Internet Research Task Force (IRTF)
Request for Comments: 5783
Category: Informational
ISSN: 2070-1721
M. Welzl
University of Oslo
W. Eddy
MTI Systems
February 2010

Congestion Control in the RFC Series

Abstract

This document is an informational snapshot taken by the IRTF's Internet Congestion Control Research Group (ICCRG) in October 2008. It provides a survey of congestion control topics described by documents in the RFC series. This does not modify or update the specifications or status of the RFC documents that are discussed. It may be used as a reference or starting point for the future work of the research group, especially in noting gaps or open issues in the current IETF standards.

Status of This Memo

Table of Contents

1. Introduction

In this document, we define congestion control as the feedback-based adjustment of the rate at which data is sent into the network. Congestion control is an indispensable set of principles and mechanisms for maintaining the stability of the Internet. Congestion control has been closely associated with TCP since 1988 [Jac88], but there has also been a great deal of congestion control work outside of TCP (e.g., for real-time multimedia applications, multicast, and router-based mechanisms). Several such proposals have been produced within the IETF and published as RFCs, along with RFCs that give architectural guidance (e.g., by pointing out the importance of performing some form of congestion control). Several of these mechanisms are in use within the Internet.

When designing a new Internet transport protocol, it is therefore important to not only understand how congestion control works in TCP but also have a broader understanding of the other congestion control RFCs -- some give guidance, some of them describe mechanisms that may have a direct influence on a newly designed protocol, and some of them may only be "related work" worth knowing about. The purpose of this document is to facilitate and encourage this search for knowledge by providing an overview of RFCs related to congestion control that have been published thus far. This document is a product of the IRTF's Internet Congestion Control Research Group (ICCRG). It was developed because a strong grasp of the existing literature should benefit further ICCRG work. The ICCRG developed consensus on the content of this document during a two-year development period based on review comments and ICCRG mailing list discussions. A list of the main review contributors is contained in the Acknowledgements section of this document.

While the ICCRG agreed to the document's production, any opinions expressed are the authors' own, and as this document is not an IETF publication, it does not update or modify the status of any published RFCs. The format of this document is similar to an annotated bibliography. Although host and router requirements for congestion control functions are discussed, this is only an informational document and does not contain any formal standards bearing of its own.

Congestion control is a large and active topic, and so the scope of this document is limited to published RFCs and a small number of current working group drafts. This allows the document to focus on congestion control principles and mechanisms that are among the most well-supported, well-accepted, or widely used. Significant contributions to this subject also exist in both the academic literature and in the form of Internet-Drafts; however, we exclude these from this study. In many cases the RFC describing some mechanism will contain references to relevant academic publications in journals or conference proceedings that presented the research and validation of the mechanism. For instance, RFC 2581 cites Jacobson's 1988 SIGCOMM paper that has a less standards-oriented but more illustrative treatment and explanation of some of the mechanisms in RFC 2581.

The majority of the documents discussed here pertain to end-host-based congestion control. Many network-based mechanisms, such as a number of queue management algorithms, do not require any protocol exchanges between elements, but merely operate within a single host or router. Thus, network-based congestion control mechanisms have often not been described in any RFC, as they generally fall under the domain of implementation details that do not influence interoperability.

There are many RFCs related to Quality of Service (QoS), especially within the Integrated Services and Differentiated Services frameworks [RFC1633] [RFC2475] [RFC2998]. These QoS RFCs themselves deserve a similar bibliography to the one that this document provides for congestion control. We specifically do not include the vast amount of QoS work into the scope of this document, as it is a full field in its own right, and deals with issues that are mostly orthogonal to end-host congestion control and router queue management. Although there can certainly be interactions between QoS and congestion control mechanisms, scheduling mechanisms used to implement QoS (on either a per-flow or an aggregate basis), for instance, can be used independently of the end-host congestion control and queue management functions also in use. Similar arguments can be made for traffic-shaping, admission control, and other functions that are intended for QoS and are only side-notes for congestion control.

A similar argument can be made for excluding consideration of the media access control (MAC) layer protocols used by the links throughout a path. Although the MAC protocols implement various forms of resolving contention for shared links (and sometimes offer QoS services), these are also distinct from end-to-end congestion control. Furthermore, MAC protocols are not typically discussed in the RFC series, but they are defined in outside documents (e.g., IEEE standards), since the IETF does not generally work on link layers themselves. Few, if any, of the RFCs that describe mappings of IP onto various link layers directly discuss congestion control.

To organize the subject matter in this document, the content is classified into several broad categories. First, we list documents relating to Internet architecture and general architectural concepts in Section 2. Next, the congestion control algorithms used in the

TCP transport protocol are discussed in Section 3. Interactions between link properties and mechanisms with the kinds of algorithms and heuristics used within end-to-end congestion control are covered in Section 4. One method that has been developed by the IETF (and deployed to some extent) for allowing network-based and host-based congestion control to interact without dropping packets is the subject of Section 5.1. The congestion control algorithms used by unicast transport protocols other than TCP are described in Section 6. Work on congestion control for multicast transports and applications is listed in Section 7. RFCs that give guidance to developers of new algorithms are discussed in Section 8. Finally, documents that have historic significance, but perhaps not current direct technical application, have been classified into Section 9. Note that the use of the term "historic" here has nothing to do with the IETF’s formal classification of documents as having "Historic" status.

2. Architectural Documents

Some documents in this section contain architectural guidance and concerns, while others specify congestion-control-related mechanisms that are broadly applicable and have impacts on more than a single class of congestion control techniques. Some of these documents are direct products of the Internet Architecture Board (IAB), giving their guidance on specific aspects of congestion control in the Internet.

