orion

Jumbo Frame support

Eric Guo

Ontario Research and Innovation Optical Network

Content

Terminology

- MTU, Jumbo Frame, MSS and Fragmentation

Jumbo frame setup

- Transport system and IP system

Implementation Segment

- Core Access and Boundary segment

Implementation Step

- R&E L2VPN Xconnect and L2VPN VPLS Service

orion

Terminology

- Jumbo Frame and Related concept:

- **MTU**: MTU is short for Maximum Transmission Unit, the largest physical packet size, measured in bytes, that a network can transmit. Any messages larger than the MTU are divided into smaller packets before transmission.

- · **Jumbo**: Jumbo frames are frames that are bigger than the standard Ethernet frame size, which is 1518 bytes (including Layer 2 (L2) header and FCS). The definition of frame size is vendor−dependent, as these are not part of the IEEE standard.

- **Baby giants**: The baby giants feature allows a switch to pass or forward packets that are slightly larger than the IEEE Ethernet MTU. Otherwise, the switch declares big frames as oversize and discards them.

- **MTU Size**

The Maximum Transmission Unit (MTU) is the largest possible frame size of a communications Protocol Data Unit (PDU) on an OSI Model Layer 2 data network. The size is governed based on the physical properties of the communications media. Historical network media were slower and more prone to errors so the MTU sizes were set smaller. For most Ethernet networks this is set to 1500 bytes and this size is used almost universally on access networks. Ethernet Version 2 networks have a standard frame size of 1518 bytes (including the 14-byte Ethernet II header and 4-byte Frame Check Sequence (FCS)). It should also be mentioned that other communications media types have different MTU sizes. For example, T3/DS3 (or E3) and SONET/SDH interfaces have an MTU size of 4470 bytes (4474 with header).

• MTU and MSS

Another method to handle the increase in MTU size due to encapsulation and the resulting fragmentation is to utilize the TCP Maximum Segment Size (MSS) parameter. The MSS is the largest amount of bytes of payload data able to be sent in a single TCP packet. In other words, the MSS is the largest amount of TCP data (in bytes) that can be transported over a computer network. This is negotiated during the TCP 3-way handshake in the SYN packet. The MSS is defined in RFC 879 for IPv4 and in RFC 2460 for IPv6. The MSS does not include the TCP header (20 bytes) or the IPv4 header (20 bytes) (IPv6 header is 40 bytes).

For IPv6-enabled interfaces we can use the same type of functions, but the IPv6 header is 40-bytes instead of IPv4's ~20-byte header. We must also consider the 20-byte TCP header which is the same size for IPv4 and IPv6.

This MSS option does not work for UDP applications because there is no way to negotiate this during the handshake because UDP is a connectionless protocol. For UDP applications that do not perform PMTUD and set the DF=1 bit, one option may be to configure a policy that sets the DF bit back to zero.

• Path MTU Discovery (PMTUD)

Routers are capable of performing fragmentation of packets to cut them down to size so they fit into the smaller MTU-size tunnels, but this is not optimal. When an incoming packet to a network device gets its size increased due to encapsulation the packet then gets sent through the outgoing interface on its way toward the destination. However, if the new total packet size exceeds the MTU of the outgoing interface, the network device may fragment the packet into two smaller packets before being able to forward the packet. The IPv4 router will fragment and forward the packet, but also send back to the source an ICMP “packet too big” error message to inform the source that it should use a smaller MTU size. IPv6 routers do not fragment the packet on behalf of the source and just drop the packet and send back the ICMPv6 error message.

The primary problem with the MTU size being reduced across the network is that some applications may not be able to work well in this environment. Some nodes that send 1500 byte packets into the DMVPN and subsequently receive an ICMPv4 “packet too big” message from the router may choose to ignore this. These nodes are not performing Path MTU Discovery (PMTUD) as prescribed by IETF Internet RFC 1191 or RFC 1981 and are therefore relying on the IPv4 routers to perform this fragmentation on behalf of the source host. RFC 2923 also covers the topic of “TCP Problems with Path MTU Discovery”. If the application cannot function properly in this environment, there could be end-user impacts. Also, if there is a firewall in the middle of the communication path somewhere that is blocking the ICMP error messages, then that would definitely prevent PMTUD from operating properly.

