TSVWG Working Group
Internet-Draft
Updates: RFC 3819 (if published) (if approved)
Intended status: Best Current Practice
Expires: September 12, 2013

G. Fairhurst
University of Aberdeen
B. Briscoe
BT

March 11, 2013

Advice on network buffering
draft-fairhurst-tsvwg-buffers-00

Abstract

This document proposes an update to the advice given in RFC 3819. Subsequent research has altered understanding of buffer sizing and queue management. Therefore this document significantly revises the previous recommendations on buffering. The advice applies to all packet buffers, whether in network equipment, end hosts or middleboxes such as firewalls or NATs. And the advice applies to packet buffers at any layer: whether subnet, IP, transport or application.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 12, 2013.

Table of Contents

1. Introduction 2
2. Terminology 4
3. Updated Recommendations on Buffering 4
3.1. Recommendations Applicable to Any Buffer 4
3.2. Buffering recommendations for end hosts 5
3.3. Buffering recommendations for edge routers and switches . 5
3.4. Buffering recommendations for core routers and switches . 6
3.5. Recommendations on Flow Isolation 6
4. Buffer Management Methods 6
4.1. Examples of subnetwork buffering 6
4.2. Examples of methods for active buffer management 7
5. Security Considerations 7
6. IANA Considerations 7
7. Acknowledgments 7
8. References 8
8.1. Normative References 8
8.2. Informative References 8
Appendix A. vious IETF guidance for configuring network buffers 9
Appendix B. Revision notes 10
Authors' Addresses 10

1. Introduction

[RFC3819] provides guidance on the design of subnetworks and networking equipment. This document updates this guidance for the topic of Internet buffer configuration and control. The guidance is aimed at both equipment designers and network operators.

All networking devices use buffers to temporarily store packets that are waiting for transmission on an out-going link during traffic bursts or at times when the capacity of the ingress/egress changes.

The congestion control algorithms in TCP (and derivatives of TCP) are designed to try to fully utilise the link that has the least available capacity on the path across the network. This is called the bottleneck link. Network link capacities are typically arranged so that it will be rare for a bottleneck to arise in the network core. However, depending on prevailing patterns of traffic, any link might become the bottleneck (within the host, at an edge router, at a core router, at a switch in the subnet between routers or at some middlebox such as a firewall or a network address translator). Modern TCP stacks are capable of filling a link of any capacity.

A buffer that simply discards incoming packets when it is full is called a tail-drop buffer. A long-running TCP flow will fill a tail-drop buffer and keep it full, so that there is no longer any space to absorb bursts. This is called a standing queue. Packets arriving at the tail of a standing queue still work their way through the buffer until they emerge onto the link, but this introduces unnecessary delay to every packet, including those from other sessions sharing the link. This can intermittently add intolerable delay to a real-time interactive media session (e.g. voice or video). Also, most Web pages involve dozens of short back-and-forth exchanges, so adding even a small amount of queuing delay to each round can accumulate considerable delay in the completion of the whole task.

The recommended way to avoid these problems is to use an active queue management (AQM) algorithm in every potential bottleneck buffer (subnet, router, middlebox or host), and to enable explicit congestion notification (ECN). However, if AQM has not been implemented in existing equipment, the next best option is to at least size the buffer so that it is no larger than needed to absorb bursts.

This document gives advice on using and configuring AQM algorithms and ECN, and advice on buffer sizing in the absence of such algorithms.

The correct buffer size depends on the link rate, so a common problem is where equipment auto-adjusts its rate, often over a wide range, so the buffer size can be badly incorrect. Advice is also given on how to relate buffer auto-sizing algorithms to rate-adjusting algorithms, and the best static buffer size to configure if auto-sizing has not been implemented.

It is difficult to test whether a network might exhibit these problems. They only appear intermittently, because they depend on four pathologies co-inciding: i) a particular buffer has become the bottleneck for a long-running TCP flow, which depends on relative traffic levels in other links, ii) the TCP flow has run for long

enough to fill this buffer, iii) the buffer lacks AQM or the AQM is badly configured and iv) the buffer has been badly over-sized. When all four conditions co-incide, the delays can be bad enough to lead to support desk calls.

This document updates section 13 of RFC 3819, which gave guidance to subnet designers on the use and sizing of buffers. Appendix A reviews that guidance, which now requires considerable revision in the light of subsequent research. Also, whereas RFC 3819 addressed subnet designers, the advice in this document is relevant to a wider audience, because it concerns buffers wherever they are, including in end-systems and middleboxes not just in subnet technology.

