Workgroup: Network Working Group
Internet-Draft: draft-templin-dtn-ltpfrag-05
Published: 11 October 2021
Intended Status: Informational
Expires: 14 April 2022
Authors: F. L. Templin, Ed.
Boeing Research & Technology

LTP Fragmentation

Abstract

The Licklider Transmission Protocol (LTP) provides a reliable datagram convergence layer for the Delay/Disruption Tolerant Networking (DTN) Bundle Protocol. In common practice, LTP is often configured over UDP/IP sockets and inherits its maximum segment size from the maximum-sized UDP datagram, however when this size exceeds the maximum IP packet size for the path a service known as IP fragmentation must be employed. This document discusses LTP interactions with IP fragmentation and mitigations for managing the amount of IP fragmentation employed.

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 https://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 14 April 2022.

Table of Contents

1. Introduction

The Licklider Transmission Protocol (LTP) [RFC5326] provides a reliable datagram convergence layer for the Delay/Disruption Tolerant Networking (DTN) Bundle Protocol (BP) [I-D.ietf-dtn-bpbis]. In common practice, LTP is often configured over the User Datagram Protocol (UDP) [RFC0768] and Internet Protocol (IP) [RFC0791] using the "socket" abstraction. LTP inherits its maximum segment size from the maximum-sized UDP datagram (i.e. 2**16 bytes minus header sizes), however when the UDP datagram size exceeds the maximum IP packet size for the path a service known as IP fragmentation must be employed.

LTP breaks BP bundles into "blocks", then further breaks these blocks into "segments". The segment size is a configurable option and represents the largest atomic portion of data that LTP will require underlying layers to deliver as a single unit. The segment size is therefore also known as the "retransmission unit", since each lost segment must be retransmitted in its entirety. Experimental and operational evidence has shown that on robust networks increasing the LTP segment size (up to the maximum UDP datagram size of slightly less than 64KB) can result in substantial performance increases over smaller segment sizes. However, the performance increases must be tempered with the amount of IP fragmentation invoked as discussed below.

When LTP presents a segment to the operating system kernel (e.g., via a sendmsg() system call), the UDP layer prepends a UDP header to create a UDP datagram. The UDP layer then presents the resulting datagram to the IP layer for packet framing and transmission over a networked path. The path is further characterized by the path Maximum Transmission Unit (Path-MTU) which is a measure of the smallest link MTU (Link-MTU) among all links in the path.

When LTP presents a segment to the kernel that is larger than the Path-MTU, the resulting UDP datagram is presented to the IP layer, which in turn performs IP fragmentation to break the datagram into fragments that are no larger than the Path-MTU. For example, if the LTP segment size is 64KB and the Path-MTU is 1280 bytes IP fragmentation results in 50+ fragments that are transmitted as individual IP packets. (Note that for IPv4 [RFC0791], fragmentation may occur either in the source host or in a router in the network path, while for IPv6 [RFC8200] only the source host may perform fragmentation.)

Each IP fragment is subject to the same best-effort delivery service offered by the network according to current congestion and/or link signal quality conditions; therefore, the IP fragment size becomes known as the "loss unit". Especially when the packet loss rate is non-negligible, however, performance can suffer dramatically when the loss unit is significantly smaller than the retransmission unit. In particular, if even a single IP fragment of a fragmented LTP segment is lost then the entire LTP segment is deemed lost and must be retransmitted. Since LTP does not support flow control or congestion control, this can result in catastrophic communication failure when fragments are systematically lost in transit.

This document discusses LTP interactions with IP fragmentation and mitigations for managing the amount of IP fragmentation employed. It further discusses methods for increasing LTP performance both with and without the aid of IP fragmentation.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119][RFC8174] when, and only when, they appear in all capitals, as shown here.

3. IP Fragmentation Issues

IP fragmentation is a fundamental service of the Internet Protocol, yet it has long been understood that its use can be problematic in some environments. Beginning as early as 1987, "Fragmentation

Considered Harmful" [FRAG] outlined multiple issues with the service including a performance-crippling condition that can occur at high data rates when the loss unit is considerably smaller than the retransmission unit during intermittent and/or steady-state loss conditions.

Later investigations also identified the possibility for undetected data corruption at high data rates due to a condition known as "ID wraparound" when the 16-bit IP identification field (aka the "IP ID") increments such that new fragments overlap with existing fragments still alive in the network and with identical ID values [RFC4963][RFC6864]. Although this issue occurs only in the IPv4 protocol (and not in IPv6 where the IP ID is 32-bits in length), the IPv4 concerns along with the fact that IPv6 does not permit routers to perform "network fragmentation" have led many to discourage its use.