- **Jumbograms**

Jumbo Frames should not be confused with jumbograms. When discussing communications protocols, "frames" are the Protocol Data Unit (PDU) used at Layer 2 (Data Link Layer) of the OSI model, and "packets" are the PDU used at Layer 3 (Network Layer) of the OSI model. The term "datagrams" is the PDU used at Layer 4 (Transport layer) of the OSI model. A jumbogram is a larger Layer 3 packet that exceeds the link MTU size. IPv4 is capable of generating payloads up to 65535 bytes, while IPv6 is capable of a 32-bit "Jumbo Payload Length" size within a Hop-by-Hop option header. Therefore, IPv6 could support a ridiculous 4.2GB payload. Clearly, that packet could not be transported on any type of common networking interface. Just imagine the repercussions of a retransmission.

Fragmentation

IPv4 routers fragment on behalf of the source node that is sending a larger packet. Routers can fragment IPv4 packets unless the Do-Not-Fragment (DF) bit is set to 1 in the IPv4 header. If the DF bit is set to 0 (the default), the router splits the packet that is too large to fit into the outgoing interface and send the two packets toward the destination. When the destination receives the two fragments, then the destination's protocol stack must perform reassembly of the fragments before processing the Protocol Data Unit (PDU). The danger is when an application sends its packets with DF=1 and does not pay attention to the ICMP “packet too big” messages and does not perform PMTUD.

All IPv6 networks must support an MTU size of 1280 bytes or greater (RFC 2460). This is because IPv6 routers do not fragment IPv6 packets on behalf of the source. IPv6 routers drop the packet and send back an ICMPv6 Type 4 packet (size exceeded) to the source indicating the proper MTU size. It then falls on the shoulders of the source to perform the fragmentation itself and cache the new reduced MTU size for that destination so future packets use the correct MTU size.

• Fragmentation (cont’d)

The primary concern with having the routers performing fragmentation on behalf of the source is the added CPU processing overhead on the router. If IPsec is being used, then the routers on both ends of the tunnel will need to handle the fragmentation and reassembly of the packets. If the routers are performing fragmentation on behalf of the source node, it may be desirable to have the encryption performed prior to encryption. This prevents the destination tunnel router from having to reassemble the fragments and then perform the decryption. In other cases, we may want to fragmentation take place after encryption. If fragmentation takes place after encryption, then the destination tunnel router will need to perform reassembly before it can decrypt the packet which can add CPU overhead. Therefore, it is advisable for most networks to fragment before encryption.

• Fragmentation- work on router

- Fragmentation occurs at L3 not a L2 (unless we deal with FR or ATM or other technologies). For sure it does not apply to Ethernet.

- Other important concept is that the MTU MUST be configured with the same value on each L3 link between 2 routers.

- router1 1500 --------- 1500 router2 9000 ------- 9000 router3 9000 ---------- 9000 router4 1500 --------- 1500 router5

- Fragmentation will occur between router4 and router5 assuming that router4 gets a frame with size 9000 from router3 and will fragment in at least 6 smaller 1500 frames to be sent to router5 (this is a scholastic exercise as since our ip packet most likely is coming from router1 its size won't be larger than 1500).

• Fragmentation- work on router

- if you have a MTU mismatch between 2 ends of a link you will have drops in the larger-to-smaller direction router1 9000 ------- 1500 router2

- router2 will drop all the frames bigger than 1500 (well a little bit more) coming from router1,this scenario is clearly a configuration mistake!!!

- if you enable jumbo support on one link you need to be sure that on the other side of it there is a l3 device which also supports jumbo frames

- L2 switches simply drop frames bigger than the port mtu.. but if this happen you clearly configured something wrong.

- Say that you set a L2 port of a L2 switch to 9000, since the mtu is calculated at l3 the real frame size is the one calculated by the L3 device (host or router) which sends the frame, therefore most likely will be 1500. So there is no risk to drop anything on the switch.