RFC 1122: "Requirements for Internet Hosts -- Communication Layers" (October 1989)

[RFC1122] formally mandates that hosts perform congestion control. For TCP, several congestion control features are described and listed as required elements of conforming implementations, and for UDP, RFC 1122 leaves congestion control as an issue for higher-layered protocols. Although sending and reacting to ICMP Source Quench packets is no longer recommended [RFC1812] [Gont10], the rest of the congestion control guidance in this RFC is still a basis for several current practices in TCP implementations.

RFC 1812: "Requirements for IP Version 4 Routers" (June 1995)

Numerous issues relevant to router behavior are discussed in [RFC1812], and requirements for routers to support are prescribed within the document. Portions of RFC 1812 that are particularly relevant to congestion control include the directive that routers SHOULD NOT originate ICMP Source Quench messages, discussion of precedence in queueing, and Section 5.3.6 titled "Congestion

Control" that recommends sizing buffers as a function of the product of the bandwidth of the link times the path delay of the flows using the link, and advises on the implementation of active queue management techniques.

RFC 1958: "Architectural Principles of the Internet" (June 1996)

Several guidelines for network systems design that have proven useful in the evolution of the Internet are sketched in [RFC1958]. Congestion control is not specifically mentioned or alluded to, but the general principles apply to congestion control. For instance, performing end-to-end functions at end nodes, lack of centralized control, heterogeneity, scalability, simplicity, avoiding options and parameters, etc., are all valid concerns in the design and assessment of congestion control schemes for the Internet.

RFC 2140: "TCP Control Block Interdependence" (April 1997)

[RFC2140] suggests that TCP connections between the same endpoints might share some information, including their congestion control state. To some degree, this is done in practice by a few current operating systems; for example, Linux currently has a destination cache with this information, but this behavior is not yet formally standardized or recognized as a best practice by the IETF.

RFC 2309: "Recommendations on Queue Management and Congestion Avoidance in the Internet" (April 1998)

[RFC2309] briefly discusses the history of congestion and the origin of congestion control in the Internet. The focus is mainly on network- or router-based queue management algorithms. This RFC recommends to test, standardize, and deploy Active Queue Management (AQM) in routers; it provides an overview of one such mechanism, Random Early Detection (RED), and explains how and why AQM mechanisms can improve the performance of the Internet. Finally, this document explains the danger of a possible "congestion collapse" from unresponsive flows and makes a strong recommendation to develop and eventually deploy router mechanisms to protect the Internet from such traffic.

Today, the advice in this document has been followed to some extent. Hardware and software vendors have been receptive, and AQM techniques are widely available in many popular dedicated commercial router products and even in more general operating systems that are sometimes used as routers. However, AQM techniques may not be enabled in default configurations of these systems, and it is often left to users and network engineers to enable and configure AQM mechanisms when desired. In some cases, enabling QoS mechanisms on a device also enables AQM mechanisms by default. The number of production routers that actually have these AQM features enabled is an open question.

RFC 2914 (BCP 41): "Congestion Control Principles" (September 2000)

[RFC2914] is an explanation of the principles of congestion control, and the IETF's Best Current Practice for congestion control design. It points out that there are an increasing number of applications that do not use TCP, and elaborates on the importance of performing congestion control for such traffic in order to prevent congestion collapse. The TCP Reno congestion control mechanisms are described as an example of end-to-end congestion control within transport protocols.

SCTP is one example of a non-TCP transport protocol that implements congestion control based on these principles. The developments of TFRC [RFC3448] and DCCP [RFC4340] are attempts to provide useful tools implementing those principles for applications with needs similar to streaming media, where TCP's reactions are too fast. It would be beneficial for users and the Internet itself if these carefully designed tools become widely deployed in place of other ad hoc schemes that may not be well-grounded in the congestion control principles. This replacement process is ongoing and not yet complete. Appropriate and usable congestion control schemes for non-TCP flows continue to be an open research area.

RFC 3124: "The Congestion Manager" (June 2001)

[RFC3124] specifies the Congestion Manager, an end-system service that realizes congestion control on a per-host-pair rather than a per-connection basis, which may be a more appropriate way to carry out congestion control. Using the Congestion Manager, multiple streams between two hosts (which may include TCP flows) can adapt to network congestion in a unified fashion.

This proposal is related to RFC 2140, discussed above, but with a wider scope than TCP. Because some pieces of its supporting architecture have not yet been specified, the Congestion Manager's techniques are not commonly used today and have not been widely implemented and deployed yet beyond experimental stacks. Sharing of congestion and path information between individual connections continues to be an open research area with branches in detecting shared bottlenecks when using multiple paths, caching of old state for faster startup, and sharing of current state and feedback.

RFC 3426: "General Architectural and Policy Considerations" (November 2002)

[RFC3426] lists a number of questions that can be answered for a particular technical solution to determine its architectural impact and desirability. These are valid for congestion control mechanisms, and end-point congestion management is used as an example case-study several times in RFC 3426. Two salient questions that RFC 3426 advises asking about proposed mechanisms are why they are needed in addition to existing protocols, and why they are needed at a certain layer rather than at other layers. These are particularly relevant for congestion control mechanisms since several already exist and since they can span network, transport, and application layers.

RFC 3439: "Some Internet Architectural Guidelines and Philosophy" (December 2002)

[RFC3439] supplements RFC 1958. Simplicity is stressed, as the unpredictable results of complexity (due to amplification and coupling) are described. Congestion control issues stemming from layering interactions between transport and lower protocols are presented, as well as other items relevant to congestion control, including asymmetry and the "myth of over-provisioning".

