Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: June 6, 2014

J. Jaeggli
Zynga
L. Colitti
W. Kumari
Google
E. Vyncke
Cisco
M. Kaeo
Double Shot Security
T. Taylor, Ed.
Huawei Technologies

December 3, 2013

Why Operators Filter Fragments and What It Implies
draft-taylor-v6ops-fragdrop-02

Abstract

This memo was written to make application developers and network operators aware of the significant possibility that IPv6 packets containing fragmentation extension headers may fail to reach their destination. Some protocol or application assumptions about the ability to use messages larger than a single packet may accordingly not be supportable in all networks or circumstances.

This memo provides observational evidence for the dropping of IPv6 fragments along a significant number of paths, explores the operational impact of fragmentation and the reasons and scenarios where drops occur, and considers the effect of fragment drops on applications where fragmentation is known to occur, particularly including DNS.

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 June 6, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Measurements of whether Internet Service Providers and edge networks deliver IPv6 fragments to their destination reveal that for IPv6 in particular, fragments are being dropped along a substantial number of paths. The filtering of IPv6 datagrams with fragmentation headers is presumed to be a non-issue in the core of the Internet, where fragments are routed just like any other IPv6 datagram. However, fragmentation can creates operational issues at the edges of the Internet that may lead to administratively imposed filtering or inadvertent failure to deliver the fragment to the end-system or application.

Section 2 begins with some observations on how often IPv6 fragment loss occurs in practice. We go on to look at the operational reasons for filtering fragments, a key aspect of which is the limitations they expose in the application of security policy, at resource bottlenecks and in forwarding decisions. Section 2.2 then looks at the impact on key applications, particularly DNS.

In the longer run, as network operators gain a better understanding of the risks and non-risks of fragmentation and as middlebox, customer premise equipment (CPE), and host implementations improve, we believe that some incidence of fragment dropping currently required will diminish. Some of the justifications for filtering will persist in the long-term, and application developers and network operators must remain aware of the implications.

This document deliberately refrains from discussing possible responses to the problem posed by the dropping of IPv6 fragments. Such a discussion will quickly turn up a number of possibilities, application-specific or more general; but the amount of time needed to specify and deploy a given resolution will be a major constraint in choosing amongst them. In any event, that discussion is likely to proceed in multiple directions, occur in different areas and is therefore considered beyond the scope of this memo.

2. Observations and Rationale

[Blackhole] is a good public reference for some empirical data on IPv6 fragment filtering. It describes experiments run to determine the incidence and location of ICMP Packet Too Big and fragment filtering. The authors used fragmented DNS packets to determine the latter, setting the servers to an IPv6 minimum of 1280 bytes to avoid any PMTU issues. The tests found for IPv6 that filtering appeared to be occurring on some 10% of the tested paths. The filtering appeared to be located at the edge (enterprise and customer networks) rather than in the core.

2.1. Possible Causes

Why does such filtering happen? One cause is non-conforming implementations in CPE and low-end routers. Some network managers filter fragments on principle, thinking this is an easier way to deter realizable attacks utilizing IPv6 fragments without thinking of other network impacts, similar to the practice of filtering ICMP Packet Too Big. Both implementations and management should improve over time, reducing the problem somewhat.

Some filtering and dropping of fragments is known to be done for hardware, performance, or topological considerations.

2.1.1. Stateful inspection

Stateful inspection devices or destination hosts can readily experience resource exhaustion if they are flooded with fragments that are not followed in a timely manner by the remaining fragments of the original datagram. Holding fragments for reassembly even on end-system firewalls can readily result in an effective denial of service by memory and CPU exhaustion even if techniques, such as virtual re-assambly exist.

2.1.2. Stateless ACLs

Stateless ACLs at layer 4 and up may be difficult to apply to fragments other than the first one in which enough of the upper layer header is present. As [Attacks] demonstrates, inconsistencies in reassembly logic between middleboxes or CPEs and hosts can cause fragments to be wrongfully discarded, or can allow exploits to pass undetected through middleboxes. Stateless load balancing schemes may hash fragmented datagrams from the same flow to different paths because the 5-tuple may be available on only the initial fragment. While rehashing has the possibility of reordering packets in ISP cores it is not disastrous. However, in front of a stateful inspection device, load balancer tier, or anycast service instance, where headers other than the L3 header -- for example, the L4 header, interface index (for traffic already rehashed onto different paths), DS fields -- are considered as part of the hash, rehashing may result in the fragments being delivered to different end-systems

2.1.3. Performance considerations

Leaving aside these incentives towards fragment dropping, other considerations may weigh on the operator’s mind. One example cited on the NANOG list was that of a router where fragment processing was done by the control plane processor rather than in the forwarding plane hardware, with a consequent hit on performance.

