Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 17, 2012

M. Levy
Hurricane Electric

November 14, 2011

Jumbo Frame Deployment at Internet Exchange Points (IXPs)
draft-mlevy-ixp-jumboframes-00.txt

Abstract

This document provides guidelines on how to deploy Jumbo Frame support on Internet Exchange Points (IXP). Jumbo Frame support allows packets larger than 1,500 Bytes to be passed between IXP customers over the IXPs layer 2 fabric. This document describes methods to enable Jumbo Frame support and keep in place existing 1,500 Byte communications.

This document strongly recommends that IXP operators choose 9,000 Bytes for their Jumbo Frame implementation.

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 May 17, 2012.

Copyright Notice

Table of Contents

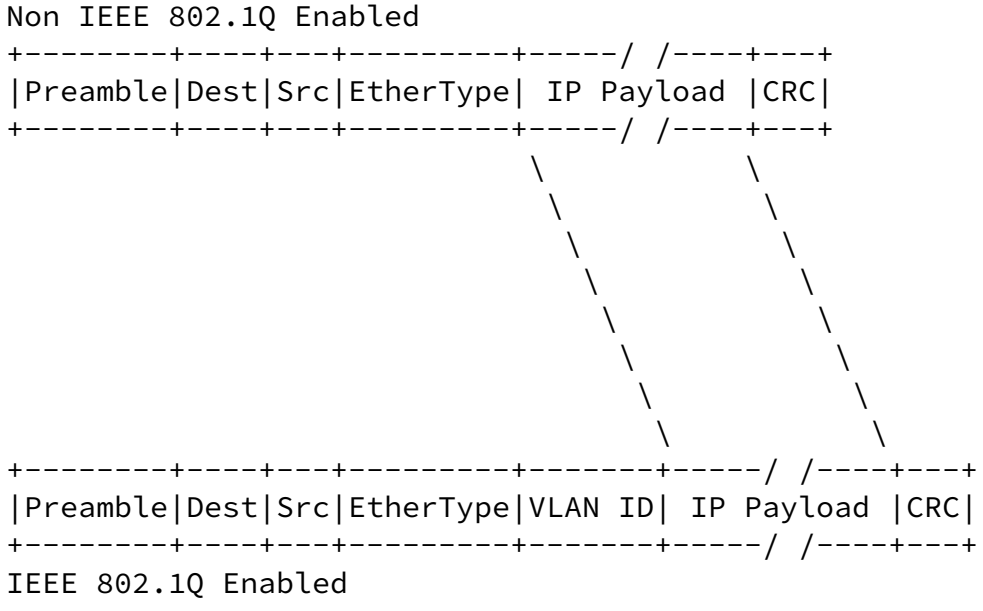
1. Introduction

The standard Maximum Transmission Unit (MTU) value, for IP packets encapsulated within an Ethernet frame, is 1,500 Bytes. This is described in RFC 894 [RFC894] and RFC 1042 [RFC1042].

The specific size of a Jumbo Frame is not defined by the IEEE. Many sizes can be chosen depending on the hardware vendor or hardware platform. This document strongly recommends that IXP operators choose 9,000 Bytes for their Jumbo Frame implementation.

1.1. Defining MTU values

All MTU sizes, including the default 1,500 Byte size, refers to the IP packet/payload size vs. the full Ethernet frame size. The standard Ethernet frame size is 1,514 Bytes (1,500 + 6 + 6 + 2) or 1,518 (1,500 + 6 + 6 + 4 + 2) Bytes depending on the use of IEEE 802.1Q (VLAN) tags [IEEE802_1Q]. The Preamble and CRC lengths are not used in the count.



All sizes listed within this document references the IP payload portion of the Ethernet frame only.

1.2. Jumbo Frames

Jumbo Frames are considered to be Ethernet frames that can carry an IP payload greater than 1,500 Bytes [MATHIS2002] [SAUVER2003]. Jumbo Frames are sometimes called "Giant Jumbo", "Mini Jumbo" or "Baby Jumbo" [TULYU2011]. This document recommends the use of the wording "Jumbo Frame" as the terminology within the IXP industry. This document only uses the wording "Jumbo Frame" to represent a frame

capable of transporting a payload above 1,500 Bytes MTU.

If customers require end-to-end Jumbo Frame support and an IXP within the path only provides 1,500 Byte MTU connections, then the end-to-end provided Path MTU (PMTU) can only be 1,500 Bytes. This document recommends ways for IXP operators to provide networks with Jumbo Frame support and potentially allowing larger end-to-end PMTU.