RFC 3714: "IAB Concerns Regarding Congestion Control for Voice Traffic in the Internet" (March 2004)

[RFC3714] can be seen as a follow-up to the concerns that were discussed in RFC 2914. It expresses the IAB’s concern over the lack of effective end-to-end congestion control for best-effort voice traffic, which is noted as being a current service with growing demand. An example of a VoIP connection between Atlanta, Georgia, USA, and Nairobi, Kenya, is given, where a single VoIP call consumed more than half of the access link capacity (which is normally shared across several different users). This example is used as the basis for further discussion, making it clear that using some form of congestion control for VoIP traffic is highly recommended.

3. TCP Congestion Control

The TCP specifications found in RFC 793 and its predecessors did not contain any discussion of using or managing a congestion window. Other than a simple retransmission timeout and flow control through the advertised receive window, TCP implementations based only on RFC 793 do not contain congestion control. As several congestion collapse events occurred on the Internet, it was later realized that congestion control was needed. The host requirements in RFC 1122 require conforming TCP implementations to implement Jacobson's slow start and congestion avoidance algorithms (later specified in RFC 2001 and then RFC 2581). RFC 1122 also recommends several other behaviors that influence congestion control like the Nagle algorithm, delayed acknowledgements, Jacobson's retransmission timeout (RTO) estimation algorithm, and exponential backoff of the retransmission timer.

Basic TCP congestion control is defined in RFC 2581, with many other RFCs that specify ancillary modifications and enhancements. RFC 2581 obsoletes the first proposed standard for TCP congestion control in RFC 2001. These two RFCs document the mechanisms that had already been in common use by TCP implementations for many years. The reader may refer to the TCP Roadmap [RFC4614] for more information on the RFCs that specifically describe TCP congestion control, as this material is not replicated here.

Recently, significant effort has been put into experimental TCP congestion control modifications for obtaining high throughput with reduced startup and recovery times. RFCs have been published on some of these modifications, including HighSpeed TCP [RFC3649], and Limited Slow-Start [RFC3742], but high-rate congestion control mechanisms are still considered an open issue in congestion control research. Other schemes have been published as Internet-Drafts or have been discussed a little by the IETF, but much of the work in this area has not been adopted within the IETF yet, so the majority of this work is outside the RFC series and may be discussed in other products of the ICCRG.

At the time of writing, the IETF's TCP Maintenance and Minor Extensions (TCPM) Working Group was developing an update to RFC 2581 to incorporate small changes from other documents and advance TCP congestion control mechanisms on the IETF Standards Track. The update also clarifies and revises some points. These include the definition of a duplicate ACK, initial congestion window and slow start threshold values, behavior in response to retransmission timeouts, the use of the limited transmit mechanism, and security with regards to misbehaving receivers that practice ACK division.

4. Challenging Link and Path Characteristics

Links with large and/or variable bandwidth-delay products have traditionally been problematic for congestion control schemes because they can distort the properties of the feedback loop. Links that either expose a high rate of packet losses to the upper layers, or use highly-persistent retransmission mechanisms to prevent losses also cause problems with some of the standard congestion control mechanisms. The documents in this section discuss challenging link characteristics; many of them were written by the Performance Implications of Link Characteristics (PILC) Working Group.

While these documents often refer to specific problems with TCP, the link characteristics that they describe can be expected to affect other congestion control mechanisms too. In particular, interactions between link properties and TCP congestion control will be shared by other protocols that use the similar congestion control behavior, such as SCTP [RFC4960] and DCCP with CCID 2 [RFC4341] (see Section 6), and should be taken into consideration by designers of congestion control mechanisms that utilize the same kind of feedback as TCP.

Some RFCs only make recommendations regarding the implementation and configuration of TCP based upon characteristics of special links. As these RFCs are so closely connected to the specification of TCP itself, they are not included in this document, but are listed in the TCP Roadmap [RFC4614].

RFC 2488 (BCP 28): "Enhancing TCP Over Satellite Channels using Standard Mechanisms" (January 1999)

The summary of recommendations in [RFC2488] came from the TCP over Satellite (TCPSAT) Working Group, whose goal was to identify the performance problems that TCP may have over satellite links and suggest mitigations. The document explains several ways that existing standards can be applied to improve the performance of basic TCP congestion control over paths with characteristics similar to those involving satellite links.

RFC 3135: "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations" (June 2001)

[RFC3135] is a survey of Performance Enhancing Proxies (PEPs) often employed to improve degraded TCP performance caused by characteristics of specific link environments, for example, in satellite, wireless WAN, and wireless LAN environments. Different types of PEPs are described as well as the mechanisms used to improve performance. While there is a specific focus on TCP in this document, PEPs can operate on any protocol, and the performance enhancements that PEPs achieve are often closely related to congestion control.

The use of PEPs has architectural implications as they sometimes violate end-to-end assumptions and can add complexity to the inner portions of a network. Certain types of PEPs are commonly used today in satellite or long-distance networking because it is easier to insert a small number of PEPs near problematic links than to upgrade the TCP implementations on all the end hosts that might use those links. One down-side is that their deployment raises some issues when introducing new or updated congestion control (CC) methods into these deployed networks, since the PEPs may be operating with undocumented algorithms, making assumptions about the end-host CC behavior, and/or altering packet fields that will affect the end-host CC behavior.

RFC 3150 (BCP 48): "End-to-end Performance Implications of Slow Links" (July 2001)

[RFC3150] makes performance-related recommendations for users of network paths that traverse "very low bit-rate" links. It includes a discussion of interactions between such links and TCP congestion control.

RFC 3155 (BCP 50): "End-to-end Performance Implications of Links with Errors" (August 2001)

Under the premise that several types of PEP have undesirable implications, [RFC3155] recommends end-to-end alternatives for improving TCP performance over paths with error-prone links.

