HURRICANE ELECTRIC
INTERNET SERVICES

Jumbo Frame Deployment at IXPs (Internet Exchange Points)

draft-mlevy-ixp-jumboframes-00.txt

9, 000 Bytes – Go big or go home!

Hurricane Electric
IPv6 Native Backbone – Massive Peering!

RIPE64 – EIX-WG

Ljubljana, Slovenia – 18th April 2012

Martin J. Levy, Director IPv6 Strategy
Hurricane Electric

NATIVE IPv6 EVERYWHERE

WE LIVE IN A 1,500 BYTE WORLD

1,500 Byte MTUs are today's norm

A sample tracepath showing MTU values across a long Internet path

```
$ tracepath www.royal.gov.uk
 1:  staff.he.net (216.218.248.2)                          0.077ms pmtu 1500
 1:  gige-g3-10.core1.fmt1.he.net (216.218.248.1)          0.563ms
 2:  10gigabitethernet1-2.core1.sjc2.he.net (72.52.92.110)  0.703ms
 3:  10gigabitethernet9-1.core1.nyc4.he.net (184.105.213.174) 70.664ms
 4:  10gigabitethernet1-2.core1.lon1.he.net (72.52.92.242)  138.744ms
 5:  lonap.thn.dedipower.net (193.203.5.118)                139.452ms
 6:  te-0-2-0.asr.thn.dedipower.net (89.151.95.93)          141.781ms asymm 12
 7:  89-151-95-28.servers.dedipower.net (89.151.95.28)      141.163ms asymm 11
 8:  89-151-95-82.servers.dedipower.net (89.151.95.82)      141.146ms
 9:  f5lb1.tvhc1.dedipower.net (89.151.95.43)               140.570ms
10:  no reply
11:  no reply
12:  no reply
13:  f5lb1.tvhc1.dedipower.net (89.151.95.43)               140.793ms !H
     Resume: pmtu 1500
$
```

Nothing incorrect with this path; this is the way the Internet works day-in-day-out

SO WHY WRITE A BCP/RFC?

Data usage has changed / Bandwidth has changed

- End-to-end packets are presently 1,500 bytes at best!
- Data usage patterns have changed (for the better)
  - Datacenter to datacenter data transfers are growing
  - Cloud storage and remote restful access are more common
  - Backbone bandwidth pipe size has increased significantly
- Global interconnections are now “the norm”
  - It’s not uncommon to see continent-to-continent data transfers
  - Disparate backbones used for communications (src & dst)
  - IXPs are in the path (which is a good thing)

Throughput vs. MTU/MSS

WARNING! MATHS AHEAD

$$W = \sqrt{\frac{8}{3p}} \quad (1)$$

Substitute W into the bandwidth equation below:

$$BW = \frac{data\ per\ cycle}{time\ per\ cycle} = \frac{MSS * \frac{3}{8}W^2}{RTT * \frac{W}{2}} = \frac{MSS/p}{RTT\sqrt{\frac{2}{3p}}} \quad (2)$$

Collect the constants in one term, $C = \sqrt{3/2}$, then we arrive at:

$$BW = \frac{MSS}{RTT}\frac{C}{\sqrt{p}} \quad (3)$$

**

** Matt Mathis - The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm http://www.psc.edu/networking/papers/model_abstract.html

Throughput vs. MTU/MSS (simplified)

NATIVE IPv6 EVERYWHERE

WARNING! STILL MORE MATHS AHEAD

Maximum Segment Size

Doesn't exist at IXPs!

Throughput <= ~0.7 * MSS / (rtt * sqrt(packet_loss))