Even in the modern era, investigators have seen fit to declare "IP Fragmentation Considered Fragile" in an Internet Engineering Task Force (IETF) Best Current Practice (BCP) reference [RFC8900]. Indeed, the BCP recommendations cite the Bundle Protocol LTP convergence layer as a user of IP fragmentation that depends on some of its properties to realize greater performance. However, the BCP summarizes by saying:

> "Rather than deprecating IP fragmentation, this document recommends that upper-layer protocols address the problem of fragmentation at their layer, reducing their reliance on IP fragmentation to the greatest degree possible."

While the performance implications are considerable and have serious implications for real-world applications, our goal in this document is neither to condemn nor embrace IP fragmentation as it pertains to the Bundle Protocol LTP convergence layer operating over UDP/IP sockets. Instead, we examine ways in which the benefits of IP fragmentation can be realized while avoiding the pitfalls. We therefore next discuss our systematic approach to LTP fragmentation.

4. LTP Fragmentation

In common LTP implementations over UDP/IP (e.g., the Interplanetary Overlay Network (ION)), performance is greatly dependent on the LTP segment size. This is due to the fact that a larger segment presented to UDP/IP as a single unit incurs only a single system call and a single data copy from application to kernel space via the sendmsg() system call. Once inside the kernel, the segment incurs UDP/IP encapsulation and IP fragmentation which again results in a loss unit smaller than the retransmission unit. However, during fragmentation, each fragment is transmitted immediately following

the previous without delay so that the fragments appear as a "burst" of consecutive packets over the network path resulting in high network utilization during the burst period. Additionally, the use of IP fragmentation with a larger segment size conserves header framing bytes since the LTP layer headers only appear in the first IP fragment as opposed to appearing in all IP packets.

In order to avoid retransmission congestion (i.e., especially when the loss probability is non-negligible), the natural choice would be to set the LTP segment size to a size that is no larger than the Path-MTU. Assuming the minimum IPv4 MTU of 576 bytes, however, transmission of 64KB of data using a 576B segment size would require well over 100 independent sendmsg() system calls and data copies as opposed to just one when the largest segment size is used. This greatly reduces the bandwidth advantage offered by IP fragmentation bursts. Therefore, a means for providing the best aspects of both large segment fragment bursting and small segment retransmission efficiency is needed.

Common operating systems such as linux provide the sendmmsg() ("send multiple messages") system call that allows the LTP application to present the kernel with a vector of up to 1024 segments instead of just a single segment. This affords the bursting behavior of IP fragmentation coupled with the retransmission efficiency of employing small segment sizes. (Note that LTP receivers can also use the recvmmsg() ("receive multiple messages") system call to receive a vector of segments from the kernel in case multiple recent packet arrivals can be combined.)

This work therefore recommends implementations of LTP to employ a large block size, a conservative segment size and a new configuration option known as the "Burst-Limit" which determines the number of segments that can be presented in a single sendmmsg() system call. When the implementation receives an LTP block, it carves Burst-Limit-many segments from the block and presents the vector of segments to sendmmsg(). The kernel will prepare each segment as an independent UDP/IP packet and transmit them into the network as a burst in a fashion that parallels IP fragmentation. The loss unit and retransmission unit will be the same, therefore loss of a single segment does not result in a retransmission congestion event.

It should be noted that the Burst-Limit is bounded only by the LTP block size and not by the maximum UDP datagram size. Therefore, each burst can in practice convey significantly more data than a single IP fragmentation event. It should also be noted that the segment size can still be made larger than the Path-MTU in low-loss environments without danger of triggering retransmission storms due to loss of IP fragments. This would result in combined UDP message

and IP fragment bursting for increased network utilization in more robust environments. Finally, both the Burst-Limit and UDP message sizes need not be static values, and can be tuned to adaptively increase or decrease according to time varying network conditions.

5. Beyond "sendmmsg()"

Implementation experience with the ION DTN distribution along with two recent studies have demonstrated modest performance increases for employing sendmmsg() for transmission over UDP/IP sockets. A first study used sendmmsg() as part of an integrated solution to produce 1M packets per second assuming only raw data transmission conditions [MPPS], while a second study focused on performance improvements for the QUIC reliable transport service [QUIC]. In both studies, the use of sendmmsg() alone produced observable increases but complimentary enhancements were identified that (when combined with sendmmsg()) produced considerable additional increases.

In [MPPS], additional enhancements such as using recvmmsg() and configuring multiple receive queues at the receiver were introduced in an attempt to achieve greater parallelism and engage multiple processors and threads. However, the system was still limited to a single thread until multiple receiving processes were introduced using the "SO_REUSEPORT" socket option. By having multiple receiving processes (each with its own socket buffer), the performance advantages of parallel processing were employed to achieve the 1M packets per second goal.