Additional protocols that exceed 1,500 Byte MTU are "FCoE", "iSCSI", "MPLS", "IEEE 802.1AS", "IEEE 802.3AE", etc. None are applicable to the IXP industry.

1.3. IXPs

An Internet Exchange Points (IXP) is a layer 2 service allowing one network to communicate with one or more networks over a shared fabric. These days an IXP is normally built using high availability Ethernet switches and historically provided the IEEE defined default Ethernet Maximum Transmission Unit (MTU) size of 1,500 Bytes for each port.

As the Internet has grown, both in geography and speed, IXPs has mainly stuck to 1,500 Byte MTU size. A study done in 2008 of the peering community showed interest in larger MTU peering [HANKINS2008].

IXP	Location	Provided MTU	Comments

AMS-IX	Amsterdam, NL	1,500	Untagged ports
Any2	US	1,500	Untagged ports
DE-CIX	Frankfurt, DE	1,500	Untagged ports
Equinix	US & others	1,500	Untagged ports
HKIX	Hong Kong, HK	1,500	Untagged ports
JPIX	Tokyo, JP	1,500	Untagged ports
JPNAP	Tokyo, JP	1,500	Untagged ports
LINX	London, UK	1,500	Untagged ports
NASA-AIX	Palo Alto, US	1,500 & 9,000	Two VLANs on request
NETNOD	Stockholm, SE	1,500 & 4,470	Two VLANs by default
Telx TIE	US	1,500	Untagged ports

Table 1: IXP MTU sizes

There is no extensive study of IXP operators and MTU values. This is just a minimal review to show it exists.

1.4. IP Backbones

Some IP backbones have implemented larger MTU sizes on backbone links [NANOG2008]; however, it's safe to say that nearly every broadband user is connected at 1,500 Byte MTU size, or less. Broadband or dialup connections using PPPoE are configured at 1,492 Bytes. See RFC 2516 [RFC2516].

The same limitation of 1,500 Bytes can be said for most sources of content. (CITATION NEEDED)

Allowing end-to-end system to communicate with larger MTUs can reduce end-system CPU usage, provide less per-packet overhead and improve TCP performance [NANOG2003] [Internet2_LSR]. Applications that do mass data transfer (backups, replication, NNTP, etc) benefit from larger MTU paths.

VPNs that require MTU sizes of 1,500 Bytes could use larger MTU paths to handle the additional header bytes. Presently VPNs provide a smaller end-to-end MTU size.

There's not expected to be much value to VoIP traffic, simple DNS

requests or other similar protocols that nearly always send small packets. (DNS zone transfers could use larger packets). Operating on a larger MTU Path should have no adverse affect on the end-to-end communications.

1.5. IP Traffic today

It's acknowledged that a majority of Internet traffic today uses small MTU size packets. A study of IP traffic at the AMS-IX IXP in Amsterdam showed the following breakdown [TULYU2011].

Size	Current	Average	Maximum	Minimum
0 - 63 Bytes	0.0%	0.0%	0.0%	0.0%
64 - 127 Bytes	41.2%	41.1%	45.7%	38.7%
128 - 255 Bytes	3.5%	3.4%	4.9%	2.8%
256 - 511 Bytes	2.1%	1.9%	2.2%	1.6%
512 - 1023 Bytes	2.7%	2.5%	2.8%	2.1%
1023 - 1513 Bytes	28.8%	27.8%	29.4%	24.8%
1514 Bytes	21.8%	23.3%	26.1%	21.5%
> 1514 Bytes	0.0%	0.0%	0.0%	0.0%

Weekly Graph - 25 October 2011 to 1 November 2011 (Note: This table is shown in Ethernet frame sizes, ie: 14 Bytes greater than IP MTU)

Table 2: AMS-IX Frame Size Distrubution

The AMS-IX IXP does not provide customer ports configured to anything other than 1,500 Bytes; hence, today AMS-IX will never measure traffic in the final row of this table. (ie: Above 1,500 Bytes IP MTU size). It's safe to say that any IXP operating at the default 1,500 Byte MTU will never see packets above 1,500 Bytes. This means that there's no way to measure the potential traffic until Jumbo Frames on the IXP are enabled.

1.6. NRENs and Jumbo Frames

Research network (NRENs etc) have long-standing operational experiences with Jumbo Frame enabled networks. They have taken the time to test and deploy larger MTU sized networks globally [JET2007] [SUMMERHILL2003].