2. Terminology

The document assumes familiarity with the terminology of RFC 3819 [RFC3819].

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].

The term active queue management (AQM) has been applied to technologies that work only at the packet level as well as technologies that identify and police flows with above average rates or that enforce flow-level or user-level policies such as fair queuing. For this document, we will use the term 'AQM' for technologies or parts of technologies that treat packets indiscriminately, and the term 'policing' for the additional technologies that attempt to enforce some level of behaviour or isolation at the flow or user level of granularity.

3. Updated Recommendations on Buffering

This section updates the rules for network buffers in section 13 of RFC 3819.

3.1. Recommendations Applicable to Any Buffer

XX Work in Progress, to be included in next revision XX

AQM is strongly recommended recommended for any buffer. Auto-tuned configuration is recommended.

Explicit Congestion Notification (ECN) [RFC3168] is also strongly recommended for any buffer (this avoids delays due to timeouts after loss). It is safe to enable ECN for routers and servers. If concerns arise over the use of ECN, this can be fully addressed by

turning off ECN support at the endpoint. If routers and servers were not to enable ECN, where it is deemed safe, it will not be possible for endpoints to turn it on.

Buffer size: if AQM is implemented, there is no harm in having a large buffer to absorb bursts. However, if there is no AQM, it is important to keep the buffer small.

- Too little buffering can result in poor utilisation of the egress link, since many traffic flows are not smooth-paced and bursts of traffic may fail to be buffered.

- Large buffers can help ensure full utilisation of the egress link, but excessive buffering results in slow response to congestion and in unnecessary delay experienced by any flow that shares the egress link. Such events are not uncommon, since a single long-lived connection using a modern TCP stack can fill any size of network buffer.

Auto-sizing is recommended if the line rate is adjustable or auto-adjusts (e.g. setting buffer time, not byte-size). If auto-sizing has not been implemented, a large buffer is not best. Too small a buffer reduces link utilisaiton. If it is necessary to find a compromise size for adjustable line rates, should consider sacrificing some utilisation at lower rates to keep the buffer delay reasonable.

3.2. Buffering recommendations for end hosts

XX Work in Progress, to be included in next revision XX

Large buffers are not best. AQM and auto-tuning/auto-sizing are as applicable in end hosts as in network equipment.

ECN may even be appropriate (e.g. on a subsystem such as a NIC), but within a host it should be possible to use back-pressure messages instead.

Buffer sizing recommendations specific to end-systems.

3.3. Buffering recommendations for edge routers and switches

XX Work in Progress, to be included in next revision XX

Large is not best.

AQM and ECN are strongly recommended.

Buffer sizing recommendations specific to edge routers, switches & middleboxes.

3.4. Buffering recommendations for core routers and switches

XX Work in Progress, to be included in next revision XX

Large is not best.

Buffer sizing recommendations specific to core routers & switches.

3.5. Recommendations on Flow Isolation

XX Work in Progress, to be included in next revision XX

Still a subject of debate and research. May be able to recommend something here, but more likely will commentate on the debate.

4. Buffer Management Methods

This section provides informative documentation of current practice.

4.1. Examples of subnetwork buffering

This section provides informative examples of buffer configuration and their impact on network traffic {TBA: to consider whether to bless, deprecate or merely state each of these practices}.

- An Ethernet subnetwork may operate over a range of speeds from a shared 10 Mbps of capacity to over 40 Gbps. The buffering required depends on the link speed and many Many device drivers and operating systems do not adjust their buffering to the available capacity. The first hop link from a host often has a higher speed than the subsequent links along a network path.

- Subnetwork flow-control can be triggered when a subnetwork link suffers congestion. An example is the use of Ethernet Pause frames (e.g. by consumer Ethernet switches) to slow a sender emitting traffic towards a congestion switch port. These methods can increase the buffering experienced by the end-to-end flow.

- Docsis 3.1 supports transmission up to 300Mbps. A current modem can be plugged into a current network. Then suppose a customers service only supports 10 Mbps, the network equipment may be 30 times over-buffered (assuming buffers are dimensioned based on the maximum bit rate). The buffer control amendment may be implemented in the modem, and in its provisioning system to address this type of issue. Similar issues apply for other link

technologies, were the offered service is often less than the maximum supported rate.

- On wireless, bandwidth (and hence network capacity) is often highly variable, unless you have a fixed point to point link. Even fixed links may use adaptive methods and propagation conditions can cause the capacity to var

4.2. Examples of methods for active buffer management

This section provides informative examples of active buffer management.