Some Random Constant

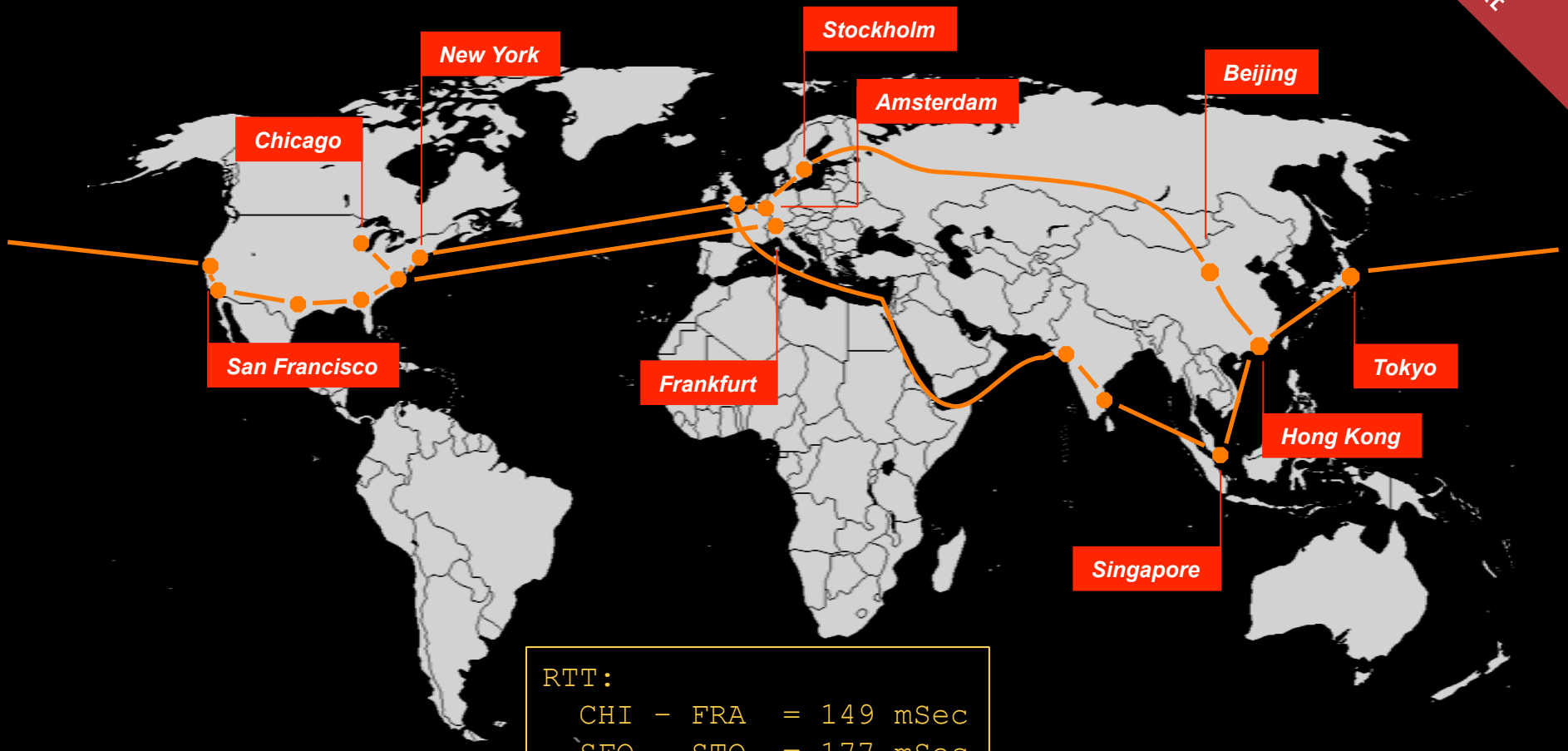
Round Trip Time

Relationship between MTU & MSS:

IPv4:	8,960 MSS = 9,000 MTU - 20 (for IPv4 header) - 20 (for TCP header)
IPv6:	8,940 MSS = 9,000 MTU - 40 (for IPv6 header) - 20 (for TCP header)

** Matt Mathis - The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm http://www.psc.edu/networking/papers/model_abstract.html

Nothing strange about ~200 mSec paths




```
RTT:
  CHI - FRA  = 149 mSec
  SFO - STO  = 177 mSec
  HKG - AMS  = 183 mSec
  LAX - SIN  = 199 mSec
  LON - SIN  = 204 mSec
```

Throughput – significant theoretical differences

- Sample MSS calculations
  - Hong Kong – Amsterdam (via TEA land cable)
  - ~3.6 Gbps @ MSS = 1,460 Bytes / RTT = 190 mSec
  - ~9.4 Gbps @ MSS = 8,960 Bytes / RTT = 190 mSec
- Significant differences in theoretical network b/w limits with larger MSS
- Not relevant to broadband/mobile users
- Well known effect; but hard to replicate outside of NRENs
- My focus is to provide this capability on commercial backbones