RFC 3366 (BCP 62): "Advice to link designers on link Automatic Repeat reQuest (ARQ)" (August 2002)

Link-layer ARQ techniques are a popular means to increase the robustness of particular links to transmission errors via retransmission and acknowledgement mechanisms. As [RFC3366] explains, ARQ techniques on a link can interact poorly with TCP’s end-to-end congestion control if they lead to additional delay variation or reordering. This RFC gives some advice on limiting the extent of these types of problematic interactions. The proper balance between end-to-end and link-layer reliability mechanisms is still an open research issue that has been explored in many academic papers outside the IETF.

RFC 3449 (BCP 69): "TCP Performance Implications of Network Path Asymmetry" (December 2002)

[RFC3449] describes performance limitations of TCP when the capacity of the ACK path is limited. Several techniques to aid TCP in these circumstances are recommended as Best Current Practices, particularly ACK congestion control and sender pacing are relevant to other non-TCP congestion control schemes, outside the scope of this document. For instance, in the design of the Reliable Multicast Transport (RMT) protocols for multicast, preventing ACK-implosion at multicast sources can be seen as a form of ACK congestion control.

RFC 3481: "TCP over Second (2.5G) and Third (3G) Generation Wireless Networks" (February 2003)

Among other issues, some mobile data systems exhibit delay spikes, handovers, and bandwidth oscillation. [RFC3481] describes the problems that these conditions cause for TCP congestion control and how some TCP extensions can be used to mitigate them.

RFC 3819 (BCP 89): "Advice for Internet Subnetwork Designers" (July 2004)

Several issues in link design and optimization for carrying IP traffic are discussed in [RFC3819], which recommends Best Current Practices. Many of these principles are motivated by properties of TCP, but most of them also apply to other transport-layer congestion control techniques as well.

5. End-Host and Router Cooperative Signaling

Some RFCs define mechanisms that allow routers to add signaling information to packets that makes the network’s congestion state less of a mystery to end-host congestion controllers. Routers supporting these can signal information about the current congestion state to flows in-band, providing faster and finer-grained information than inference-based methods. Two examples of this are discussed in this section; the first directs sources to slow down in order to avoid losses, and the other assists in determining an appropriate starting rate for new flows.

5.1. Explicit Congestion Notification

Traditionally, under congestion, IP routers enqueue packets until some limit is reached, at which point packets are dropped. TCP, and other IETF transport protocols, use a stream of acknowledgements to infer these losses and take congestion control action. This section describes a more advanced way to signal congestion to sources before packet-dropping is required.

There are two Explicit Congestion Notification (ECN) bits in the IP header that enable an AQM mechanism (see [RFC2309] or Section 2) to convey congestion information to endpoints without dropping packets. This can significantly reduce the losses experienced by transport endpoints if they are responsive to ECN. While ECN is most frequently discussed in the context of TCP (and therefore included in the TCP Roadmap [RFC4614]), its applicability is broader, and ECN use has also been specified for protocols such as DCCP and SCTP.

RFC 2481: "A Proposal to add Explicit Congestion Notification (ECN) to IP" (January 1999) - Obsoleted by RFC 3168

[RFC2481] introduced ECN into the RFC series, describing when the Congestion Experienced (CE) bit in the IP header should be set in routers, and what modifications are needed to TCP to make it ECN-capable. It includes a discussion of issues related to nodes and routers that are non-compliant, IPsec tunnels, and dropped or corrupted packets, as well as a summary of related work. Many of these issues will also be faced by operators trying to deploy other network-based congestion control methods. RFC 2481 has been obsoleted by RFC 3168.

RFC 2884: "Performance Evaluation of Explicit Congestion Notification (ECN) in IP Networks" (July 2000)

[RFC2884] presents a performance study of ECN as specified in [RFC2481] using an implementation on the Linux operating system. The experiments focused on ECN for both bulk and transactional transfers, showing that there is improvement in throughput over TCP without ECN in the case of bulk transfers and substantial improvement for transactional transfers. Studies like this help to build the community’s confidence that extensions like ECN are both safe and valuable. Similar RFCs helped the community accept larger initial windows for TCP [RFC2414] [RFC2415] [RFC2416].

RFC 3168: "The Addition of Explicit Congestion Notification (ECN) to IP" (September 2001)

[RFC3168], which obsoletes [RFC2481], specifies the incorporation of ECN into TCP and IP. One notable change in this significantly extended specification is the definition of a bit combination that was not defined in [RFC2481], which can be used to realize a nonce that would prevent a receiver from falsely claiming that there was no congestion. Potential issues related to ECN are discussed at length, including those already included in [RFC2481] and backwards compatibility with implementations that would follow the specification in the obsoleted document.

ECN, as specified in RFC 3168, is implemented in several popular router and end-host platforms. It is in active use, to at least some extent. Problems with ECN "blackholes" (Internet routers misconfigured to discard packets with ECN-capable bits set) were discovered when ECN was enabled by default in some end-host operating systems. Fears about the persisting presence of these blackholes currently may be keeping ECN from being used by default in many end-host operating systems even though it is implemented as an option within them. Some measurements on ECN support and usability are available [PF01] [MAF04] [MAF05].

RFC 3540: "Robust Explicit Congestion Notification (ECN) Signaling with Nonces" (June 2003)

[RFC3540] specifies a nonce mechanism that uses an ECN bit combination that is not used in [RFC2481], but that is specified in [RFC3168] to allow a one-bit ECN nonce. This nonce mechanism includes a Nonce Sum (NS) field in the TCP header so that senders can ensure that ACKs that do not indicate congestion are credible. The mechanism improves the robustness of congestion control by preventing receivers from exploiting ECN to gain an unfair share of network bandwidth.

This nonce technique is not understood to have been widely implemented or deployed, and there has been some discussion as to whether the mechanism is really effective or is the best use of these bits (see emails to the IETF Transport Area Working Group (TSVWG) mailing list, in the thread "ECN nonce snag in TCP ESTATS MIB" from December 2006 - January 2007, or [MBJ07]).