1.7. Requirements Language

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 RFC 2119 [RFC2119].

2. The Property of an IXPs Switch Fabric

An IXP configuration can vary dramatically. It can be a very simple switch without monitoring or it can be a multi-site multi-terabit infrastructure with 24/7 NOC support and extensive portal support for network customers.

This document only addresses Ethernet based IXPs (which is today the

near de-facto technology). Ethernet ports can be configured in two ways:

a. Untagged ports with all traffic destined for the shared fabric.

b. Tagged ports with traffic controlled by a Virtual LAN (VLAN) identifier. Frames are placed into whatever configured virtual fabric the switch is configured with. This could include some

configurations where only two customer ports communicate privately.

Customers connecting to an IXP need to be operating in the correct tagged or untagged mode. Untagged packets sent into a tagged port will not propagate. This should be considered part of an IXPs standard customer configuration review and install testing process.

This document assumes that the IXP is operating a hardware platform that can provide its customers with a large MTU service. Most modern hardware provides support for Jumbo Frames.

If an IXP can only operate at 1,500 Byte MTU, then this document is not appropriate till the IXP upgrades the hardware platform.

Its quite possible that an existing IXP is operating today with an MTU value above 1,500 Bytes; but has never told its customers. This is not recommended; but is known to work. It is not recommended that customers take advantage of this without the coordination of the IXP operator. See below.

3. MTU Size Considerations

The default payload MTU on Ethernet is 1,500 Bytes. This is defined by the IEEE 802 specification. There is normally no configuration required by network or IXP operators to ensure that clean communications is provided to interconnected networks (IXP customer-to-customer communications). All Ethernet hardware operates at 1,500 Byte MTU, including switches, routers, servers, end-user computers, etc.

Jumbo Frame support is provided by many hardware vendors and some non-Ethernet based systems also have greater than 1,500 Byte MTUs. As IP packets can be transported by many different media types (Ethernet, Token rings or FDDI rings, POS, Radio links, VPNs, Tunnels, etc), the IP protocol can handle nearly any MTU side.

IXPs mainly use Ethernet fabrics and layer 2 communications on Ethernet fabrics require matching MTU sizes.

Adding support for Jumbo Frames means that a higher MTU value needs

to be picked.

1,500 Bytes This is the default from the IEEE 802 specifications.

4,352 Bytes FDDI as defined in RFC 1390 [RFC1390].

4,470 Bytes SONET POS links along with older switches use this.

9,000 Bytes Less than an absolute maximum value; but a number that's easy to remember [JET2007].

9,170 Bytes Used by some hardware.

9,174 Bytes Used by some hardware. Used by CERN.

9,180 Bytes Used by some hardware. Used by Internet2/Abilene Backbone [SUMMERHILL2003], CalREN, etc.

9,192 Bytes Used by some hardware.

9,216 Bytes Used by some hardware.

An extensive study of Jumbo Frame sizes can be found in a presentation by Joe St Sauver in 2003 [SAUVER2003].

The MTU size picked needs to also address the potential of a frame being transported via an encapsulation protocol that reduces overall frame size. Encapsulation could exist within the transport from the router to the IXP and reduce the customers MTU. This means that using the absolute maximum value of the hardware platform could cause issues for customers.

The IXP operators can choose from many hardware vendors. There's no industry standard for an exact Jumbo Frame size; so it varies by vendor and sometimes even by platform. Add to that, an IXP operator and can configure the fabric to nearly any size below their hardware maximum.

3.1. Jumbo Frame size recommendation

It's RECOMMENDED that Jumbo Frames are defined as 9,000 Bytes.

The choice of 9,000 Bytes is based on experience at the Networking and Information Technology Research and Development (NITRD) - Large Scale Network (LSN) Joint Engineering Team (JET) community [JET2007]. It's considered to be an easy to recall number and hence reduces misconfiguration.

If an IXP operator is going to introduce a Jumbo Frame service, it's RECOMMENDED that they pick 9,000 Bytes. Smaller numbers are not useful anymore (the 4,470 value is a legacy value). While values substantially over 9,000 Bytes may be supported by some vendors, support for substantialy larger values is incomplete at best.

9,000 Bytes easily provides support for a TCP or UDP payload of 8,192 Bytes. Protocols like NFS and iSCSI use 8,192 Bytes for data as this matches multiples of physical disk sector sizes along with CPU virtual memory mapping systems.