- **Fragmentation -on switch side**

Bridged and Routed Traffic Size Check at Ingress 10, 10/100, and 100 Mbps Ethernet and 10-Gigabit Ethernet Ports Jumbo frame support compares ingress traffic size with the global LAN port MTU size at ingress 10, 10/100, and 100 Mbps Ethernet and 10-Gigabit Ethernet LAN ports that have a non-default MTU size configured.

Bridged and Routed Traffic Size Check at Ingress Gigabit Ethernet Ports

Gigabit Ethernet LAN ports configured with a non-default MTU size accept frames containing packets of any size larger than 64 bytes. With a non-default MTU size configured, Gigabit Ethernet LAN ports do not check for oversize ingress frames.

• Fragmentation - on switch side

Routed Traffic Size Check on the Policy Feature Card

For traffic that needs to be routed, Jumbo frame support on the PFC compares traffic sizes to the configured MTU sizes and provides Layer 3 switching for jumbo traffic between interfaces configured with MTU sizes large enough to accommodate the traffic. Between interfaces that are not configured with large enough MTU sizes, if the "do not fragment bit" is not set, the PFC sends the traffic to the RP to be fragmented and routed in software. If the "do not fragment bit" is set, the PFC drops the traffic.

Bridged and Routed Traffic Size Check at Egress 10, 10/100, and 100 Mbps Ethernet Ports

10, 10/100, and 100 Mbps Ethernet LAN ports configured with a nondefault MTU size transmit frames containing packets of any size larger than 64 bytes. With a nondefault MTU size configured, 10, 10/100, and 100 Mbps Ethernet LAN ports do not check for oversize egress frames.

Bridged and Routed Traffic Size Check at Egress Gigabit Ethernet and 10-Gigabit Ethernet Ports

Jumbo frame support compares egress traffic size with the global egress LAN port MTU size at egress Gigabit Ethernet and 10-Gigabit Ethernet LAN ports that have a nondefault MTU size configured. The port drops traffic that is oversized.

IOS XR forwarding frames

- Cisco IOS XR software supports two types of frame forwarding processes:
- •Fragmentation for IPV4 packets—In this process, IPv4 packets are fragmented as necessary to fit within the MTU of the next-hop physical network.
- MTU discovery process determines largest packet size—This process is available for all IPV6 devices, and for originating IPv4 devices. In this process, the originating IP device determines the size of the largest IPv6 or IPV4 packet that can be sent without being fragmented. The largest packet is equal to the smallest MTU of any network between the IP source and the IP destination devices. If a packet is larger than the smallest MTU of all the networks in its path, that packet will be fragmented as necessary. This process ensures that the originating device does not send an IP packet that is too large.
- Jumbo frame support is automatically enable for frames that exceed the standard frame size. The default value is 1514 for standard frames and 1518 for 802.1Q tagged frames. These numbers exclude the 4-byte frame check sequence (FCS).

Note:IPv6 does not support fragmentation.

Jumbo frame setup

Orion Transport system:

Transport system	ALU 1830	JDSU 3500F WRT 740	Nortal long haul 1600	Canaria-Nortal ActivFlex 6500
Link /MTU	no limit	STCH- Niagara no limit (protocol aware)	SDBR-TBAY 9582	Toro-TBAY 9600 Toro-OTWA no limit Toro-WNDR no limit

Orion IP system:

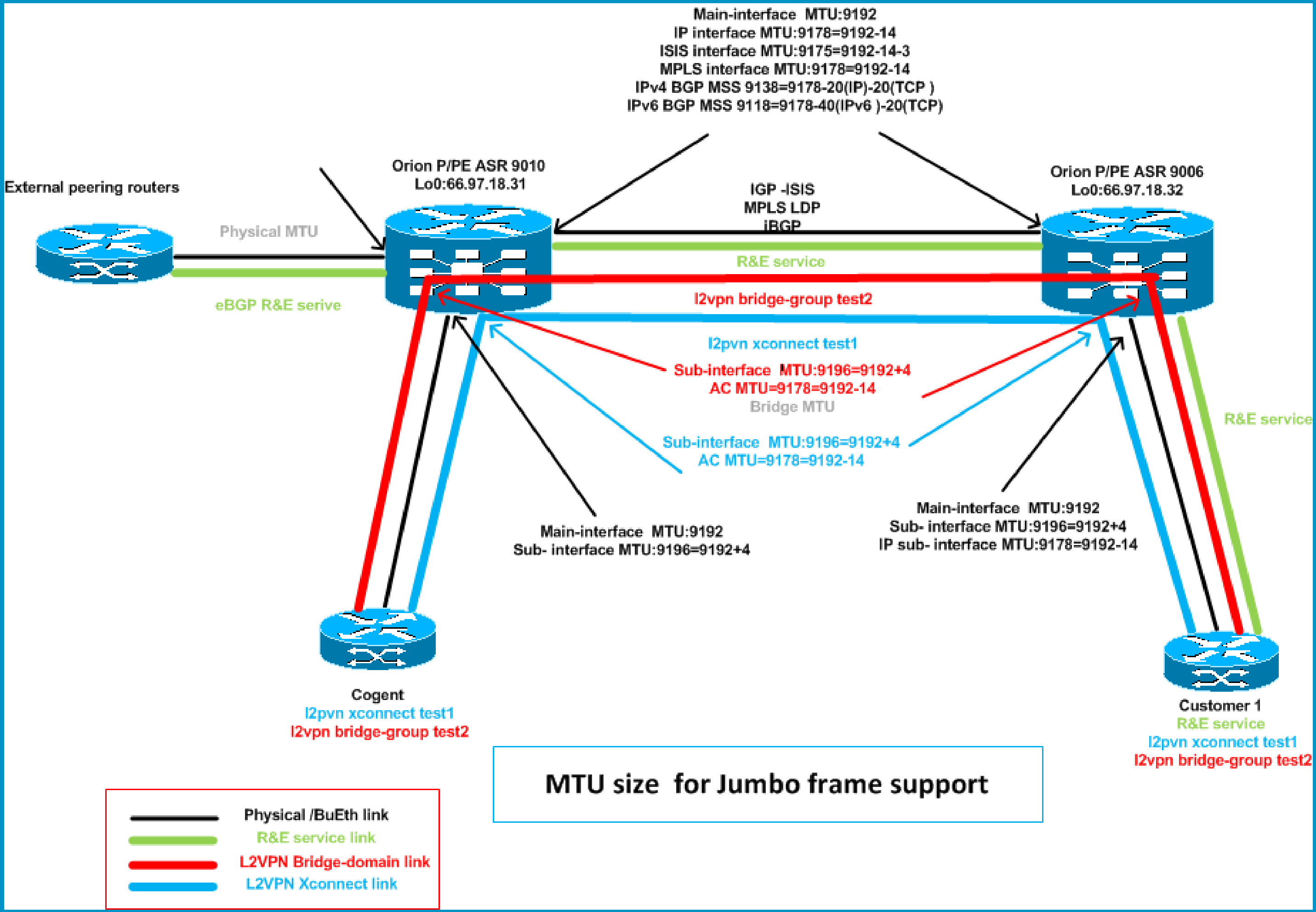
platform	ASR9K	cisco 10700		7600 IOS Router		7200 IOS router	
interface type	all	Giga	FastEthernet	Giga/ port-channel	vlan	Giga/ port-channel	FastEthernet
MTU size	64-9216 (65535)	1500-9100	1500-2000	9216-9216	64-9216	1500-4470	64-4294967295
IP MTU	68-65535	68-1500	68-1500	none	68-1000000	68-1000000	68-1000000
platform	Cisco 3850	ME3600		Catalyst 3560	Catalyst 3550	Catalyst2950	BayStack
interface type	system mtu 1500-9198	10Gi/1Giga	vlan	system 1500-1998	system	system	1500 -non configurable
MTU size		1500-9800	64-9800	jumbo 1500-9000	1500-1546	1500-1530	
IP MTU		68-1500	68-9216	routing 1500-1500			

Jumbo frame setup

Orion MTU size

Interface	MTU size
BuEth/TenGi/Gi	9192
Sub-interface	9196
ISIS-interface	9175
mpls interface	9178
R&E interface	9178
L2vpn interface	9178
IPv6-interface	9178