While large buffers can lead to an increase in experienced network delay, they do not necessarily impact the flow delay. The issue is not how how much buffering is provided, but how the provided buffers are used to manage the flow of traffic.

Several active buffer/queue management methods have been proposed that can significantly improve performance of flows using a (potentially) congested bottleneck.

- RED
- CoDel
- Pi
- etc

5. Security Considerations

Decisions on queue management and buffer sizing are neutral to security considerations if they act indiscriminately over all packets. Recommendations on treatment or lack of treatment at the flow or user-level can have security considerations, which are TBA.

The question of whether end-systems respond to congestion signals is a valid security concern, but outside the scope of this document.

6. IANA Considerations

This document does not require any IANA considerations.

[RFC-ED]: Please remove this section prior to publication.

7. Acknowledgments

This work was part-funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700). The views expressed are solely those of the author.

The authors acknowledge contributions from: Jim Gettys.

8. References

8.1. Normative References

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

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

[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.

8.2. Informative References

[Appenzeller] Appenzeller, G., Keslassy, I., and N. McKeown, "Sizing router buffers; ACM SIGCOMM '04, pages 281-292, New York, NY, USA.", 2004.

[Ganjali] Ganjali, Y. and N. McKeown, "Update on Buffer Sizing in Internet Routers; ACM SIGCOMM Computer Communication Review 36 ACM", October 2006.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

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

[Villamizar] Villamizar, C. and C. Song, "High Performance TCP in ANSNET; ACM Computer Communications Review, 24(5):45-60", 1994.

[Wischik] , "TCP Buffer Sizing Advice", .

Appendix A. vious IETF guidance for configuring network buffers

This section reviews previous guidance for configuring network buffers and motivates the need to update these recommendations.

Guidance for the use of buffers was provided in section 13 of RFC 3819:

"each node should have enough buffering to hold one link_bandwidth*link_delay product's worth of data for each TCP connection sharing the link."

However, in today's Internet, a deployment following this recommendation would overly allocate buffering for a network link that supports multiple flows. This is discussed in the observations below:

- This buffering recommendation is appropriate for a device that supports a single or small number of bulk TCP flows [Villamizar].

- The buffering is unduly large when there are more than a small number of flows (e.g. >10). The goal of sharing between TCP flows requires only that the buffering is sufficient to hold one link_bandwidth*path_delay product's worth of data for the longest path flow. The more flows share a link, the less buffering is needed [Appenzeller], unless the egress link becomes congested with so many flows that there are only a few packets per flow buffered.

- Many egress links have a higher level of multiplexing (e.g. >100 of uncorrelated flows). This is often found beyond the edge of a network. In this case, the buffer size may be inversely proportional to the square root of the number of flows (for medium numbers . For still higher levels of multiplexing, this may be of the order of the logarithm of the number of flows [Wischik][Ganjali].

- Note that while optimal buffering may be a function of the number of concurrent flows, it is not recommended to tune buffering by dynamically estimating the number of flows sharing a network device or path, or by attempting to classify flows as "long", "short", etc. Such estimates are difficult, due to the wide variety of flow behaviours and the use of aggregation methods (such as tunnels) that hide the traffic of individual flows.

- In deployed scenarios (apart from restricted deployments in operator-controlled subnetworks), it is usually impossible for a router or other network middlebox to know the experienced by a

flow. In the Internet service model this information is only available to end points (e.g. using feedback provided by TCP [RFC0793] or RTCP [RFC3550]. It is therefore not usually possibly for operators to use the end-to-end path delay calculation to determine the size of buffering when configuring network equipment.

The discussion in section 13 of RFC 3819 summarises:

"In general, it is wise to err in favor of too much buffering rather than too little."

While this advice may have been appropriate when routers and subnetworks with small numbers of flows and low buffer memory [Villamizar], this advice is now not appropriate for many modern networks.

Section 13 of RFC 3819 also motivates using methods such as Active Queue Management, AQM and [RFC3168]. However, at the time of writing there was little deployment experience, and little understanding of how to configure these methods. We now argue that these methods should be considered for deployment in operational networks.

Appendix B. Revision notes

RFC-Editor: Please remove this section prior to publication

Draft 00

o This contains the first draft for comment.

Authors' Addresses

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

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

Bob Briscoe
BT
B54/77, Adastral Park
Martlesham Heath, Ipswich IP5 3RE
UK

Phone: +44 1473 645196
Email: bob.briscoe@bt.com
URI: http://bobbriscoe.net/