The value 9,100 Bytes SHOULD NOT be used as this can not be supported by all hardware (even if it's also an easy number to recall).

3.2. Jumbo Frame size example router configurations

Cisco example.

```
!
interface gigabitethernet 1/1
 mtu 9216
 ip mtu 9000
 ipv6 mtu 9000
!

!
interface vlan 1000
 mtu 9216
 ip mtu 9000
 ipv6 mtu 9000
!
```

Juniper example.

```
interface xe-0/1/0
  mtu 9000
  unit 0
    family inet
      mtu 9000
    family inet6
      mtu 9000
```

Brocade/Foundry example.

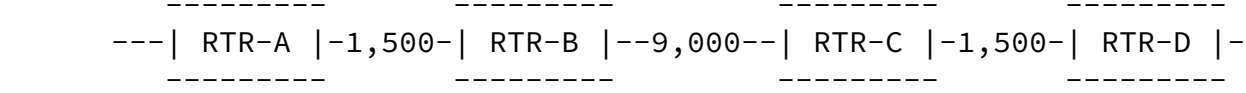
```
!
default-max-frame-size 9216
!
interface ve 81
 ip mtu 9000
 ipv6 mtu 9000
!
```

3.3. Jumbo Frame size limitations

There is a maximum to the size of an Ethernet frame as long at its represented within the link layer size field. Hardware design normally dictates that a memory buffer needs to be reserved or configured into hardware of a specific size. This usually limits the maximum size of a packet.

Every Ethernet frame has a calculated CRC value to make sure the data does not get a bit-level error. With the size of the CRC used by the IEEE 802 Ethernet specifications it's not clear than frames larger than approximately 9,000 Bytes are well protected. Updates to the IEEE 802 specification to implement larger CRCs could allow protection of larger frames; however this subject is outside of the scope of this document.

Jumbo frame links that are surrounded by standard MTU valued links will never be used by end-to-end communications. For example a 9,000 Byte MTU link surrounded by 1,500 Byte MTU links will never see a packet greater than 1,500 Bytes pass via the IXP.



This could simply be put down to future-proofing a network link. In fact many IP backbones operate with 4,470 Byte or ~9,000 Byte long-haul links without any detrimental issues, even if customer only see a 1,500 Byte end-to-end service.

3.4. Consistent MTU Sizes

A maximum sized packet can be sent from a device with a smaller MTU to a device with a larger MTU; however a larger MTU device can't send to a smaller MTU device. A frame sent that's larger than the receivers MTU will produce an incoming error.

A vast majority of Ethernet users have never experienced this issue, as it's unique to the Jumbo Frame configurations. Users have simply lived with the default 1,500 Byte packet size preconfigured on each

and every device.

When two devices communicate over a shared fabric, it's important that both entities have the same MTU value. On an IXP fabric where all peering networks are using the default MTU value of 1,500 Bytes, there's no issue with communications. Should a network configure a different MTU value than other devices on a shared fabric, there's a possibility of a packet not being received by the destination device.

That means IXP operator have to coordinate with every customer any change to the fabrics MTU. If an additional MTU is provided it must be keep on different hardware-platform, specific ports or specific VLANs.

4. Methods of coordinating MTU changes or adding a larger MTU values

Various methods exist for IXPs to operate with more than one MTU value.

a. Provide two untagged ports, one with the de-facto MTU of 1,500 Byte packets and one for the larger MTU value. The IXP fabric should be configured so the two different MTUs are kept seperate. This assumes the IXP and customer has additional network ports to support the larger MTU. Billing for additional ports is not within the scope of this document.

b. Add a duplicate IXP hardware platform configured with the larger MTU value. With this configuration the two different MTU values never touch. This assumes the IXP operator has additional hardware for the new fabric and that the customer has additional network ports to connect to that new IXP fabric. This assumes

the IXP operator has additional space and power for the new fabric; along with the additonal operational overhead required. Billing for additional fabric and ports is not within the scope of this document.

c. Coordinate a specific cutover date/time and have all IXP customers reconfigure at that cutover time. Customers that don't reconfigure will run the risk of loosing operational abilities. This also assumes that every customer has network hardware capable of the larger MTU value. This is not a recommended solution as it removes support for 1,500 Byte MTU communications.

d. Add a second IP range on the existing switch fabric dedicated for the larger MTU range and coordinate a time to increase all switch interfaces to the larger MTU size. Existing 1,500 Byte MTU communications can continue as-is using the existing IP range.