In [QUIC], a new feature available in recent linux kernel versions was employed. The feature, known as "Generic Segmentation Offload (GSO) / Generic Receive Offload (GRO)" allows an application to provide the kernel with a "super-buffer" containing up to 64 separate upper layer protocol segments. When the application presents the super-buffer to the kernel, GSO segmentation then sends 64 separate UDP/IP packets in a burst. If each packet is larger than the Path-MTU, then IP fragmentation will be invoked for each packet leading to high network utilization (at the risk of IP fragment loss and retransmission storms). The GSO facility can be invoked by either sendmsg() (i.e., a single super-buffer) or sendmmsg() (i.e., multiple super-buffers), and the study showed a substantial performance increase over using just sendmsg() and sendmmsg() alone.

For LTP fragmentation, our ongoing efforts explore using these techniques in a manner that parallels the effort undertaken for QUIC. Using these higher-layer segmentation management facilities is

consistent with the guidance in "IP Fragmentation Considered Fragile" that states:

> "Rather than deprecating IP fragmentation, this document recommends that upper-layer protocols address the problem of fragmentation at their layer, reducing their reliance on IP fragmentation to the greatest degree possible."

By addressing fragmentation at their layer, the LTP/UDP functions can then be tuned to minimize IP fragmentation in environments where it may be problematic or to adaptively engage IP fragmentation in environments where performance gains can be realized without risking data corruption.

6. LTP Performance Enhancement Using GSO/GRO

Some modern operating systems include Generic Segment Offload (GSO) and Generic Receive Offload (GRO) services. For example, GSO/GRO support has been included in linux beginning with kernel version 4.18. Some network drivers and network hardware also support GSO/GRO at or below the operating system network-to-driver interface layer to provide benefits of delayed segmentation and/or early reassembly. The following sections discuss LTP interactions with GSO and GRO.

6.1. LTP and GSO

GSO allows LTP implementations to present the sendmsg() or sendmmsg() system calls with "super-buffers" that include up to 64 LTP segments which the kernel will subdivide into individual UDP datagrams. LTP implementations enable GSO on a per-socket basis using the "setsockopt()" system call as follows:

```
unsigned integer gso_size = SEGSIZE;
setsockopt(fd, SOL_UDP, UDP_SEGMENT, &gso_size, sizeof(gso_size)));
```

Implementations must set SEGSIZE to an initial value no larger than the MTU of the underlying network interface minus the UDP and IP header sizes; this ensures that UDP datagrams generated during GSO segmentation will not incur local IP fragmentation prior to transmission (NB: the linux kernel returns EINVAL if SEGSIZE is set to a value that would exceed the MTU of the underlying interface). For paths that traverse multiple links, implementations should also dynamically adjust SEGSIZE according to the per-destination path MTU to avoid sustained in-the-network fragmentation resulting in a loss unit smaller than the retransmission unit.

Implementations should therefore dynamically determine SEGSIZE for paths that traverse multiple links through Packetization Layer Path MTU Discovery for Datagram Transports [RFC8899] (DPMTUD). For IPv4 paths, implementations may initially set SEGSIZE according to the

MTU of the underlying interface and invoke DPMTUD while initial packets are flowing, then should dynamically reduce SEGSIZE without service interruption if the discovered path MTU is smaller. For IPv6 paths, implementations should initially set SEGSIZE according to the minimum IPv6 Path MTU (i.e., 1280) then may dynamically increase SEGSIZE without service interruption if the discovered path MTU is larger.

6.2. LTP and GRO

GRO allows the kernel to return "super-buffers" that contain multiple concatenated received segments to the LTP implementation in recvmsg() or recvmmsg() system calls, where each concatenated segment is distinguished by an LTP segment header per [RFC5326]. LTP implementations enable GRO on a per-socket basis using the "setsockopt()" system call as follows:

```
unsigned integer gro_size = 0;
setsockopt(fd, SOL_UDP, UDP_GRO, &gro_size, sizeof(gro_size)));
```

Implementations include a pointer to a gro_size variable as a boolean indication to the kernel using any arbitrary initialization value (e.g., '0'), as GRO will accept received segments of any size; the only interoperability requirement therefore is that each UDP packet includes one or more properly-formed LTP segments. The kernel and/or underlying network hardware will first coalesce multiple received segments into a larger single segment whenever possible and/or return multiple coalesced or singular segments to the LTP implementation so as to maximize the amount of data returned in a single system call.

Implementations that invoke recvmsg( ) and/or recvmmsg() will therefore receive "super-buffers" that include one or more concatenated received LTP segments. The LTP implementation accepts all received LTP segments and identifies any segments that may be missing. The LTP protocol then engages segment report procedures if necessary to request retransmission of any missing segments.

6.3. LTP GSO/GRO Over OMNI Interfaces

LTP engines produce UDP/IP packets that can be forwarded over an underlying network interface as the head-end of a "link-layer service that transits IP packets". UDP/IP packets that enter the link near-end are deterministically delivered to the link-far end modulo loss due to corruption, congestion or disruption. The link-layer service is associated with an MTU that deterministically establishes the maximum packet size that can transit the link. The link-layer service may further support a segmentation and reassembly function with fragment retransmissions at a layer below IP; in many

cases, these timely link-layer retransmissions can reduce dependency on (slow) end-to-end retransmissions.

