Internet Engineering Task Force (IETF)
Request for Comments: 8084
BCP: 208
Category: Best Current Practice
ISSN: 2070-1721

G. Fairhurst
University of Aberdeen

March 2017

Network Transport Circuit Breakers

Abstract

This document explains what is meant by the term "network transport Circuit Breaker". It describes the need for Circuit Breakers (CBs) for network tunnels and applications when using non-congestion-controlled traffic and explains where CBs are, and are not, needed. It also defines requirements for building a CB and the expected outcomes of using a CB within the Internet.

Status of This Memo

This memo documents an Internet Best Current Practice.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on BCPs is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc8084.

Table of Contents

1. Introduction

The term "Circuit Breaker" originates in electricity supply, and has nothing to do with network circuits or virtual circuits. In electricity supply, a Circuit Breaker (CB) is intended as a protection mechanism of last resort. Under normal circumstances, a CB ought not to be triggered; it is designed to protect the supply network and attached equipment when there is overload. People do not expect an electrical CB (or fuse) in their home to be triggered, except when there is a wiring fault or a problem with an electrical appliance.

In networking, the CB principle can be used as a protection mechanism of last resort to avoid persistent excessive congestion impacting other flows that share network capacity. Persistent congestion was a feature of the early Internet of the 1980s. This resulted in excess traffic starving other connections from access to the Internet. It

was countered by the requirement to use congestion control (CC) in the Transmission Control Protocol (TCP) [Jacobson88]. These mechanisms operate in Internet hosts to cause TCP connections to "back off" during congestion. The addition of a congestion control to TCP (currently documented in [RFC5681]) ensured the stability of the Internet, because it was able to detect congestion and promptly react. This was effective in an Internet where most TCP flows were long lived (ensuring that they could detect and respond to congestion before the flows terminated). Although TCP was, by far, the dominant traffic, this is no longer the always the case, and non-congestion-controlled traffic, including many applications using the User Datagram Protocol (UDP), can form a significant proportion of the total traffic traversing a link. To avoid persistent excessive congestion, the current Internet therefore requires consideration of the way that non-congestion-controlled traffic is forwarded.

A network transport CB is an automatic mechanism that is used to continuously monitor a flow or aggregate set of flows. The mechanism seeks to detect when the flow(s) experience persistent excessive congestion. When this is detected, a CB terminates (or significantly reduces the rate of) the flow(s). This is a safety measure to prevent starvation of network resources denying other flows from access to the Internet. Such measures are essential for an Internet that is heterogeneous and for traffic that is hard to predict in advance. Avoiding persistent excessive congestion is important to reduce the potential for "Congestion Collapse" [RFC2914].

There are important differences between a transport CB and a congestion control method. Congestion control (as implemented in TCP, Stream Control Transmission Protocol (SCTP), and Datagram Congestion Control Protocol (DCCP)) operates on a timescale on the order of a packet Round-Trip Time (RTT): the time from sender to destination and return. Congestion at a network link can also be detected using Explicit Congestion Notification (ECN) [RFC3168], which allows the network to signal congestion by marking ECN-capable packets with a Congestion Experienced (CE) mark. Both loss and reception of CE-marked packets are treated as congestion events. Congestion control methods are able to react to a congestion event by continuously adapting to reduce their transmission rate. The goal is usually to limit the transmission rate to a maximum rate that reflects a fair use of the available capacity across a network path. These methods typically operate on individual traffic flows (e.g., a 5-tuple that includes the IP addresses, protocol, and ports).

In contrast, CBs are recommended for non-congestion-controlled Internet flows and for traffic aggregates, e.g., traffic sent using a network tunnel. They operate on timescales much longer than the packet RTT, and trigger under situations of abnormal (excessive)

congestion. People have been implementing what this document characterizes as CBs on an ad hoc basis to protect Internet traffic. This document therefore provides guidance on how to deploy and use these mechanisms. Later sections provide examples of cases where CBs may or may not be desirable.

A CB needs to measure (meter) some portion of the traffic to determine if the network is experiencing congestion and needs to be designed to trigger robustly when there is persistent excessive congestion.

A CB trigger will often utilize a series of successive sample measurements metered at an ingress point and an egress point (either of which could be a transport endpoint). The trigger needs to operate on a timescale much longer than the path RTT (e.g., seconds to possibly many tens of seconds). This longer period is needed to provide sufficient time for transport congestion control or applications to adjust their rate following congestion, and for the network load to stabilize after any adjustment. Congestion events can be common when a congestion-controlled transport is used over a network link operating near capacity. Each event results in reduction in the rate of the transport flow experiencing congestion. The longer period seeks to ensure that a CB is not accidentally triggered following a single (or even successive) congestion event(s).

Once triggered, the CB needs to provide a control function (called the "reaction"). This removes traffic from the network, either by disabling the flow or by significantly reducing the level of traffic. This reaction provides the required protection to prevent persistent excessive congestion being experienced by other flows that share the congested part of the network path.

Section 4 defines requirements for building a CB.

The operational conditions that cause a CB to trigger ought to be regarded as abnormal. Examples of situations that could trigger a CB include:

- anomalous traffic that exceeds the provisioned capacity (or whose traffic characteristics exceed the threshold configured for the CB);
- traffic generated by an application at a time when the provisioned network capacity is being utilized for other purposes;
- routing changes that cause additional traffic to start using the path monitored by the CB;

- misconfiguration of a service/network device where the capacity available is insufficient to support the current traffic aggregate;

- misconfiguration of an admission controller or traffic policer that allows more traffic than expected across the path monitored by the CB.