5.2. Quick-Start

RFC 4782: "Quick-Start for TCP and IP" (January 2007)

Quick-Start provides a way for hosts to ask routers to help them select an initial sending rate, and use this rate rather than the traditional small initial congestion window and slow-start algorithm. [RFC4782] describes the Quick-Start mechanism and its use with TCP. In addition to discussing the benefits of Quick-Start, the document also discusses several limitations of the Quick-Start technique with respect to some types of tunnels in use over the Internet today and other potential costs of Quick-Start including those related to router design. Analysis of the effects of misbehaving entities and appendices containing design rationale and related work are also notably present in this RFC.

Many of the issues discussed in RFC 4782, including router architecture, network design / tunnels, and misbehaving agents are all challenges relevant to other proposals that try to add router assistance into the network. The consideration of these issues can be illustrative for other protocol designers, even if they are not interested in Quick-Start itself.

6. Non-TCP Unicast Congestion Control

In the past, TCP dominated Internet traffic, as it was used for many of the popular applications (email, web browsing, file transfer, remote login, etc.). The majority of early congestion control work focused on TCP, and the introduction of congestion control into TCP alone is often credited with saving the Internet from additional congestion collapse events. Today, TCP has been joined by other transport protocols (e.g., custom UDP-based protocols, SCTP, DCCP, RTP over UDP [RFC3550], etc.), and so having properly functioning congestion control within these other protocols is important for the Internet’s health (as explained in RFC 3714, for instance, or see the discussion of the "congestion control arms race" scenario in RFC 2914). Documents that describe unicast congestion control methods for non-TCP transport protocols have been grouped into this section.

RFC 4960: "Stream Control Transmission Protocol" (September 2007)

SCTP congestion control is very similar to TCP with Selective Acknowledgements, but there are some differences, as described in Section 7.1 of [RFC4960]. The major difference lies in the fact that SCTP supports multihoming, whereas TCP does not. Thus, SCTP keeps a different set of congestion control parameters for each destination address within an association, whereas TCP only keeps a single set of congestion control parameters per connection.

RFC 5348: "TCP Friendly Rate Control (TFRC): Protocol Specification" (September 2008)

[RFC5348], which obsoletes [RFC3448], specifies TCP-Friendly Rate Control (TFRC), a rate-based congestion control mechanism for unicast flows operating in a best-effort Internet environment where flows are competing with standard TCP traffic. TFRC ensures conformance with TCP by continuously calculating the rate that a TCP sender would obtain under similar circumstances using a slightly simplified version of the TCP Reno throughput equation in [PFTK98]. Its sending rate is smoother than the rate of TCP, making it suitable for multimedia applications. TFRC is not a wire protocol but rather a mechanism that could, for instance, be used within a UDP-based application, in a transport protocol such as RTP, or in the context of endpoint congestion management [RFC3124].

RFC 3550: "RTP: A Transport Protocol for Real-Time Applications" (July 2003)

[RFC3550] specifies the real-time transport protocol RTP along with its control protocol RTCP. RTP/RTCP does not prescribe a specific congestion control behavior, but it is recommended that such a behavior be specified in each RTP profile (which is due to the fact that the potential for reducing the sending rate is often content dependent in the case of real-time streams). Specifically, [RFC3550] states: "For some profiles, it may be sufficient to include an applicability statement restricting the use of that profile to environments where congestion is avoided by engineering. For other profiles, specific methods such as data rate adaptation based on RTCP feedback may be required". [RFC4585], which discusses RTCP feedback and adaptation mechanisms, points out that RTCP feedback may operate on much slower timescales than transport layer feedback mechanisms, and that additional mechanisms are therefore required to perform proper congestion control. One way to make use of such additional mechanisms is to run RTP over DCCP.

RFC 4336: "Problem Statement for the Datagram Congestion Control Protocol (DCCP)" (March 2006)

[RFC4336] provides the motivation leading to the design of DCCP. In doing so, other possibilities of implementing similar functionality are discussed, including unreliable extensions of SCTP, RTP-based congestion control, and providing congestion control above or below UDP.

RFC 4340: "Datagram Congestion Control Protocol" (March 2006)

[RFC4340] specifies DCCP, the Datagram Congestion Control Protocol. This protocol provides bidirectional unicast connections of congestion-controlled unreliable datagrams. It is suitable for applications that can benefit from control over the tradeoff between timeliness and reliability. The core DCCP specification does not include a specific congestion control behavior; rather, it functions as a framework for such mechanisms, which can be selected via the Congestion Control Identifier (CCID).

RFC 4341: "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 2: TCP-like Congestion Control" (March 2006)

[RFC4341] is the specification of TCP-like congestion control within DCCP. This should be used by senders who would like to take advantage of the available bandwidth in an environment with rapidly changing conditions, and who are able to adapt to the abrupt changes in the congestion window typical of TCP's Additive Increase Multiplicative Decrease (AIMD) congestion control. ECN is also supported within RFC 4341.

RFC 4342: "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 3: TCP-Friendly Rate Control (TFRC)" (March 2006)

[RFC4342] is the specification of TFRC congestion control as described in [RFC3448] for DCCP. This should be used by senders who want a TCP-friendly sending rate, possibly with Explicit Congestion Notification (ECN), while minimizing abrupt rate changes.

7. Multicast Congestion Control

In the IETF, congestion control for multicast (one-to-many) communication has primarily been tackled in the Reliable Multicast Transport (RMT) Working Group. Except for [RFC2357] and [RFC3208], all the documents in this section were written by this group. Since a "one size fits all" protocol cannot meet the requirements of all possible applications in this space, the approach taken is a modular one, consisting of "protocol cores" and "building blocks". Multiple congestion control building blocks have been defined, providing both sender-driven and receiver-driven congestion control methods that differ widely in their assumptions and behavior.