LTP engines that connect to networks traversed by paths consisting of multiple concatenated links must be prepared to adapt their segment sizes to match the minimum MTU of all links in the path. This could result in a small SEGSIZE that would interfere with the benefits of GSO/GRO layering. However, nodes that configure LTP engines can establish an Overlay Multilink Network Interface (OMNI) [I-D.templin-6man-omni] that spans the multiple concatenated links while presenting an assured 9180 byte MTU to the LTP engine.

The OMNI interface internally uses IP fragmentation as a link-layer adaptation service not visible to the LTP engine, including timely link-layer retransmissions of lost fragments where the retransmission unit matches the loss unit. The LTP engine can then dynamically vary its SEGSIZE (up to a maximum value of 9180 bytes) to determine the size that produces the best performance given the combined operational factors at all layers of the multi-layer architecture. This dynamic factoring coupled with the ideal link properties provided by the OMNI interface support an effective layering solution for many DTN networks.

7. Implementation Status

Supporting code for invoking the sendmmsg() facility is included in the official ION source code distribution, beginning with release ion-4.0.1.

8. IANA Considerations

This document introduces no IANA considerations.

9. Security Considerations

Communications networking security is necessary to preserve confidentiality, integrity and availability.

10. Acknowledgements

The NASA Space Communications and Networks (SCaN) directorate coordinates DTN activities for the International Space Station (ISS) and other space exploration initiatives.

Madhuri Madhava Badgandi, Keith Philpott, Bill Pohlchuck, Vijayasarathy Rajagopalan and Eric Yeh are acknowledged for their significant contributions. Tyler Doubrava was the first to mention the "sendmmsg()" facility. Scott Burleigh provided review input, and David Zoller provided useful perspective.

11. References

11.1. Normative References

[RFC0768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <https://www.rfc-editor.org/info/rfc768>.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, DOI 10.17487/RFC0791, September 1981, <https://www.rfc-editor.org/info/rfc791>.

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

[RFC5326] Ramadas, M., Burleigh, S., and S. Farrell, "Licklider Transmission Protocol - Specification", RFC 5326, DOI 10.17487/RFC5326, September 2008, <https://www.rfc-editor.org/info/rfc5326>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

11.2. Informative References

[FRAG] Mogul, J. and C. Kent, "Fragmentation Considered Harmful, ACM Sigcomm 1987", August 1987.

[I-D.ietf-dtn-bpbis] Burleigh, S., Fall, K., and E. J. Birrane, "Bundle Protocol Version 7", Work in Progress, Internet-Draft, draft-ietf-dtn-bpbis-31, 25 January 2021, <https://www.ietf.org/archive/id/draft-ietf-dtn-bpbis-31.txt>.

[I-D.templin-6man-omni] Templin, F. L. and T. Whyman, "Transmission of IP Packets over Overlay Multilink Network (OMNI) Interfaces", Work in Progress, Internet-Draft, draft-

templin-6man-omni-47, 8 September 2021, <https://www.ietf.org/archive/id/draft-templin-6man-omni-47.txt>.

[MPPS] Majkowski, M., "How to Receive a Million Packets Per Second, https://blog.cloudflare.com/how-to-receive-a-million-packets/", June 2015.

[QUIC] Ghedini, A., "Accelerating UDP Packet Transmission for QUIC, https://calendar.perfplanet.com/2019/accelerating-udp-packet-transmission-for-quic/", December 2019.

[RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, DOI 10.17487/RFC4963, July 2007, <https://www.rfc-editor.org/info/rfc4963>.

[RFC6864] Touch, J., "Updated Specification of the IPv4 ID Field", RFC 6864, DOI 10.17487/RFC6864, February 2013, <https://www.rfc-editor.org/info/rfc6864>.

[RFC8899] Fairhurst, G., Jones, T., Tüxen, M., Rüngeler, I., and T. Völker, "Packetization Layer Path MTU Discovery for Datagram Transports", RFC 8899, DOI 10.17487/RFC8899, September 2020, <https://www.rfc-editor.org/info/rfc8899>.

[RFC8900] Bonica, R., Baker, F., Huston, G., Hinden, R., Troan, O., and F. Gont, "IP Fragmentation Considered Fragile", BCP 230, RFC 8900, DOI 10.17487/RFC8900, September 2020, <https://www.rfc-editor.org/info/rfc8900>.

Author's Address

Fred L. Templin (editor)
Boeing Research & Technology
P.O. Box 3707
Seattle, WA 98124
United States of America

Email: fltemplin@acm.org