New larger MTU communications can use a new IP range. It's unclear this configuration works in the real world as MTU values are defined by port or virtual port vs. by IP. This is not a configuration recommended by this document.

e. Provide each customer a tagged port with one VLAN setup for 1,500 Byte MTU services and another VLAN setup for the larger MTU service. Existing customers, who want to implement Jumbo Frame support, can choose a cutover time to move from untagged to tagged ports. Existing 1,500 Byte MTU sessions will continue on a VLAN on that tagged port. New customers can be enabled with tagged ports at service delivery time. This is the configuration recommended by this document.

All methods require coordination with the customer to verify configuration correctness. All methods assumes the IXP operator has the additonal operational overhead required to support this offering. IXPs that presently use quarantine ports or VLANs already have processes in place to verify new customers are configured correctly. Providing Jumbo Frame support requires the customer to adjust their configuration and be in-sync with the IXP configuration.

Whatever method is chosen; it's in the interest of the IXP and it's customers to encourage customers to enable Jumbo Frame support.

5. Changing MTU using a Flag-Day approach

IXP operators can assign a flag-day to coordinate a change to the MTU value. This requires communications and coordination with all customers. It also assumes all customers on that fabric are capable of Jumbo Frames.

One advantage of a flag-day is that it allows the IXP provider to remove legacy setups rather than support them forever.

This is also needed if a current Jumbo Frame enabled VLAN is being updated from one size Jumbo Frame to a different one (e.g., from 4,470 bytes to 9,000 bytes).

6. Testing customer MTU values

An IXP operator can test the customer port MTU setting via a simple ping [PING] packet. ICMP filtering on the customers router could impeed this testing. Assuming the test host is connected via a large MTU size path to the IXP, the testing setup can check each customer port to confirm the MTU configuration is correct.

To use a ping packet with IPv4 you are required to set the DF bit. For IPv6 there's no fragmentation during transmission of packets, it's only done at the host level. If you use a server for testing, then the "ping6 -m" (or equivalent option) should be used to control the kernel packet processing and force no fragmentation at the packet level.

Assuming the customer responds to an ICMP ping packet, then a ping with a incrementing packet size will measure the customer-configured MTU value. Commands like tracepath or tracepath6 [TRACEPATH] can be used for these tests.

It's important that the IXP provider has each-and-every customer setup with the identical MTU value.

6.1. MTU Testing Example

A existing IXP did a review of it's Jumbo Frame enabled customers.

The IXP has a 4,470 Byte MTU VLAN and had informed all its customers to operate at 4,470 Bytes MTU.

Customer	Measured MTU	Correct?	Works?
Most	4,470	Yes	Yes
Cust-X	1,500	No	Incorrect!
Cust-Y	4,484	No	Incorrect (but works)
Cust-Z	9,000	No	Incorrect (but works)

Data from testing on a Jumbo Frame enabled IXP

Table 3: Testing IXP customers

The customer responding with the 1,500 Byte MTU should be having operational issues with other peers at that IXP. Any packet greater than 1,500 Bytes sent towards that customer port will be dropped. A small MTU router can send a packet to a large MTU router; however, if a large MTU router sends a packet to a small MTU router and that packet is greater than the receiver MTU; then the packet will be dropped by the receiver with a layer 2 framing error. The customer operating with an MTU of 4,484 or 9,000 Bytes may have it's IP MTU set at 4,470 Bytes and hence operate correctly. Or they may just be lucky and never see a large packet flow across their links.

Further investigating showed that with at least one IP router platform there's a Maximum Receive Unit (MRU) size on the Ethernet interfaces that's based on the physical interfaces memory size. This

allows inbound packets that are larger than the MTU setting. In the case of a ping packet with the DF bit set, the response is fragmented to match the routers MTU.

7. Customer affecting issues

Customers may not like changes within the IXP setup. IXP operators have various choices when it comes to implementing Jumbo Frames.

a. Decide to completely ignore the requirement and define the IXP as

a 1,500 Byte MTU only IXP.

b. Decide to implement Jumbo Frames at the point when the IXP operator announces and creates the IXP (this assumes we are talking about a new IXP).

c. Allow customer to pick how they connect to the IXP. Customer can choose to connect with only one port and only one MTU size, from two or more ports (untagged) each set and allowing access to the MTU values operated by the IXP, one single port (tagged) allowing access to the MTU values operated by the IXP or some other method specific to the IXP.

d. For IXP operators that allow for private VLAN between customers, the MTU value should be defined and if the IXP implements Jumbo Frames, then the value should be communicated to the customers at each port associated with the private VLAN.