Other mechanisms could also be available to network operators to detect excessive congestion (e.g., an observation of excessive utilization for a port on a network device). Utilizing such information, operational mechanisms could react to reduce network load over a shorter timescale than those of a network transport CB. The role of the CB over such paths remains as a method of last resort. Because it acts over a longer timescale, the CB ought to be triggered only when other reactions did not succeed in reducing persistent excessive congestion.

In many cases, the reason for triggering a CB will not be evident to the source of the traffic (user, application, endpoint, etc.). A CB can be used to limit traffic from applications that are unable, or choose not, to use congestion control or in cases in which the congestion control properties of the traffic cannot be relied upon (e.g., traffic carried over a network tunnel). In such circumstances, it is all but impossible for the CB to signal back to the impacted applications. In some cases, applications could therefore have difficulty in determining that a CB has been triggered and where in the network this happened.

Application developers are therefore advised, where possible, to deploy appropriate congestion control mechanisms. An application that uses congestion control will be aware of congestion events in the network. This allows it to regulate the network load under congestion, and it is expected to avoid triggering a network CB. For applications that can generate elastic traffic, this will often be a preferred solution.

1.1. Types of CBs

There are various forms of network transport CBs. These are differentiated mainly on the timescale over which they are triggered, but also in the intended protection they offer:

- Fast-Trip CBs: The relatively short timescale used by this form of CB is intended to provide protection for network traffic from a single flow or related group of flows.

- Slow-Trip CBs: This CB utilizes a longer timescale and is designed to protect network traffic from congestion by traffic aggregates.

- Managed CBs: Utilize the operations and management functions that might be present in a managed service to implement a CB.

Examples of each type of CB are provided in Section 4.

2. Terminology

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

3. Design of a CB (What makes a good CB?)

Although CBs have been talked about in the IETF for many years, there has not yet been guidance on the cases where CBs are needed or upon the design of CB mechanisms. This document seeks to offer advice on these two topics.

CBs are RECOMMENDED for IETF protocols and tunnels that carry non-congestion-controlled Internet flows and for traffic aggregates. This includes traffic sent using a network tunnel. Designers of other protocols and tunnel encapsulations also ought to consider the use of these techniques as a last resort to protect traffic that shares the network path being used.

This document defines the requirements for the design of a CB and provides examples of how a CB can be constructed. The specifications of individual protocols and tunnel encapsulations need to detail the protocol mechanisms needed to implement a CB.

Section 3.1 describes the functional components of a CB and Section 3.2 defines requirements for implementing a CB.

3.1. Functional Components

The basic design of a CB involves communication between an ingress point (a sender) and an egress point (a receiver) of a network flow or set of flows. A simple picture of operation is provided in Figure 1. This shows a set of routers (each labeled R) connecting a set of endpoints.

A CB is used to control traffic passing through a subset of these routers, acting between the ingress and a egress point network devices. The path between the ingress and egress could be provided by a tunnel or other network-layer technique. One expected use would

be at the ingress and egress of a service, where all traffic being considered terminates beyond the egress point; hence, the ingress and egress carry the same set of flows.