RFC 2357: "IETF Criteria for Evaluating Reliable Multicast Transport and Application Protocols" (June 1998)

Some early multicast content dissemination proposals did not incorporate proper congestion control; this is pointed out as being a severe mistake in [RFC2357], as large-scale multicast applications have the potential to do vast congestion-related damage. This document clearly makes the case that congestion control mechanisms should be developed and incorporated into multicast content dissemination protocols intended for use over the Internet.

RFC 2887: "The Reliable Multicast Design Space for Bulk Data Transfer" (August 2000)

Several classes of potential congestion control schemes for single-sender multicast protocols are briefly sketched as possibilities, but no specific protocols are developed or selected in [RFC2887].

RFC 3048: "Reliable Multicast Transport Building Blocks for One-to-Many Bulk-Data Transfer" (January 2001)

[RFC3048] discusses the building block approach to RMT protocols and mentions that several different congestion control building blocks may be required in order to deal with different situations. Some of the possible interactions between building blocks for congestion control and those for Forward Error Correction (FEC), acknowledgement, and group management are also mentioned.

RFC 3208: "PGM Reliable Transport Protocol Specification" (December 2001)

Pragmatic General Multicast (PGM) is a reliable multicast transport protocol for applications that require ordered or unordered, duplicate-free, multicast data delivery from multiple sources to multiple receivers. As discussed in [RFC3208]'s Appendix B, a PGM protocol source can request congestion control feedback from both network elements (routers) and receivers (end hosts). These reports can indicate the load on the worst link in a particular path, or the load on the worst path. The actual procedure used in response to this feedback is not part of RFC 3208, but the notion of using multicast routers to assist in congestion control is significant.

RFC 3450: "Asynchronous Layered Coding (ALC) Protocol Instantiation" (December 2002)

[RFC3450] specifies ALC, a rough header format using the RMT building blocks, that can be used by multicast content dissemination protocols. ALC is intended to use a multi-rate congestion control building block, where the sender does not require any feedback, but where multiple multicast groups with different transmission rates are available within and ALC session, and receivers control their rates by joining or leaving groups.

RFC 3738: "Wave and Equation Based Rate Control (WEBRC) Building Block" (April 2004)

The WEBRC mechanism defined in [RFC3738] is a receiver-driven form of congestion control, where each receiver in a multicast group can determine the individual rate at which packets are delivered to it. WEBRC senders create a base channel for control information and several multicast channels for data transmission that each send packets at a varying rate in the form of a wave. The receivers dynamically join and leave channels at chosen points within the wave of sending rates to obtain the desired overall receive rate based on an equation using the estimated loss probability and round-trip time within an epoch. WEBRC is compatible for use within ALC.

RFC 4654: "TCP-Friendly Multicast Congestion Control (TFMCC): Protocol Specification" (August 2006)

TFMCC, as described in [RFC4654], is a sender-driven congestion control mechanism, where the received rate for the entire multicast group is determined by the worst-connected receiver. TFMCC builds upon TFRC, but scales down the feedback to prevent ACK-implosion effects by having receivers suppress their feedback unless they perceive it to be the worst among the reception group.

8. Guidance for Developing and Analyzing Congestion Control Techniques

Some recently published RFCs discuss the properties of congestion control protocols that are "safe" for Internet deployment, as well as how to measure the properties of congestion control mechanisms and transport protocols. These documents are particularly relevant to the ICCRG as some of the group’s activities involve reviewing congestion control proposals that have been brought to the IETF for publication (see http://www.ietf.org/iesg/statement/congestion-control.html).

RFC 5033 (BCP 133): "Specifying New Congestion Control Algorithms" (August 2007)

The concurrent development of multiple TCP modifications for high-rate use and the deployments of these modifications on the Internet prompted [RFC5033] to be written. RFC 5033 comes from the Transport Area Working Group (TSVWG), and gives guidance on the classes of Experimental RFC that can be published to document algorithms that are either encouraged for investigation on the Internet, and those that are only encouraged for experimentation in less-critical environments. It has been described as a list of things for people to think about when creating new congestion control techniques that they are planning to widely deploy.

RFC 5166: "Metrics for the Evaluation of Congestion Control Mechanisms" (March 2008)

The IRTF Transport Modeling Research Group (TMRG) produced [RFC5166] to describe the set of metrics and related tradeoffs between metrics that can be used to compare, contrast, and evaluate congestion control techniques. This RFC gives an overview of many such metrics, and gives references to their detailed descriptions.

9. Historic Interest

Early in the RFC series, there are many documents that represent an author's thoughts on a subject or brief summaries from measurement and experimentation, rather than the result of a long formal IETF process. Some of the RFCs listed in this section have this distinction.

RFC 889: "Internet Delay Experiments" (December 1983)

Based on reported measurement experiments, changes to the TCP retransmission timeout (RTO) calculation are suggested in [RFC0889]. It is noted that the original TCP RTO calculation leads to congestion when a delay spike occurs because it takes too long for the RTO to adapt, leading to superfluous retransmissions.

RFC 896: "Congestion Control in IP/TCP Internetworks" (January 1984)

[RFC0896] is the first document known to the authors where the term "congestion collapse" was used. Here, it refers to the stable state that was observed when a sudden load on the net caused the round-trip time to rise faster than the sending hosts measured round-trip time could be updated. Two problems are discussed: the "small-packet problem" (now commonly known by the name "silly window syndrome") and the "source-quench problem", which is about inappropriately deciding when to send and how to react to ICMP Source Quench messages. Solutions for these problems are presented.