There's no need to provide each customer with the same setup; however, operational issues should be addressed if customer configuration is not consistent. Clear documentation and provisioning process will be required.

8. Addressing Plans

Adding support for Jumbo Frames within an IXP could require additional addressing schemes for layer 2 and layer 3. This assumes the existing 1,500 Byte MTU customer-connection stays.

8.1. IPv4/IPv6 Addressing Plans

Technically a large MTU path between two networks could be parallel to the same connection as a standard 1,500 Byte MTU. If that is the case, then it's useful for the IXP operator to provide a different IP network range; but using a similar IP addressing schemes for each path. This means that if a specific prefix is used for an IPv4 /24

or an IPv6 /64 allocated to an exchange fabric with the rest of the address allocated to the customer; then the same final part of the address should be used for the large MTU connection. For example repeat the last octet if it's an IPv4 address or the last 64 bits

with an IPv6 address.

For example, if the IXP used 192.0.2.0/24 (or 2001:DB8:10::/64) today and has 198.51.100.0/24 (or 2001:DB8:11::/64) allocated for the new Jumbo Frame services; then:

192.0.2.NN for customer NN

198.51.100.NN for customer NN on Jumbo Frame service

Or for IPv6:

2001:DB8:10::NN for customer NN

2001:DB8:11::NN for customer NN on Jumbo Frame service

The goal is to make sure that customers always communicate with customers setup with a like MTU value.

It's noted that IXP operators will have to acquire additional IP space for the Jumbo Frame network addressing. This is left outside the scope of this document.

8.2. VLAN Numbering Plans

If the IXP operator provides tagged ports to implement different MTU values; then the operator should allocate VLAN numbers that are compatible with the customer base.

IXP operators can choose to:

a. Some IXP hardware platforms will require the same VLAN number to be used for all customer ports.

b. Some will allow the VLAN number to be set on a per-port per-customer basis.

Allowing the VLAN to be set on a per-port per-customer basis could cause confusion and/or provisioning issues. This is for the IXP operator to decide.

Customers may have limited choices on their VLAN configuration. Some customer hardware platforms do not allow the same VLAN number to be used for different purposes on the same router.

IXP operators should consider coordinating with other IXP operators in their region so the VLAN numbers are not overlapping.

The IXP operator can choose an arbitrary VLAN numbers from the IEEE 802.1Q [IEEE802_1Q] specification range. VLAN number 0 and 4,095 are reserved, as per the specification. VLAN number 1 is used by many platforms to denote the default VLAN and hence should also be avoided.

The IEEE 802.1ad [IEEE802_1AD] Provider Bridges standard, commonly called Q-in-Q, is not applicable to IXP operators implementing Jumbo Frames.

9. IXPs Operating Route Server Configuration

If a route server is provided by the IXP operator on the 1,500 Byte MTU fabric, then another instance of the route server has to operate on the Jumbo Frame MTU fabric and be configured with the correct Jumbo Frame MTU. Hence the Jumbo Frame route server hardware needs to support Jumbo Frames on it's Ethernet interface.

It's important that a customer network is never provided a next hop that's on a port that would drop an incorrectly sized packet.

BGP sessions have the possibility of using larger MSS and MTU sizes when a peering session is initiated. The ability to choose a different MSS is very dependent on the configuration each side of the BGP configuration. For IXPs that implement Jumbo Frames on their route servers; they should report the negotiated MSS size for each BGP session.

10. Known issues for IXPs to consider

Increasing the MTU size has a cost at the network layer. These issues should be considered by the IXP operation for performance, reliability, cost and operational issues.

a. As stated above, it's not clear that frames larger than approximately 9,000 Bytes are well-protected by the existing IEEE 802 checksum method. IXP operators that measure error counters on interfaces should consider providing customers access to their port error statistics (along with their traffic statistics).

b. Jumbo Frames do not have a defined size by the IEEE and hence the strong recommendation that IXP operators choose 9,000 Bytes for

their Jumbo Frame implementation. It's true that each IXP can

choose a different number; however, consistency amongst IXP operators will be a plus.

c. IXP operators should understand that a larger MTU packet will potentially require additional transmission time and buffer memory. Packets may have a larger packet delay and potentially a different or greater jitter value.

d. IXP operators should realize that any mis-configured customer-to-customer communications, with disparate MTU values, will have a potential of failing without any useful reporting at the IP or layer 4 level. No PMTU (Path MTU) packet will be generated should a large MTU packet be sent to a port configured with a smaller MTU.

e. Jumbo Frame support is not intended to change existing end-to-end packet communications if the end-nodes are configured at 1,500 Byte MTU (or lower). Only end-to-end communications where a larger MTU path exists along the whole source to destination path will take advantage of IXPs with larger MTUs.