** http://www.switch.ch/network/tools/tcp_throughput/index.html

THE BCP/RFC

draft-mlevy-ixp-jumboframes-00

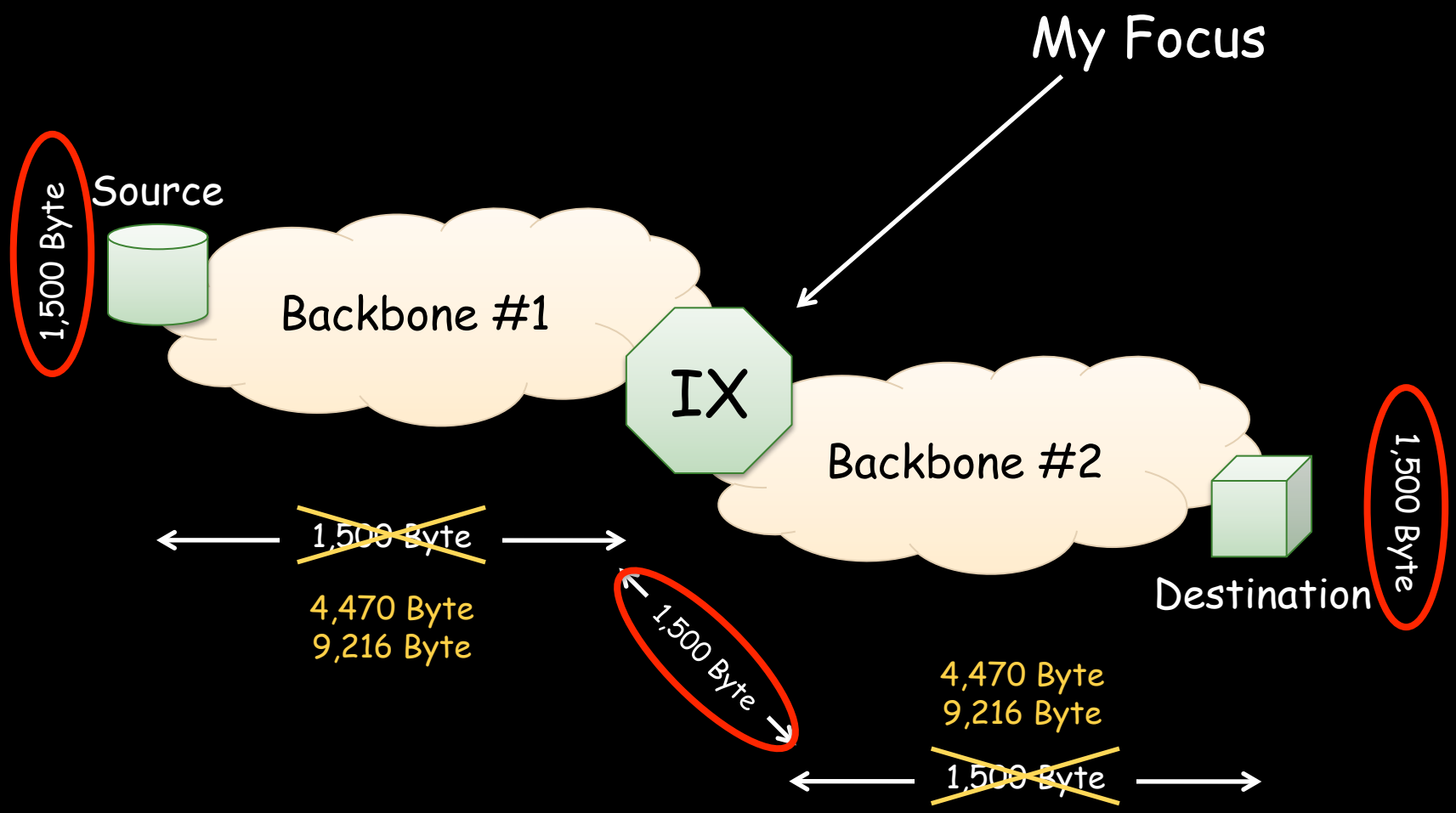
Abstract

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.