RFC 970: "On Packet Switches with Infinite Storage" (December 1985)

Using a thought experiment based on a router with infinite buffering capacity, [RFC0970] develops a different kind of congestion collapse scenario, where few useful packet transmissions occur due to the queue being longer than the time-to-live of the packets within it. As described in RFC 970, this scenario was also demonstrated using real equipment by the author.

The document also includes discussion of game-theoretic analysis of congestion control and obtaining fairness between behaving and non-behaving flows, by focusing on the order of scheduling packets within the buffer rather than the actual allocation of buffer space between flows.

RFC 1016: "Something a Host Could Do with Source Quench: The Source Quench Introduced Delay (SQuID)" (July 1987)

[RFC1016] outlines a rate-based congestion control mechanism where end-hosts use Source Quench packets from routers to adjust their sending rates. RFC 1016 also suggests sending congestion notifications before queues are actually full, at a rate that increases with the current queue occupancy. This strategy has been used in several other AQM mechanisms, notably RED [FJ93].

RFC 1254: "Gateway Congestion Control Survey" (August 1991)

[RFC1254] is a survey of congestion control approaches in routers that first discusses general congestion control performance goals (such as fairness), and then elaborates on the use of Source Quench messages (which are now discouraged, as they have been found ineffective), Random Drop (which would now be called "Active Queue Management"), Congestion Indication (DEC Bit; an early form of ECN), "Selective Feedback Congestion Indication" (one particular method for applying ECN), and Fair Queuing. Finally, end-system congestion control policies are discussed, including Jacobson’s well-known algorithms [Jac88] and their predecessor -- "CUTE" [Jain86].

10. Security Considerations

This document introduces no new security considerations. Each RFC listed in this document discusses the security considerations of the specification it contains.

11. Acknowledgements

Several participants in the ICCRG contributed useful comments in the development of this document, including Rex Buddenberg, Mitchell Erblichs, Lachlan Andrew, Sally Floyd, Stephen Farrell, Gorry Fairhurst, Lars Eggert, Mark Allman, and Juergen Schoenwaelder.

12. Informative References

[FJ93] Floyd, S. and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, volume 1, number 4, August 1993.

[Gont10] Gont, F., "ICMP attacks against TCP", Work in Progress, January 2010.

[Jac88] Jacobson, V., "Congestion Avoidance and Control", Proceedings of ACM SIGCOMM 1988, in ACM Computer Communication Review, 18 (4), pp. 314-329.

[Jain86] Jain, R., "A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks", IEEE Journal on Selected Areas in Communications, volume 4, number 7, October 1986.

[MAF04] Medina, A., Allman, M., and S. Floyd, "Measuring Interactions Between Transport Protocols and Middleboxes", Proceedings of the Internet Measurement Conference 2004, August 2004.

[MAF05] Medina, A., Allman, M., and S. Floyd, "Measuring the Evolution of Transport Protocols in the Internet", ACM Computer Communications Review, volume 35, issue 2, April 2005.

[MBJ07] Moncaster, T., Briscoe, B., and A. Jacquet, "A TCP Test to Allow Senders to Identify Receiver Non-Compliance", Work in Progress, November 2007.

[PF01] Padhye, J. and S. Floyd, "On Inferring TCP Behavior", Proceedings of ACM SIGCOMM 2001, August 2001.

[PFTK98] Padhye, J., Firoiu, V., Towsley, D., and J. Kurose, "Modeling TCP Throughput: A Simple Model and its Empirical Validation", Proceedings of ACM SIGCOMM 1998.

[RFC0889] Mills, D., "Internet delay experiments", RFC 889, December 1983.

[RFC0896] Nagle, J., "Congestion control in IP/TCP internetworks", RFC 896, January 1984.

[RFC0970] Nagle, J., "On packet switches with infinite storage", RFC 970, December 1985.

[RFC1016] Prue, W. and J. Postel, "Something a host could do with source quench: The Source Quench Introduced Delay (SQuID)", RFC 1016, July 1987.

[RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.

[RFC1254] Mankin, A. and K. Ramakrishnan, "Gateway Congestion Control Survey", RFC 1254, August 1991.

[RFC1633] Braden, B., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[RFC1812] Baker, F., "Requirements for IP Version 4 Routers", RFC 1812, June 1995.

[RFC1958] Carpenter, B., "Architectural Principles of the Internet", RFC 1958, June 1996.

[RFC2001] Stevens, W., "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC 2001, January 1997.

[RFC2140] Touch, J., "TCP Control Block Interdependence", RFC 2140, April 1997.

[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, April 1998.

[RFC2357] Mankin, A., Romanov, A., Bradner, S., and V. Paxson, "IETF Criteria for Evaluating Reliable Multicast Transport and Application Protocols", RFC 2357, June 1998.

[RFC2414] Allman, M., Floyd, S., and C. Partridge, "Increasing TCP's Initial Window", RFC 2414, September 1998.

[RFC2415] Poduri, K., "Simulation Studies of Increased Initial TCP Window Size", RFC 2415, September 1998.

[RFC2416] Shepard, T. and C. Partridge, "When TCP Starts Up With Four Packets Into Only Three Buffers", RFC 2416, September 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2481] Ramakrishnan, K. and S. Floyd, "A Proposal to add Explicit Congestion Notification (ECN) to IP", RFC 2481, January 1999.

[RFC2488] Allman, M., Glover, D., and L. Sanchez, "Enhancing TCP Over Satellite Channels using Standard Mechanisms", BCP 28, RFC 2488, January 1999.

[RFC2581] Allman, M., Paxson, V., and W. Stevens, "TCP Congestion Control", RFC 2581, April 1999.

[RFC2884] Hadi Salim, J. and U. Ahmed, "Performance Evaluation of Explicit Congestion Notification (ECN) in IP Networks", RFC 2884, July 2000.