IXPs should consider recommending existing and new customers enable the larger MTU connection along with the existing 1,500 Byte connections as this provides a potential larger MTU should an end-to-end packet require it.

This document does not address how an IXP will present these issues to its customers or charge for any mitigation of these issues.

In order to encourage the deployment of Jumbo Frames, it's recommended that IXP operators only charge customers if there is a physical difference in their offering.

10.1. PMTU (Path MTU) issues

The IP protocol has two Path MTU Discovery (PMTU) mechanisms to handle packets traveling along a path with varying MTU values for various links in the path.

The IPv4 Path MTU Discovery protocol, RFC 1191 [RFC1191], is

considered often NOT to work. See RFC 2923 [RFC2923] [SAUVER2003]. In IPv6, Path MTU Discovery protocol, RFC 1981 [RFC1981], is considered to work.

However neither the IPv4 or IPv6 PMTU methods will work if the layer 2 fabric has a mismatched value.

10.2. IXP Customer BGP sessions

IXP Customers setup BGP session via an IXP to enable inter-customer routing. For Jumbo Frame enabled IXPs the customers can setup one session or more than one session depending on the MTU match between the two customers.

Customer-A MTU	Customer-B MTU	Choices
1,500 Byte	1,500 Byte	Can only do 1,500 Byte
1,500 Byte	9,000 Byte	Can't communicate
1,500 Byte	9,000 & 1,500 Byte	Can only do 1,500 Byte
9,000 Byte	1,500 Byte	Can't communicate
9,000 & 1,500 Byte	1,500 Byte	Can only do 1,500 Byte
9,000 Byte	9,000 Byte	Can only do 9,000 Byte
9,000 & 1,500 Byte	9,000 & 1,500 Byte	Can do one or both

Table 4: BGP session setup for IXP customers

If the two customers at on both the 1,500 Byte and 9,000 Byte fabrics; then special care should be taken by the IXP customers to confirm their path prefers the 9,000 Byte fabric. This is done so the advantages of the Jumbo Frame fabric will be realized.

This can be done by only enabling the Jumbo Frame BGP session or by keeping the 1,500 Byte BGP session active; but with a lower priority so the routes prefer the next-hop associated with the Jumbo Frame fabric.

IXP customers should note that an extra BGP session will require additional BGP resources; but provide resilience should the Jumbo

Frame fabric fail for any reason.

Outside of the IXPs general operating rules, the BGP session configuration is not within the control of the IXP.

10.3. IXP Operator Service Level Agreements (SLAs)

This document does not state if an IXP operator has to change its SLA to handle Jumbo Frames. That's within the control of the IXP operator.

11. Customer Requirements outside of the IXP operator's control

Many Customers may opt to implement Jumbo Frame services from an IXP,

even if they never will send a packet greater than 1,500 Bytes. The IXP operator should not discourage this behavior as it could be considered as future-proofing their network.

If the IXP has a higher charge for Jumbo Frames and a customer decides to accept those additional charges; but never send a large packet, then this is also acceptable. The customer is allowed to do anything they want, within technical reason.

Customers may have requirement from their own customer-base to provide where possible end-to-end large MTU services even if their customer-base never sends a large packet. This is very hierarchal nature of the Internet and is not the concern of the IXP operator as long as the IXP operator is satisfied with the service level they are providing.

12. IANA Considerations

This memo includes no request to IANA.

13. Security Considerations

The support of Jumbo Frames at IXPs doesn't have any direct impact on Internet infrastructure security.

If there was a security issue related to using Jumbo Frames then providing Jumbo Frame support within IXPs simply extends the potential source location of that thread. Firewalling, filtering or protection at any point on the path does not change when Jumbo Frames on IXPs is provided.

It's possible that security monitoring facilities should be upgraded to be tolerant of and handle Jumbo Frames. Existing hardware may only capture and report on packets up to 1,500 Byte.