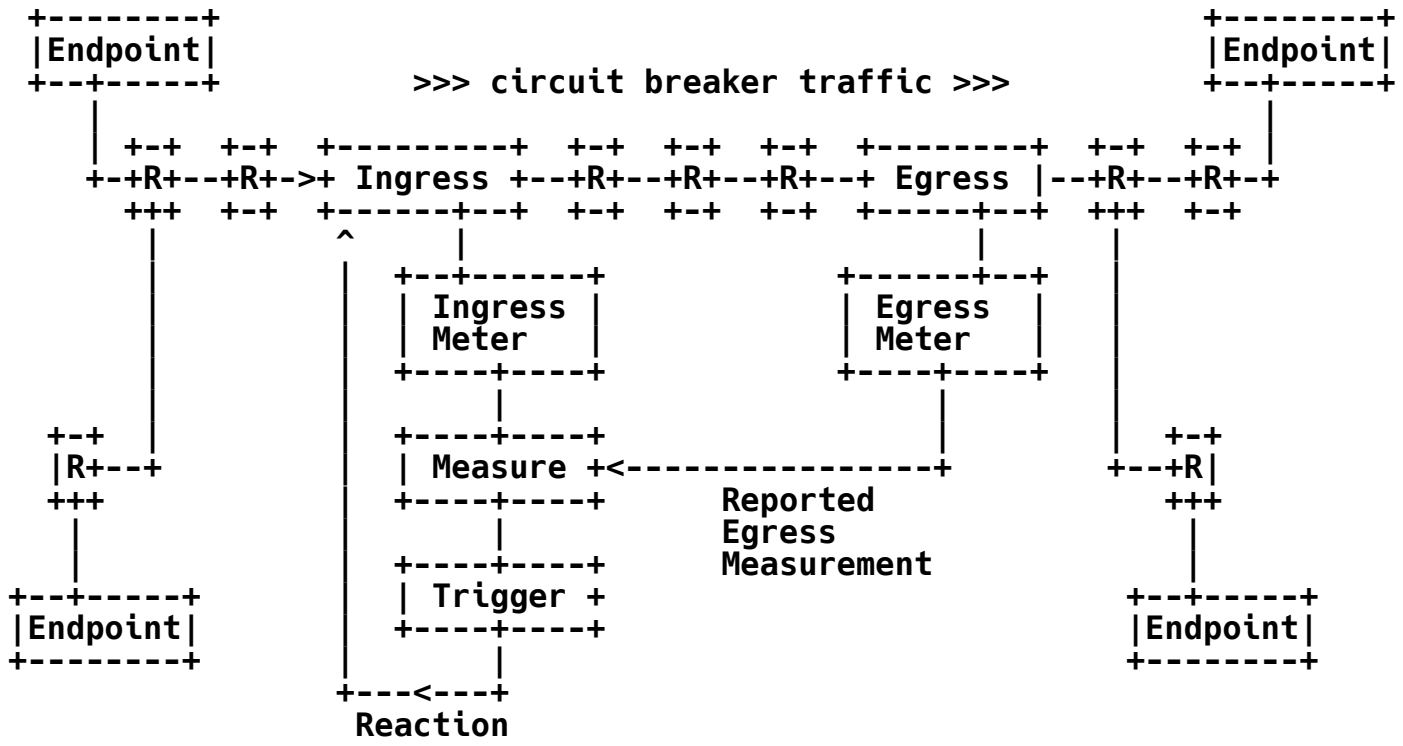

Figure 1: A CB controlling the part of the end-to-end path between an ingress point and an egress point. Note in some cases, the trigger and measurement functions could alternatively be located at other locations (e.g., at a network operations center).

In the context of a CB, the ingress and egress functions could be implemented in different places. For example, they could be located in network devices at a tunnel ingress and at the tunnel egress. In some cases, they could be located at one or both network endpoints (see Figure 2), implemented as components within a transport protocol.

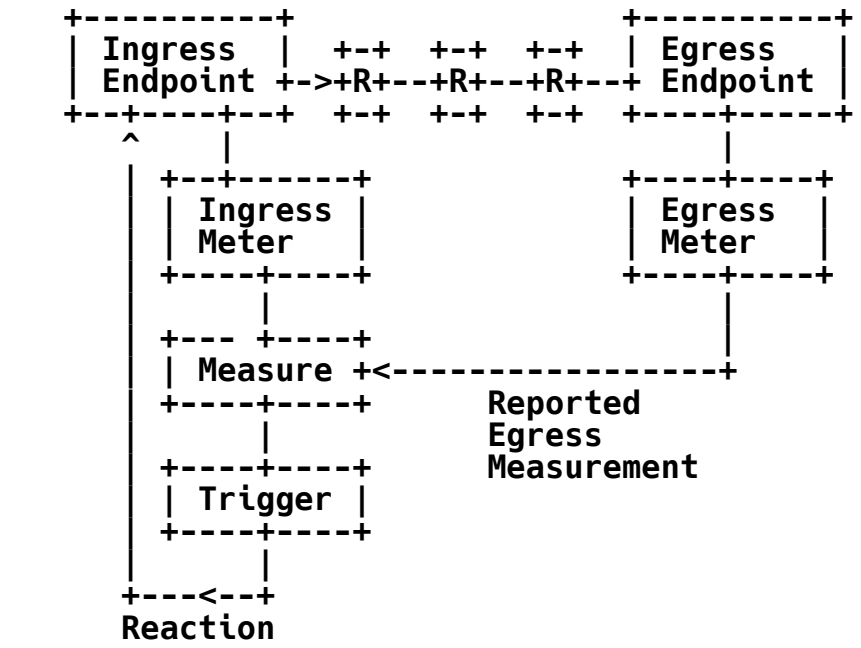

Figure 2: An endpoint CB implemented at the sender (ingress) and receiver (egress).

The set of components needed to implement a CB are:

1. An ingress meter (at the sender or tunnel ingress) that records the number of packets/bytes sent in each measurement interval. This measures the offered network load for a flow or set of flows. For example, the measurement interval could be many seconds (or every few tens of seconds or a series of successive shorter measurements that are combined by the CB Measurement function).

2. An egress meter (at the receiver or tunnel egress) that records the number/bytes received in each measurement interval. This measures the supported load for the flow or set of flows, and it could utilize other signals to detect the effect of congestion (e.g., loss/congestion marking [RFC3168] experienced over the path). The measurements at the egress could be synchronized (including an offset for the time of flight of the data, or referencing the measurements to a particular packet) to ensure any counters refer to the same span of packets.

3. A method that communicates the measured values at the ingress and egress to the CB Measurement function. This could use several methods including sending return measurement packets (or control messages) from a receiver to a trigger function at the sender; an implementation using Operations, Administration and Management (OAM); or sending an in-band signaling datagram to the trigger function. This could also be implemented purely as a control-plane function, e.g., using a software-defined network controller.

4. A measurement function that combines the ingress and egress measurements to assess the present level of network congestion. (For example, the loss rate for each measurement interval could be deduced from calculating the difference between ingress and egress counter values.) Note the method does not require high accuracy for the period of the measurement interval (or therefore the measured value, since isolated and/or infrequent loss events need to be disregarded).

5. A trigger function that determines whether the measurements indicate persistent excessive congestion. This function defines an appropriate threshold for determining that there is persistent excessive congestion between the ingress and egress. This preferably considers a rate or ratio, rather than an absolute value (e.g., more than 10% loss, but other methods could also be based on the rate of transmission as well as the loss rate). The CB is triggered when the threshold is exceeded in multiple measurement intervals (e.g., three successive measurements). Designs need to be robust so that single or spurious events do not trigger a reaction.

6. A reaction that is applied at the ingress when the CB is triggered. This seeks to automatically remove the traffic causing persistent excessive congestion.

7. A feedback control mechanism that triggers when either the ingress and egress measurements are not available, since this also could indicate a loss of control packets (also a symptom of heavy congestion or inability to control the load).

3.2. Other Network Topologies

A CB can be deployed in networks with topologies different from that presented in Figures 1 and 2. This section describes examples of such usage and possible places where functions can be implemented.