http://tools.ietf.org/html/draft-mlevy-ixp-jumboframes-00

draft-mlevy-ixp-jumboframes-00



draft-mlevy-ixp-jumboframes-00

A sample tracepath showing MTU values changing on a long path

```
$ tracepath6 www.netnod.se
 1:  [LOCALHOST]                                          0.020ms pmtu 9000
 1:  2001:470:0:238::1                                    0.401ms
 1:  2001:470:0:238::1                                    0.397ms
 2:  gige-g4-24.core2.fmt1.he.net                         1.096ms asymm  3
 3:  gige-g6-18.core1.fmt2.he.net                         4.972ms
 4:  10gigabitethernet1-1.core1.sjc2.he.net               3.914ms
 5:  10gigabitethernet3-3.core1.den1.he.net              28.139ms
 6:  10gigabitethernet8-2.core1.chi1.he.net              52.438ms
 7:  10gigabitethernet7-2.core1.nyc4.he.net              77.025ms asymm  5
 8:  10gigabitethernet1-2.core1.lon1.he.net             144.911ms asymm  6
 9:  10gigabitethernet5-2.core1.ams1.he.net             146.002ms asymm  7
10:  10gigabitethernet4-1.core1.sto1.he.net             167.826ms asymm  8
11:  10gigabitethernet4-1.core1.sto1.he.net             168.686ms pmtu 4470
11:  ve2.10ge-2-3.outer-b-gw.sth.netnod.se              172.276ms asymm  9
12:  ve2.10ge-2-3.outer-b-gw.sth.netnod.se              168.107ms pmtu 1500
12:  www.netnod.se                                      168.592ms reached
     Resume: pmtu 1500 hops 12 back 54
$
```

draft-mlevy-ixp-jumboframes-00

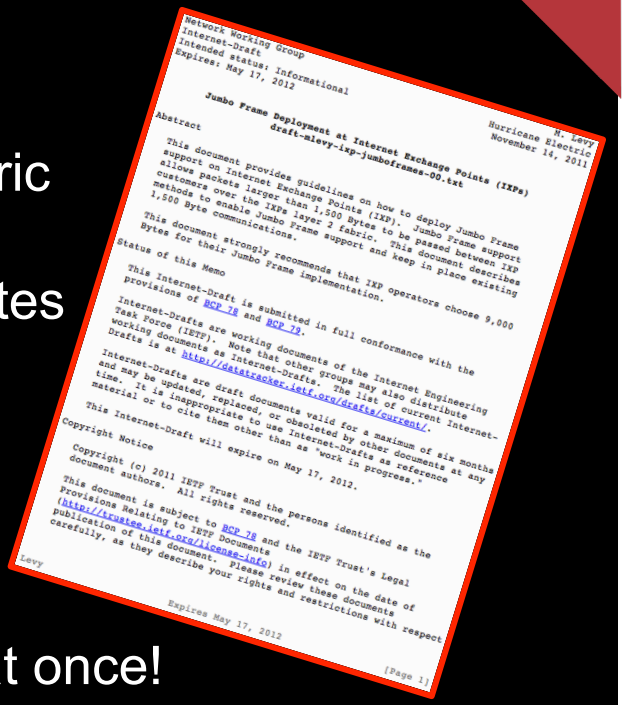
A sample tracepath showing MTU values changing on a long path

```
$ tracepath6 www.stupi.se
 1:  [LOCALHOST]                                          0.030ms pmtu 9000
 1:  2001:470:0:238::1                                    0.411ms
 1:  2001:470:0:238::1                                    0.394ms
 2:  gige-g4-24.core2.fmt1.he.net                         0.815ms asymm  3
 3:  gige-g6-18.core1.fmt2.he.net                         0.723ms
 4:  10gigabitethernet1-1.core1.sjc2.he.net               1.161ms
 5:  10gigabitethernet3-3.core1.den1.he.net              32.046ms
 6:  10gigabitethernet8-2.core1.chi1.he.net              59.390ms
 7:  10gigabitethernet7-2.core1.nyc4.he.net              69.979ms asymm  5
 8:  10gigabitethernet1-2.core1.lon1.he.net             143.209ms asymm  6
 9:  10gigabitethernet5-2.core1.ams1.he.net             143.141ms asymm  7
10:  10gigabitethernet4-1.core1.sto1.he.net             168.682ms asymm  8
11:  10gigabitethernet4-1.core1.sto1.he.net             167.818ms pmtu 4470
11:  2001:7f8:d:fb::34                                  168.571ms asymm  9
12:  2001:440:1880:3000::60                             168.191ms asymm  9
13:  2001:440:1880:1016:202:e3ff:fe00:17ff              168.382ms reached
     Resume: pmtu 4470 hops 13 back 55
$
```

draft-mlevy-ixp-jumboframes-00

- Internet Exchange Points (IXPs) are vital IP backbones interconnect
  - All IXPs are Ethernet layer-2 based these days
  - 99% of IXPs default to 1,500 Bytes MTU (1,514 Frame)
  - A few IXPs that run high MTUs (NETNOD & NASA AIX)
- No documentation exists on how IXPs should provide large MTU services
  - No standard on large MTU size
  - Even the name “Jumbo Frames” is not agreed upon
- Draft has written with input from various operators
  - Been “sitting around for nearly a year” mulling it over
  - Sent for review within IXP industry before submitting to IETF

draft-mlevy-ixp-jumboframes-00


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 then 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