14. Acknowledgements

I would like to thank the encouragement and many contributions I received from people with large MTU experience. Bobby Cates (NASA), Greg Hankins (Brocade, was Force10 [HANKINS2008]), Kurt-Erik Lindqvist (NETNOD). Peter Lothberg (STUPI and now DTAG), Kevin Oberman (retired from ESnet), Joe St Sauver, Ph.D. (University of Oregon [SAUVER2003]), Maksym Tulyu (AMS-IX [TULYU2011]) and Mathias Wolkert (NETNOD).

A special thanks goes out to Selina Lo, whom in the late 90's introduced me to the wonders of a working Ethernet Jumbo Frame implementation.

I would also like to also thank the contributions from people with extensive global peering experience: Andy Davidson (LoNAP & Hurricane Electric), Roque Gagliano (Cisco), Mike Leber (Hurricane Electric) and Doug Wilson (Yahoo!).

15. References

15.1. Normative References

[RFC1042] Postel, J. and J. Reynolds, "Standard for the transmission of IP datagrams over IEEE 802 networks", STD 43, RFC 1042, February 1988.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191,

November 1990.

[RFC1390] Katz, D., "Transmission of IP and ARP over FDDI Networks", STD 36, RFC 1390, January 1993.

[RFC1981] McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery for IP version 6", RFC 1981, August 1996.

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

[RFC2516] Mamakos, L., Lidl, K., Evarts, J., Carrel, D., Simone, D., and R. Wheeler, "A Method for Transmitting PPP Over Ethernet (PPPoE)", RFC 2516, February 1999.

[RFC2923] Lahey, K., "TCP Problems with Path MTU Discovery", RFC 2923, September 2000.

[RFC894] Hornig, C., "A Standard for the Transmission of IP Datagrams over Ethernet Networks", RFC 894, April 1984.

15.2. Informative References

[HANKINS2008]
Hankins, G., Provo, R., and T. Scholl, "Peering Survey 2008 Results", May 2008, <http://meetings.ripe.net/ripe-56/presentations/eix/Hankins-Peering_Survey_2008_Results.pdf>.

[IEEE802_1AD]
"802.1ad - Provider Bridges", May 2006, <http://www.ieee802.org/1/pages/802.1ad.html>.

[IEEE802_1Q]
"802.1Q - Virtual LANs", November 2006, <http://www.ieee802.org/1/pages/802.1Q.html>.

[Internet2_LSR]
"Internet2 Land Speed Record", November 2011, <http://www.internet2.edu/lsr/>.

[JET2007] "Recommendation on IP MTU for the JET community", April 2007, <http://www.nitrd.gov/subcommittee/lsn/jet/9000_mtu_statement.pdf>.

[MATHIS2002] Mathis, M., "Raising the Internet MTU", November 2002, <http://staff.psc.edu/mathis/MTU/>.

[NANOG2003] Cottrell, L., "Achieving Record Speed TransAtlantic End-to-end TCP Throughput", June 2003, <http://www.nanog.org/meetings/nanog28/presentations/cottrell.pdf>.

[NANOG2008] Scholl, T., "NANOG42 - Increasing the MTU of the Internet", February 2008, <http://www.nanog.org/meetings/nanog42/presentations/scholl.pdf>.

[PING] "ping, ping6 - send ICMP ECHO_REQUEST to network hosts", November 2007, <http://www.unix.com/man-page/Linux/8/ping>.

[SAUVER2003] St Sauver, J., "Practical Issues Associated With 9K MTUs", February 2003, <http://pages.uoregon.edu/joe/jumbo-frames.ppt>.

[SUMMERHILL2003] Summerhill, R., ""Jumbo" Frames and Internet2", February 2003, <http://globalnoc.iu.edu/i2network/maps--documentation/policy-statements.html>.

[TRACEPATH] Kuznetsov, A., "tracepath, tracepath6 - traces path to a network host discovering MTU along this path", November 2007, <http://www.unix.com/man-page/Linux/8/tracepath>.

[TULYU2011] Tulyu, M., "Jumbo Frames in AMS-IX version 0.3", November 2011, <http://ripe63.ripe.net/presentations/

129-Jumbo_Frames_RIPE63_Nov2011.pdf>.

Author's Address

Martin J. Levy
Hurricane Electric
760 Mission Court
Fremont, CA 94359
US

Phone: +1 510 580-4100
Email: martin@he.net
URI: http://he.net/