3.2.1. Use with a Multicast Control/Routing Protocol

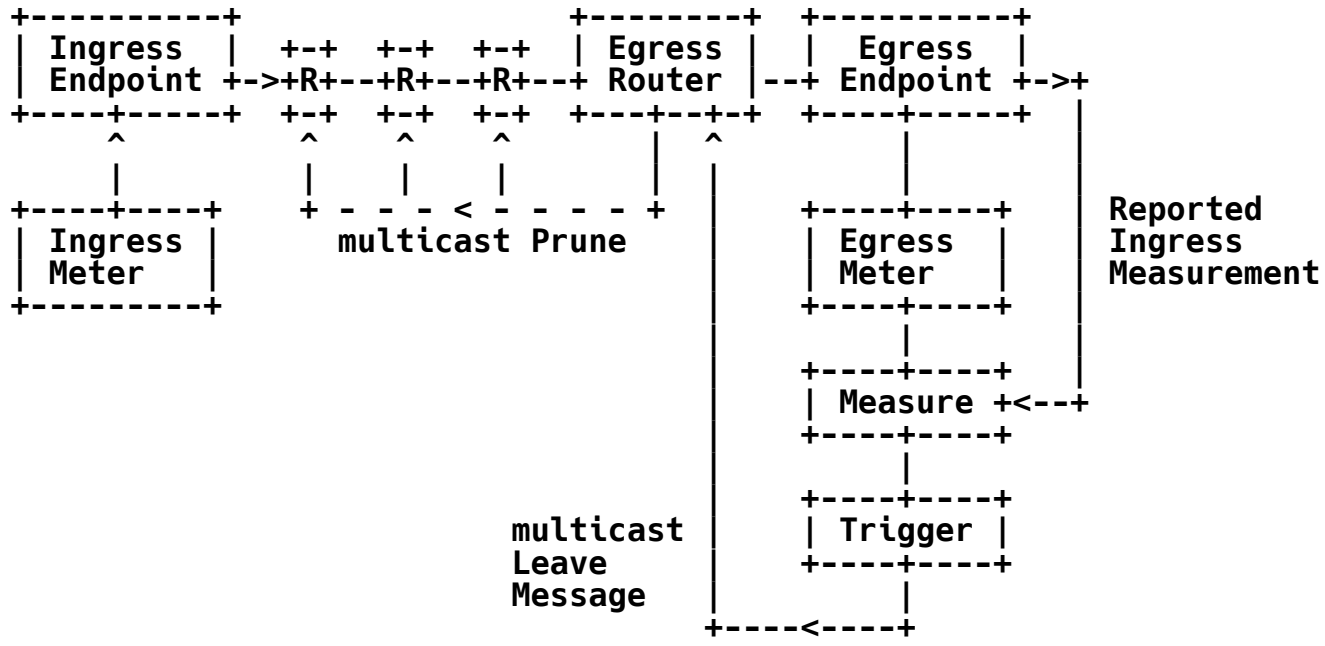

Figure 3: An example of a multicast CB controlling the end-to-end path between an ingress endpoint and an egress endpoint.

Figure 3 shows one example of how a multicast CB could be implemented at a pair of multicast endpoints (e.g., to implement a Fast-Trip CB, Section 5.1). The ingress endpoint (the sender that sources the multicast traffic) meters the ingress load, generating an ingress measurement (e.g., recording timestamped packet counts), and it sends this measurement to the multicast group together with the traffic it has measured.

Routers along a multicast path forward the multicast traffic (including the ingress measurement) to all active endpoint receivers. Each last hop (egress) router forwards the traffic to one or more egress endpoints.

In Figure 3, each endpoint includes a meter that performs a local egress load measurement. An endpoint also extracts the received ingress measurement from the traffic and compares the ingress and egress measurements to determine if the CB ought to be triggered. This measurement has to be robust to loss (see the previous section). If the CB is triggered, it generates a multicast leave message for the egress (e.g., an IGMP or MLD message sent to the last-hop router), which causes the upstream router to cease forwarding traffic to the egress endpoint [RFC1112].

Any multicast router that has no active receivers for a particular multicast group will prune traffic for that group, sending a prune message to its upstream router. This starts the process of releasing the capacity used by the traffic and is a standard multicast routing function (e.g., using Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol [RFC7761]). Each egress operates autonomously, and the CB "reaction" is executed by the multicast control plane (e.g., by PIM) requiring no explicit signaling by the CB along the communication path used for the control messages. Note there is no direct communication with the ingress; hence, a triggered CB only controls traffic downstream of the first-hop multicast router. It does not stop traffic flowing from the sender to the first-hop router; this is common practice for multicast deployment.

The method could also be used with a multicast tunnel or subnetwork (e.g., Section 5.2, Section 5.3), where a meter at the ingress generates additional control messages to carry the measurement data towards the egress where the egress metering is implemented.

3.2.2. Use with Control Protocols Supporting Pre-provisioned Capacity

Some paths are provisioned using a control protocol, e.g., flows provisioned using the Multiprotocol Label Switching (MPLS) services, paths provisioned using the Resource Reservation Protocol (RSVP), networks utilizing Software-Defined Network (SDN) functions, or admission-controlled Differentiated Services. Figure 1 shows one expected use case, where in this usage a separate device could be used to perform the measurement and trigger functions. The reaction generated by the trigger could take the form of a network-control message sent to the ingress and/or other network elements causing these elements to react to the CB. Examples of this type of use are provided in Section 5.3.