Jumbo frame setup



Main-interface MTU:9192
IP interface MTU:9178=9192-14
ISIS interface MTU:9175=9192-14-3
MPLS interface MTU:9178=9192-14
IPv4 BGP MSS 9138=9178-20(IP)-20(TCP)
IPv6 BGP MSS 9118=9178-40(IPv6)-20(TCP)

External peering routers

Physical MTU

eBGP R&E serive

Orion P/PE ASR 9010
Lo0:66.97.18.31

IGP -ISIS
MPLS LDP
iBGP

R&E service

l2vpn bridge-group test2

Orion P/PE ASR 9006
Lo0:66.97.18.32

l2pvn xconnect test1

Sub-interface MTU:9196=9192+4
AC MTU=9178=9192-14
Bridge MTU

Sub-interface MTU:9196=9192+4
AC MTU=9178=9192-14

R&E service

Main-interface MTU:9192
Sub- interface MTU:9196=9192+4

Main-interface MTU:9192
Sub- interface MTU:9196=9192+4
IP sub- interface MTU:9178=9192-14

Cogent
l2pvn xconnect test1
l2vpn bridge-group test2

Customer 1
R&E service
l2pvn xconnect test1
l2vpn bridge-group test2

MTU size for Jumbo frame support

━	Physical /BuEth link
🟩	R&E service link
🟥	L2VPN Bridge-domain link
🟦	L2VPN Xconnect link

Jumbo frame setup

MTU size always match as the lowest size by main interface and sub-interface MTU, ISIS MTU, MPLS MTU size setup

IPv4 MTU size change will come with interface MTU size change automatically, IPv4 MTU size will limit IP interface and sub-interface, no impacted l2vpn interface as not process IP protocol; IPv4 MTU will not impact ISIS MTU, MPLS MTU.

MTU size need to be match for ISIS adjacency and OSPF neighbour; ISIS adjacency will be impacted once MTU/ISIS MTU/CLNS MTU size change and only happened on ASR9K-ASR9K platform, ASR9K –IOS router will be impacted once we reset ISIS adjacency session;

MPLS LDP neighbour status won’t be impacted by MTU size change

ISIS status will be impacted MPLS and BGP status; MPLS won’t impacted BGP, but forwarding packet will be impacted by MPLS MTU size changed

Jumbo frame setup

BGP session no be impacted by MTU size change until BGP session reset, the MSS will be changed based on path-mtu-discovery enable after BGP session reset

L2VPN Xconnect AC interface MTU size need to be match on both ends, MTU size limitation is working by main-interface MTU size , sub-interface MTU size using control signal, same function for L2vpn bridge-group group MTU setting, and still need to size matching on both ends

L2VPN bridge group AC interface MTU on both ends could be different size, but packet size will be limited by lower MTU size.

MTU size not support on loopback interface

Jumbo frame setup

MTU change won’t impact IPv6 BGP session. If Ping test start from ASR router, ASR router works as a source host and can then fragment the packets. The ASR9k will create multiple frames with fragmentation header and send them across the link. Each fragment will be smaller than or equal to the ipv6 MTU.

The ipv6 mtu would imposed a limit for transit traffic. For instance if the egress interface ipv6 mtu is lower than the ingress interface and received packet is too big to be forwarded out the egress interface, the router would discard the packet and send an ICMPv6 Packet Too Big (PTB) message to the source

IPv6 does not support fragmentation.

Implementation Segment

Orion Core Segment

- Including P router –PE router and P/PE router – Border router for Layer 2/3 interconnection

Orion Access Segment

- Including the L2/L3 interface that connecting to our members and local RANs network

