255.3, 1990.

[21] "Economics and Scalability of QoS Solutions", BT Technology Journal Vol 23 No 2, April 2005.

[22] Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.

Authors' Addresses

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

Email: khchan@nortel.com

Anna Charny
Cisco Systems
14164 Massachusetts Ave
Boxborough, MA 01719
USA

Email: acharny@cisco.com

Philip Eardley
BT Research
B54/77, Sirius House Adastral Park Martlesham Heath
Ipswich, Suffolk IP5 3RE
United Kingdom

Email: philip.eardley@bt.com

Full Copyright Statement

Intellectual Property

Acknowledgment

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