3.2.3. Unidirectional CBs over Controlled Paths

A CB can be used to control unidirectional UDP traffic, providing that there is a communication path that can be used for control messages to connect the functional components at the ingress and egress. This communication path for the control messages can exist in networks for which the traffic flow is purely unidirectional. For example, a multicast stream that sends packets across an Internet path and can use multicast routing to prune flows to shed network load. Some other types of subnetwork also utilize control protocols that can be used to control traffic flows.

4. Requirements for a Network Transport CB

The requirements for implementing a CB are:

1. There needs to be a communication path for control messages to carry measurement data from the ingress meter and from the egress meter to the point of measurement. (Requirements 16-18 relate to the transmission of control messages.)

2. A CB is REQUIRED to define a measurement period over which the CB Measurement function measures the level of congestion or loss. This method does not have to detect individual packet loss, but it MUST have a way to know that packets have been lost/marked from the traffic flow.

3. An egress meter can also count ECN [RFC3168] Congestion Experienced (CE) marks as a part of measurement of congestion, but in this case, loss MUST also be measured to provide a complete view of the level of congestion. For tunnels, [CONGESTION-FEEDBACK] describes a way to measure both loss and ECN-marking; these measurements could be used on a relatively short timescale to drive a congestion control response and/or aggregated over a longer timescale with a higher trigger threshold to drive a CB. Subsequent bullet items in this section discuss the necessity of using a longer timescale and a higher trigger threshold.

4. The measurement period used by a CB Measurement function MUST be longer than the time that current Congestion Control algorithms need to reduce their rate following detection of congestion. This is important because end-to-end Congestion Control algorithms require at least one RTT to notify and adjust the traffic when congestion is experienced, and congestion bottlenecks can share traffic with a diverse range of end-to-end RTTs. The measurement period is therefore expected to be significantly longer than the RTT experienced by the CB itself.

5. If necessary, a CB MAY combine successive individual meter samples from the ingress and egress to ensure observation of an average measurement over a sufficiently long interval. (Note when meter samples need to be combined, the combination needs to reflect the sum of the individual sample counts divided by the total time/volume over which the samples were measured. Individual samples over different intervals cannot be directly combined to generate an average value.)

6. A CB MUST be constructed so that it does not trigger under light or intermittent congestion (see requirements 7-9).

7. A CB is REQUIRED to define a threshold to determine whether the measured congestion is considered excessive.

8. A CB is REQUIRED to define the triggering interval, defining the period over which the trigger uses the collected measurements. CBs need to trigger over a sufficiently long period to avoid additionally penalizing flows with a long path RTT (e.g., many path RTTs).

9. A CB MUST be robust to multiple congestion events. This usually will define a number of measured persistent congestion events per triggering period. For example, a CB MAY combine the results of several measurement periods to determine if the CB is triggered (e.g., it is triggered when persistent excessive congestion is detected in three of the measurements within the triggering interval when more than three measurements were collected).

10. The normal reaction to a trigger SHOULD disable all traffic that contributed to congestion (otherwise, see requirements 11 and 12).

11. The reaction MUST be much more severe than that of a Congestion Control algorithm (such as TCP's congestion control [RFC5681] or TCP-Friendly Rate Control, TFRC [RFC5348]), because the CB reacts to more persistent congestion and operates over longer timescales (i.e., the overload condition will have persisted for a longer time before the CB is triggered).

12. A reaction that results in a reduction SHOULD result in reducing the traffic by at least an order of magnitude. A response that achieves the reduction by terminating flows, rather than randomly dropping packets, will often be more desirable to users of the service. A CB that reduces the rate of a flow, MUST continue to monitor the level of congestion and MUST further react to reduce the rate if the CB is again triggered.

13. The reaction to a triggered CB MUST continue for a period that is at least the triggering interval. Operator intervention will usually be required to restore a flow. If an automated response is needed to reset the trigger, then this needs to not be immediate. The design of an automated reset mechanism needs to be sufficiently conservative that it does not adversely interact with other mechanisms (including other CB algorithms that control traffic over a common path). It SHOULD NOT perform an automated reset when there is evidence of continued congestion.

14. A CB trigger SHOULD be regarded as an abnormal network event. As such, this event SHOULD be logged. The measurements that lead to triggering of the CB SHOULD also be logged.

15. The control communication needs to carry measurements (requirement 1) and, in some uses, also needs to transmit trigger messages to the ingress. This control communication may be in or out of band. The use of in-band communication is RECOMMENDED when either design would be possible. The preferred CB design is one that triggers when it fails to receive measurement reports that indicate an absence of congestion, in contrast to relying on the successful transmission of a "congested" signal back to the sender. (The feedback signal could itself be lost under congestion).

    In Band: An in-band control method SHOULD assume that loss of control messages is an indication of potential congestion on the path, and repeated loss ought to cause the CB to be triggered. This design has the advantage that it provides fate-sharing of the traffic flow(s) and the control communications. This fate-sharing property is weaker when some or all of the measured traffic is sent using a path that differs from the path taken by the control traffic (e.g., where traffic and control messages follow a different path due to use of equal-cost multipath routing, traffic engineering, or tunnels for specific types of traffic).

    Out of Band: An out-of-band control method SHOULD NOT trigger a CB reaction when there is loss of control messages (e.g., a loss of measurements). This avoids failure amplification/ propagation when the measurement and data paths fail independently. A failure of an out-of-band communication path SHOULD be regarded as an abnormal network event and be handled as appropriate for the network; for example, this event SHOULD be logged, and additional network operator action might be appropriate, depending on the network and the traffic involved.