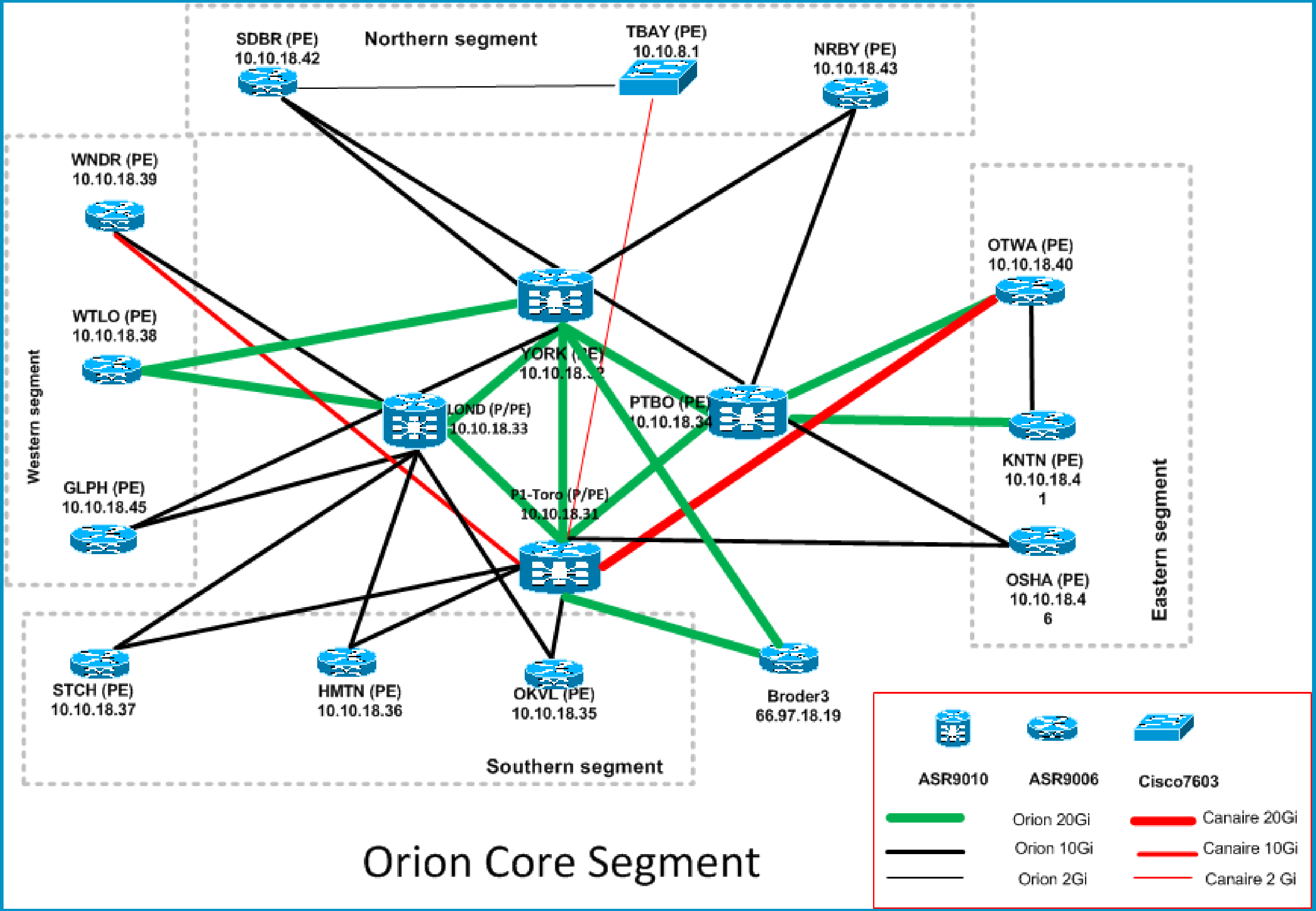
Orion Boundary segment

- Including the interface is facing to our Private/Public peering/Global R&E

Deployment strategic:

Setup from Core segment, then spread out to access segment and edge segment.