Copyright (c) 2011 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

Levy

Expires May 17, 2012

[Page 1]

- Points covered:
  - Force MTU to be consistent over layer-2 fabric
  - Propose a consistent MTU value – 9,000 Bytes
  - Propose various solutions to implement
    - VLAN based (NETNOD & NASA AIX works)
    - Duplicate fabric hardware (too expensive!)
  - Explain how to do a “flag day” – all change at once!
  - Explain how to do BGP – prefer 9,000 Byte path
  -
  - Explain pitfalls – misconfigured layer-2 failures

Why 9,000 Bytes?

- Too many choices for size based on h/w vendor or h/w version …
  - 9,000 Bytes
  - 9,170 Bytes
  - 9,174 Bytes
  - 9,180 Bytes
  - 9,192 Bytes
  - 9,216 Bytes

  } Confusion or misconfiguration or both
- The number 9,000 is easy to remember – less confusion
  - 9,000 Bytes – handle at least 8,192 Bytes user data plus TCP/IP header
- See references in draft [JET2007]

Why 9,000 Bytes?

- Large packets are needed for:
  - Mass data replication – datacenter X to datacenter Y
  - Storage – Amazon S3, MS Azure Storage, iCloud by Apple
  - NNTP, DNS zone xfer's, etc
- IXPs are in the path for any-to-any communications
  - This draft addresses only ONE element within a data path
- Also allows BGP sessions at IXPs to operate with large MSS

draft-mlevy-ixp-jumboframes-00

Does a large MTU BGP peering session get a large TCP MSS value?

MSS	IP ADDRESS	DNS NAME	ASN	AS-NAME	ROUTER
1240	194.68.128.189	netnod-ix-ge-b-sth-1500.edpnet.net	AS9031	EDPNET	Cisco
...					
1440	195.69.119.34	pro4-gw.stupi.net	AS1880	STUPI	Cisco
1440	194.68.123.26	netnod-ix-ge-a-sth.stupi.net	AS1880	STUPI	Cisco
1420	2001:7f8:d:fb::34	-	AS1880	STUPI	Cisco
1420	2001:7f8:d:ff::26	-	AS1880	STUPI	Cisco
...					
1460	195.69.119.172	10ge-2-3-4470.outer-b-gw.sth.netnod.se	AS8674	NETNOD-IX	Brocade
...					
4410	2001:7f8:d:fb::19	-	AS1653	SUNET	Juniper
4410	2001:7f8:d:fb::24	mtu4470.ge-b.netnod.nordu.net	AS2603	NORDUNET	Juniper
4430	195.69.119.143	netnod-ix-ge-b-sth-4470.port80.se	AS16150	PORT80	Cisco
4430	195.69.119.181	netnod-ix-ge-b-sth-4470.microsoft.com	AS8075	MICROSOFT	Juniper
4430	195.69.119.19	-	AS1653	SUNET	Juniper
4430	195.69.119.24	se-fre.nordu.net	AS2603	NORDUNET	Juniper

?

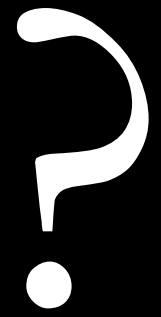
Relationship between MTU & MSS:

IPv4: 4,430 MSS = 4,470 MTU - 20 (for IPv4 header) - 20 (for TCP header)
IPv6: 4,410 MSS = 4,470 MTU - 40 (for IPv6 header) - 20 (for TCP header)

draft-mlevy-ixp-jumboframes-00

- Summary:
  - Large MTU end-to-end traffic can optimize traffic flow and reduce packet overhead
  - IXPs are a key part of the global routing system
  - Someone has to document how to do this
  - Not an end-to-end panacea – but a good start

Q&A



Contact:

Martin J. Levy
Director, IPv6 Strategy
Hurricane Electric
760 Mission Court
Fremont, CA 94539, USA
http://he.net/
martin at he dot net
+1 (510) 580 4167