16. The control communication MUST be designed to be robust to packet loss. A control message can be lost if there is a failure of the communication path used for the control messages, loss is likely also to be experienced during congestion/ overload. This does not imply that it is desirable to provide reliable delivery (e.g., over TCP), since this can incur additional delay in responding to congestion. Appropriate mechanisms could be to duplicate control messages to provide increased robustness to loss and/or to regard a lack of control traffic as an indication that excessive congestion could be

being experienced [RFC8085]. If control message traffic is sent over a shared path, it is RECOMMENDED that this control traffic is prioritized to reduce the probability of loss under congestion. Control traffic also needs to be considered when provisioning a network that uses a CB.

17. There are security requirements for the control communication between endpoints and/or network devices (Section 7). The authenticity of the source and integrity of the control messages (measurements and triggers) MUST be protected from off-path attacks. When there is a risk of an on-path attack, a cryptographic authentication mechanism for all control/measurement messages is RECOMMENDED.

5. Examples of CBs

There are multiple types of CB that could be defined for use in different deployment cases. There could be cases where a flow becomes controlled by multiple CBs (e.g., when the traffic of an end-to-end flow is carried in a tunnel within the network). This section provides examples of different types of CB.

5.1. A Fast-Trip CB

[RFC2309] discusses the dangers of congestion unresponsive flows and states that "all UDP-based streaming applications should incorporate effective congestion avoidance mechanisms." Some applications do not use a full-featured transport (TCP, SCTP, DCCP). These applications (e.g., using UDP and its UDP-Lite variant) need to provide appropriate congestion avoidance. Guidance for applications that do not use congestion-controlled transports is provided in [RFC8085]. Such mechanisms can be designed to react on much shorter timescales than a CB, that only observes a traffic envelope. Congestion control methods can also interact with an application to more effectively control its sending rate.

A Fast-trip CB is the most responsive form of CB. It has a response time that is only slightly larger than that of the traffic that it controls. It is suited to traffic with well-understood characteristics (and could include one or more trigger functions specifically tailored the type of traffic for which it is designed). It is not suited to arbitrary network traffic and could be unsuitable for traffic aggregates, since it could prematurely trigger (e.g., when the combined traffic from multiple congestion-controlled flows leads to short-term overload).

Although the mechanisms can be implemented in RTP-aware network devices, these mechanisms are also suitable for implementation in endpoints (e.g., as a part of the transport system) where they can also complement end-to-end congestion control methods. A shorter response time enables these mechanisms to triggers before other forms of CB (e.g., CBs operating on traffic aggregates at a point along the network path).

5.1.1. A Fast-Trip CB for RTP

A set of Fast-Trip CB methods have been specified for use together by a Real-time Transport Protocol (RTP) flow using the RTP/AVP Profile [RFC8083]. It is expected that, in the absence of severe congestion, all RTP applications running on best-effort IP networks will be able to run without triggering these CBs. An RTP Fast-Trip CB is therefore implemented as a fail-safe that, when triggered, will terminate RTP traffic.

The sending endpoint monitors reception of in-band RTP Control Protocol (RTCP) reception report blocks, as contained in sender report (SR) or receiver report (RR) packets, that convey reception quality feedback information. This is used to measure (congestion) loss, possibly in combination with ECN [RFC6679].

The CB action (shutdown of the flow) triggers when any of the following trigger conditions are true:

1. An RTP CB triggers on reported lack of progress.

2. An RTP CB triggers when no receiver reports messages are received.

3. An RTP CB triggers when the long-term RTP throughput (over many RTTs) exceeds a hard upper limit determined by a method that resembles TCP-Friendly Rate Control (TFRC).

4. An RTP CB includes the notion of Media Usability. This CB is triggered when the quality of the transported media falls below some required minimum acceptable quality.

5.2. A Slow-Trip CB

A Slow-Trip CB could be implemented in an endpoint or network device. This type of CB is much slower at responding to congestion than a Fast-Trip CB. This is expected to be more common.

One example where a Slow-Trip CB is needed is where flows or traffic-aggregates use a tunnel or encapsulation and the flows within the tunnel do not all support TCP-style congestion control (e.g., TCP, SCTP, TFRC), see [RFC8085], Section 3.1.3. A use case is where tunnels are deployed in the general Internet (rather than "controlled environments" within an Internet service provider or enterprise network), especially when the tunnel could need to cross a customer access router.

5.3. A Managed CB

A managed CB is implemented in the signaling protocol or management plane that relates to the traffic aggregate being controlled. This type of CB is typically applicable when the deployment is within a "controlled environment".

A CB requires more than the ability to determine that a network path is forwarding data or to measure the rate of a path -- which are often normal network operational functions. There is an additional need to determine a metric for congestion on the path and to trigger a reaction when a threshold is crossed that indicates persistent excessive congestion.

The control messages can use either in-band or out-of-band communications.

5.3.1. A Managed CB for SAToP Pseudowires

Section 8 of [RFC4553], SAToP Pseudowire Emulation Edge-to-Edge (PWE3), describes an example of a managed CB for isochronous flows.

If such flows were to run over a pre-provisioned (e.g., Multiprotocol Label Switching, MPLS) infrastructure, then it could be expected that the PW would not experience congestion, because a flow is not expected to either increase (or decrease) their rate. If, instead, PW traffic is multiplexed with other traffic over the general Internet, it could experience congestion. [RFC4553] states: "If SAToP PWs run over a PSN providing best-effort service, they SHOULD monitor packet loss in order to detect 'severe congestion'." The currently recommended measurement period is 1 second, and the trigger operates when there are more than three measured Severely Errored Seconds (SES) within a period. [RFC4553] goes on to state that "If such a condition is detected, a SAToP PW ought to shut down bi-directionally for some period of time...".