[RFC2887] Handley, M., Floyd, S., Whetten, B., Kermode, R., Vicisano, L., and M. Luby, "The Reliable Multicast Design Space for Bulk Data Transfer", RFC 2887, August 2000.

[RFC2914] Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, September 2000.

[RFC2998] 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.

[RFC3048] Whetten, B., Vicisano, L., Kermode, R., Handley, M., Floyd, S., and M. Luby, "Reliable Multicast Transport Building Blocks for One-to-Many Bulk-Data Transfer", RFC 3048, January 2001.

[RFC3124] Balakrishnan, H. and S. Seshan, "The Congestion Manager", RFC 3124, June 2001.

[RFC3135] Border, J., Kojo, M., Griner, J., Montenegro, G., and Z. Shelby, "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations", RFC 3135, June 2001.

[RFC3150] Dawkins, S., Montenegro, G., Kojo, M., and V. Magret, "End-to-end Performance Implications of Slow Links", BCP 48, RFC 3150, July 2001.

[RFC3155] Dawkins, S., Montenegro, G., Kojo, M., Magret, V., and N. Vaidya, "End-to-end Performance Implications of Links with Errors", BCP 50, RFC 3155, August 2001.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC3208] Speakman, T., Crowcroft, J., Gemmell, J., Farinacci, D., Lin, S., Leshchiner, D., Luby, M., Montgomery, T., Rizzo, L., Tweedly, A., Bhaskar, N., Edmonstone, R., Sumanasekera, R., and L. Vicisano, "PGM Reliable Transport Protocol Specification", RFC 3208, December 2001.

[RFC3366] Fairhurst, G. and L. Wood, "Advice to link designers on link Automatic Repeat reQuest (ARQ)", BCP 62, RFC 3366, August 2002.

[RFC3426] Floyd, S., "General Architectural and Policy Considerations", RFC 3426, November 2002.

[RFC3439] Bush, R. and D. Meyer, "Some Internet Architectural Guidelines and Philosophy", RFC 3439, December 2002.

[RFC3448] Handley, M., Floyd, S., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 3448, January 2003.

[RFC3449] Balakrishnan, H., Padmanabhan, V., Fairhurst, G., and M. Sooriyabandara, "TCP Performance Implications of Network Path Asymmetry", BCP 69, RFC 3449, December 2002.

[RFC3450] Luby, M., Gemmell, J., Vicisano, L., Rizzo, L., and J. Crowcroft, "Asynchronous Layered Coding (ALC) Protocol Instantiation", RFC 3450, December 2002.

[RFC3481] Inamura, H., Montenegro, G., Ludwig, R., Gurtov, A., and F. Khafizov, "TCP over Second (2.5G) and Third (3G) Generation Wireless Networks", BCP 71, RFC 3481, February 2003.

[RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit Congestion Notification (ECN) Signaling with Nonces", RFC 3540, June 2003.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC3649] Floyd, S., "HighSpeed TCP for Large Congestion Windows", RFC 3649, December 2003.

[RFC3714] Floyd, S. and J. Kempf, "IAB Concerns Regarding Congestion Control for Voice Traffic in the Internet", RFC 3714, March 2004.

[RFC3738] Luby, M. and V. Goyal, "Wave and Equation Based Rate Control (WEBRC) Building Block", RFC 3738, April 2004.

[RFC3742] Floyd, S., "Limited Slow-Start for TCP with Large Congestion Windows", RFC 3742, March 2004.

[RFC3819] Karn, P., Bormann, C., Fairhurst, G., Grossman, D., Ludwig, R., Mahdavi, J., Montenegro, G., Touch, J., and L. Wood, "Advice for Internet Subnetwork Designers", BCP 89, RFC 3819, July 2004.

[RFC4336] Floyd, S., Handley, M., and E. Kohler, "Problem Statement for the Datagram Congestion Control Protocol (DCCP)", RFC 4336, March 2006.

[RFC4340] Kohler, E., Handley, M., and S. Floyd, "Datagram Congestion Control Protocol (DCCP)", RFC 4340, March 2006.

[RFC4341] Floyd, S. and E. Kohler, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 2: TCP-like Congestion Control", RFC 4341, March 2006.

[RFC4342] Floyd, S., Kohler, E., and J. Padhye, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 3: TCP-Friendly Rate Control (TFRC)", RFC 4342, March 2006.

[RFC4585] Ott, J., Wenger, S., Sato, N., Burmeister, C., and J. Rey, "Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF)", RFC 4585, July 2006.

[RFC4614] Duke, M., Braden, R., Eddy, W., and E. Blanton, "A Roadmap for Transmission Control Protocol (TCP) Specification Documents", RFC 4614, September 2006.

[RFC4654] Widmer, J. and M. Handley, "TCP-Friendly Multicast Congestion Control (TFMCC): Protocol Specification", RFC 4654, August 2006.

[RFC4782] Floyd, S., Allman, M., Jain, A., and P. Sarolahti, "Quick-Start for TCP and IP", RFC 4782, January 2007.

[RFC4960] Stewart, R., Ed., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC5033] Floyd, S. and M. Allman, "Specifying New Congestion Control Algorithms", BCP 133, RFC 5033, August 2007.

[RFC5166] Floyd, S., "Metrics for the Evaluation of Congestion Control Mechanisms", RFC 5166, March 2008.

[RFC5348] Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 5348, September 2008.

Authors' Addresses

Michael Welzl
University of Oslo
Department of Informatics
PO Box 1080 Blindern
N-0316 Oslo, Norway

Phone: +47 22 85 24 20
EMail: michawe@ifi.uio.no

Wesley M. Eddy
MTI Systems
NASA Glenn Research Center
21000 Brookpark Rd, MS 500-ASRC
Cleveland, OH 44135

Phone: (216) 433-6682
EMail: wes@mti-systems.com