2.1.4. Other considerations

Another incentive toward dropping of fragments is the disproportionate number of software errors still being encountered in fragment processing. Since this code is exercised less frequently than the rest of the stack, bugs remain longer in the code before they are detected. Some of these software errors can introduce vulnerabilities subject to exploitation. It is common practice [RFC6192] to recommend that control-plane ACLs protecting routers and network devices be configured to drop all fragments.

2.1.5. Conclusions

Operators weigh the risks associated with each of the considerations just enumerated, and come up with the most suitable policy for their circumstances. It is likely that at least some operators will find it desirable to drop fragments in at least some cases.

The IETF and operators can help this effort by identifying specific classes of fragments that do not represent legitimate use cases and hence should always be dropped. Examples of this work are given by [RFC6946] and [I-D.ietf-6man-oversized-header-chain]. The problem of inconsistent implementations may also be mitigated by providing further advice on the more difficult points. However, some cases will remain where legitimate fragments are discarded for legitimate reasons. The potential problems these cases pose for applications is our next topic.

2.2. Impact on Applications

Some applications can live without fragmentation, some cannot. UDP DNS is one application that has the potential to be impacted when fragment dropping occurs. EDNS0 extensions [RFC2671] allow for responses in UDP PDUs that are greater than 512 bytes. Particularly with DNSSEC [RFC4033], responses may be larger than the link MTU and fragmentation would therefore occur at the sending host in order to respond using UDP. The current choices open to the operators of DNS servers in this situation are to defer deployment of DNSSEC, fragment responses, or use TCP if there are cases where the rrset would be expected to exceed the MTU. The use of fallback to TCP will impose a major resource and performance hit and increases vulnerability to denial of service attacks.

Other applications, such as the Network File System, NFS, are also known to fragment large UDP packets for datagrams larger than the MTU. NFS is most often restricted to the internal networks of organizations. In general, managing NFS connectivity should not be impacted by decisions mananging fragment drops at network borders or end-systems.

3. Acknowledgements

The authors of this document would like to thank the RIPE Atlas project and NLNetlabs whose conclusions ignited this document.

4. IANA Considerations

This memo includes no request to IANA.

5. Security Considerations

The potential for denial of service attacks, as well as limitations inherent in upper-layer filtering when dealing with non-initial fragments are significant issues under consideration by operators and end-users filtering fragments. This document does not offer alternative solutions to that problem, it does describe the impact of those filtering practices.

6. Informative References

[Attacks] Atlasis, A., "Attacking IPv6 Implementation Using Fragmentation", March 2012.

http://media.blackhat.com/bh-eu-12/Atlasis/bh-eu-12-Atlasis-Attacking_IPv6-WP.pdf

[Blackhole] de Boer, M. and J. Bosma, "Discovering Path MTU black holes on the Internet using RIPE Atlas", July 2012.

http://www.nlnetlabs.nl/downloads/publications/pmtu-black-holes-msc-thesis.pdf

[I-D.ietf-6man-oversized-header-chain] Gont, F., Manral, V., and R. Bonica, "Implications of Oversized IPv6 Header Chains", draft-ietf-6man-oversized-header-chain-08 (work in progress), October 2013.

[RFC2671] Vixie, P., "Extension Mechanisms for DNS (EDNS0)", RFC 2671, August 1999.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[RFC6192] Dugal, D., Pignataro, C., and R. Dunn, "Protecting the Router Control Plane", RFC 6192, March 2011.

[RFC6946] Gont, F., "Processing of IPv6 "Atomic" Fragments", RFC 6946, May 2013.

Authors’ Addresses

Joel Jaeggli
Zynga
630 taylor ct #10
Mountain View, CA 94043
USA

Email: jjaeggli@zynga.com

Lorenzo Colitti
Google

Email: lorenzo@google.com

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: warren@kumari.net

Eric Vyncke
Cisco
De Kleetlaan 6A
Diegem 1831
Belgium

Email: evyncke@cisco.com

Merike Kaeo
Double Shot Security

Email: merike@doubleshotsecurity.com

Tom Taylor (editor)
Huawei Technologies
Ottawa, Ontario
Canada

Email: tom.taylor.stds@gmail.com