The concept was that when the packet-loss ratio (congestion) level increased above a threshold, the PW was, by default, disabled. This use case considered fixed-rate transmission, where the PW had no reasonable way to shed load.

The trigger needs to be set at a rate at which the PW is likely to experience a serious problem, possibly making the service noncompliant. At this point, triggering the CB would remove the traffic preventing undue impact on congestion-responsive traffic (e.g., TCP). Part of the rationale was that high-loss ratios typically indicated that something was "broken" and ought to have already resulted in operator intervention and therefore now need to trigger this intervention.

An operator-based response to the triggering of a CB provides an opportunity for other action to restore the service quality (e.g., by shedding other loads or assigning additional capacity) or to consciously avoid reacting to the trigger while engineering a solution to the problem. This could require the trigger function to send a control message to a third location (e.g., a network operations center, NOC) that is responsible for operation of the tunnel ingress, rather than the tunnel ingress itself.

5.3.2. A Managed CB for Pseudowires (PWs)

Pseudowires (PWs) [RFC3985] have become a common mechanism for tunneling traffic, and they could compete for network resources both with other PWs and with non-PW traffic, such as TCP/IP flows.

[RFC7893] discusses congestion conditions that can arise when PWs compete with elastic (i.e., congestion responsive) network traffic (e.g., TCP traffic). Elastic PWs carrying IP traffic (see [RFC4448]) do not raise major concerns because all of the traffic involved responds, reducing the transmission rate when network congestion is detected.

In contrast, inelastic PWs (e.g., a fixed-bandwidth Time Division Multiplex, TDM [RFC4553] [RFC5086] [RFC5087]) have the potential to harm congestion-responsive traffic or to contribute to excessive congestion because inelastic PWs do not adjust their transmission rate in response to congestion. [RFC7893] analyses TDM PWs, with an initial conclusion that a TDM PW operating with a degree of loss that could result in congestion-related problems is also operating with a degree of loss that results in an unacceptable TDM service. For that reason, the document suggests that a managed CB that shuts down a PW when it persistently fails to deliver acceptable TDM service is a useful means for addressing these congestion concerns. (See Appendix A of [RFC7893] for further discussion.)

6. Examples in Which CBs May Not Be Needed

A CB is not required for a single congestion-controlled flow using TCP, SCTP, TFRC, etc. In these cases, the congestion control methods are already designed to prevent persistent excessive congestion.

6.1. CBs over Pre-provisioned Capacity

One common question is whether a CB is needed when a tunnel is deployed in a private network with pre-provisioned capacity.

In this case, compliant traffic that does not exceed the provisioned capacity ought not to result in persistent congestion. A CB will hence only be triggered when there is noncompliant traffic. It could be argued that this event ought never to happen -- but it could also be argued that the CB equally ought never to be triggered. If a CB were to be implemented, it will provide an appropriate response, if persistent congestion occurs in an operational network.

Implementing a CB will not reduce the performance of the flows, but in the event that persistent excessive congestion occurs, it protects network traffic that shares network capacity with these flows. It also protects network traffic from a failure when CB traffic is (re)routed to cause additional network load on a non-pre-provisioned path.

6.2. CBs with Tunnels Carrying Congestion-Controlled Traffic

IP-based traffic is generally assumed to be congestion controlled, i.e., it is assumed that the transport protocols generating IP-based traffic at the sender already employ mechanisms that are sufficient to address congestion on the path. Therefore, a question arises when people deploy a tunnel that is thought to carry only an aggregate of TCP traffic (or traffic using some other congestion control method): Is there an advantage in this case in using a CB?

TCP (and SCTP) traffic in a tunnel is expected to reduce the transmission rate when network congestion is detected. Other transports (e.g., using UDP) can employ mechanisms that are sufficient to address congestion on the path [RFC8085]. However, even if the individual flows sharing a tunnel each implement a congestion control mechanism, and individually reduce their transmission rate when network congestion is detected, the overall traffic resulting from the aggregate of the flows does not necessarily avoid persistent congestion. For instance, most congestion control mechanisms require long-lived flows to react to reduce the rate of a flow. An aggregate of many short flows could result in many flows terminating before they experience congestion.

It is also often impossible for a tunnel service provider to know that the tunnel only contains congestion-controlled traffic (e.g., Inspecting packet headers might not be possible). Some IP-based applications might not implement adequate mechanisms to address congestion. The important thing to note is that if the aggregate of the traffic does not result in persistent excessive congestion (impacting other flows), then the CB will not trigger. This is the expected case in this context -- so implementing a CB ought not to reduce performance of the tunnel, but in the event that persistent excessive congestion occurs, the CB protects other network traffic that shares capacity with the tunnel traffic.

6.3. CBs with Unidirectional Traffic and No Control Path

A one-way forwarding path could have no associated communication path for sending control messages; therefore, it cannot be controlled using a CB (compare with Section 3.2.3).

A one-way service could be provided using a path with dedicated pre-provisioned capacity that is not shared with other elastic Internet flows (i.e., flows that vary their rate). A forwarding path could also be shared with other flows. One way to mitigate the impact of traffic on the other flows is to manage the traffic envelope by using ingress policing. Supporting this type of traffic in the general Internet requires operator monitoring to detect and respond to persistent excessive congestion.

7. Security Considerations

All CB mechanisms rely upon coordination between the ingress and egress meters and communication with the trigger function. This is usually achieved by passing network-control information (or protocol messages) across the network. Timely operation of a CB depends on the choice of measurement period. If the receiver has an interval that is overly long, then the responsiveness of the CB decreases. This impacts the ability of the CB to detect and react to congestion. If the interval is too short, the CB could trigger prematurely resulting in insufficient time for other mechanisms to act and potentially resulting in unnecessary disruption to the service.

A CB could potentially be exploited by an attacker to mount a Denial-of-Service (DoS) attack against the traffic being controlled by the CB. Therefore, mechanisms need to be implemented to prevent attacks on the network-control information that would result in DoS.

The authenticity of the source and integrity of the control messages (measurements and triggers) MUST be protected from off-path attacks. Without protection, it could be trivial for an attacker to inject

fake or modified control/measurement messages (e.g., indicating high packet loss rates) causing a CB to trigger and therefore to mount a DoS attack that disrupts a flow.

Simple protection can be provided by using a randomized source port, or equivalent field in the packet header (such as the RTP SSRC value and the RTP sequence number) expected not to be known to an off-path attacker. Stronger protection can be achieved using a secure authentication protocol to mitigate this concern.

An attack on the control messages is relatively easy for an attacker on the control path when the messages are neither encrypted nor authenticated. Use of a cryptographic authentication mechanism for all control/measurement messages is RECOMMENDED to mitigate this concern, and would also provide protection from off-path attacks. There is a design trade-off between the cost of introducing cryptographic security for control messages and the desire to protect control communication. For some deployment scenarios, the value of additional protection from DoS attacks will therefore lead to a requirement to authenticate all control messages.

Transmission of network-control messages consumes network capacity. This control traffic needs to be considered in the design of a CB and could potentially add to network congestion. If this traffic is sent over a shared path, it is RECOMMENDED that this control traffic be prioritized to reduce the probability of loss under congestion. Control traffic also needs to be considered when provisioning a network that uses a CB.

The CB MUST be designed to be robust to packet loss that can also be experienced during congestion/overload. Loss of control messages could be a side-effect of a congested network, but it also could arise from other causes Section 4.

The security implications depend on the design of the mechanisms, the type of traffic being controlled and the intended deployment scenario. Each design of a CB MUST therefore evaluate whether the particular CB mechanism has new security implications.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <http://www.rfc-editor.org/info/rfc3168>.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <http://www.rfc-editor.org/info/rfc8085>.

8.2. Informative References

[CONGESTION-FEEDBACK] Wei, X., Zhu, L., and L. Deng, "Tunnel Congestion Feedback", Work in Progress, draft-ietf-tsvwg-tunnel-congestion-feedback-04, January 2017.

[Jacobson88] Jacobson, V., "Congestion Avoidance and Control", SIGCOMM Symposium proceedings on Communications architectures and protocols, August 1988.

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <http://www.rfc-editor.org/info/rfc1112>.

[RFC2309] 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, DOI 10.17487/RFC2309, April 1998, <http://www.rfc-editor.org/info/rfc2309>.

[RFC2914] Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, DOI 10.17487/RFC2914, September 2000, <http://www.rfc-editor.org/info/rfc2914>.

[RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <http://www.rfc-editor.org/info/rfc3985>.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <http://www.rfc-editor.org/info/rfc4448>.

[RFC4553] Vainshtein, A., Ed. and YJ. Stein, Ed., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, DOI 10.17487/RFC4553, June 2006, <http://www.rfc-editor.org/info/rfc4553>.

[RFC5086] Vainshtein, A., Ed., Sasson, I., Metz, E., Frost, T., and P. Pate, "Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)", RFC 5086, DOI 10.17487/RFC5086, December 2007, <http://www.rfc-editor.org/info/rfc5086>.

[RFC5087] Stein, Y(J)., Shashoua, R., Insler, R., and M. Anavi, "Time Division Multiplexing over IP (TDMoIP)", RFC 5087, DOI 10.17487/RFC5087, December 2007, <http://www.rfc-editor.org/info/rfc5087>.

[RFC5348] Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 5348, DOI 10.17487/RFC5348, September 2008, <http://www.rfc-editor.org/info/rfc5348>.

[RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, DOI 10.17487/RFC5681, September 2009, <http://www.rfc-editor.org/info/rfc5681>.

[RFC6679] Westerlund, M., Johansson, I., Perkins, C., O'Hanlon, P., and K. Carlberg, "Explicit Congestion Notification (ECN) for RTP over UDP", RFC 6679, DOI 10.17487/RFC6679, August 2012, <http://www.rfc-editor.org/info/rfc6679>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <http://www.rfc-editor.org/info/rfc7761>.

[RFC7893] Stein, Y(J)., Black, D., and B. Briscoe, "Pseudowire Congestion Considerations", RFC 7893, DOI 10.17487/RFC7893, June 2016, <http://www.rfc-editor.org/info/rfc7893>.

[RFC8083] Perkins, C. and V. Singh, "Multimedia Congestion Control: Circuit Breakers for Unicast RTP Sessions", RFC 8083, DOI 10.17487/RFC8083, March 2017, <http://www.rfc-editor.org/info/rfc8083>.

Acknowledgments

There are many people who have discussed and described the issues that have motivated this document. Contributions and comments included: Lars Eggert, Colin Perkins, David Black, Matt Mathis, Andrew McGregor, Bob Briscoe, and Eliot Lear. This work was partly funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700).

Author's Address

Godred Fairhurst
University of Aberdeen
School of Engineering
Fraser Noble Building
Aberdeen, Scotland AB24 3UE
United Kingdom

Email: gorry@erg.abdn.ac.uk
URI: http://www.erg.abdn.ac.uk