A.Orion Core segment

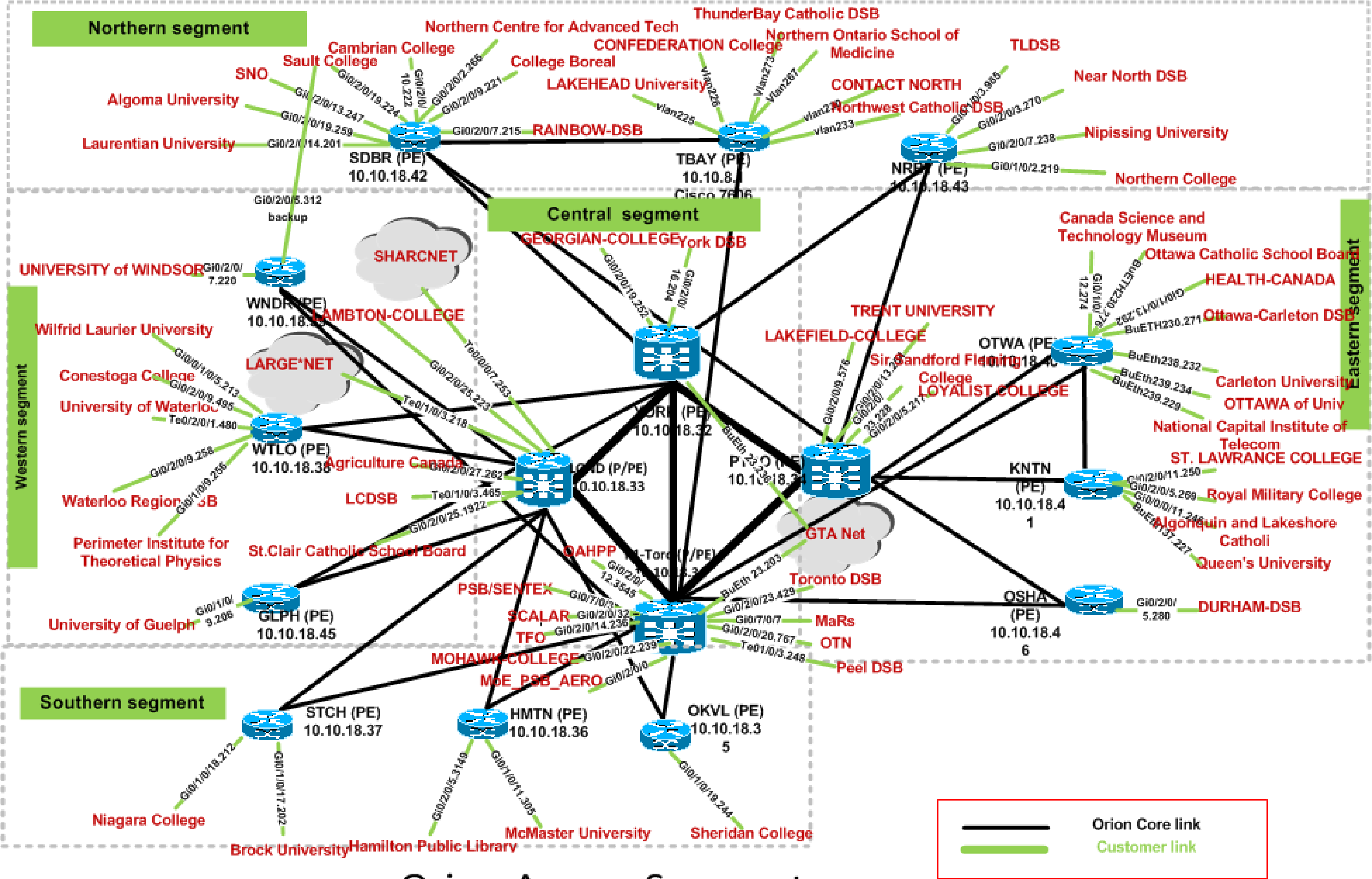


Orion Core Segment

Setup on Core segment

- Between P-P and P-PE
- Conf t
- tcp path-mtu-discovery
- Interface bundle-Ethernet XX
- MTU 9192
- Verification:
- Show int Bundle-Ethernet XX
- Show int TenGi YY
- Show isis interface Bundle-Ether XX
- Show mpls interface Bundle-Ethernet XX location 0/0/CPU0
- Show ipv4 or ipv6 interface Bundle-Ether XX
- Show tcp brief
- Show tcp detail pcb XXXXXX
- Ping peering IP size 9000 donnotfrag

B. Orion Access Segment



Orion Access Segment

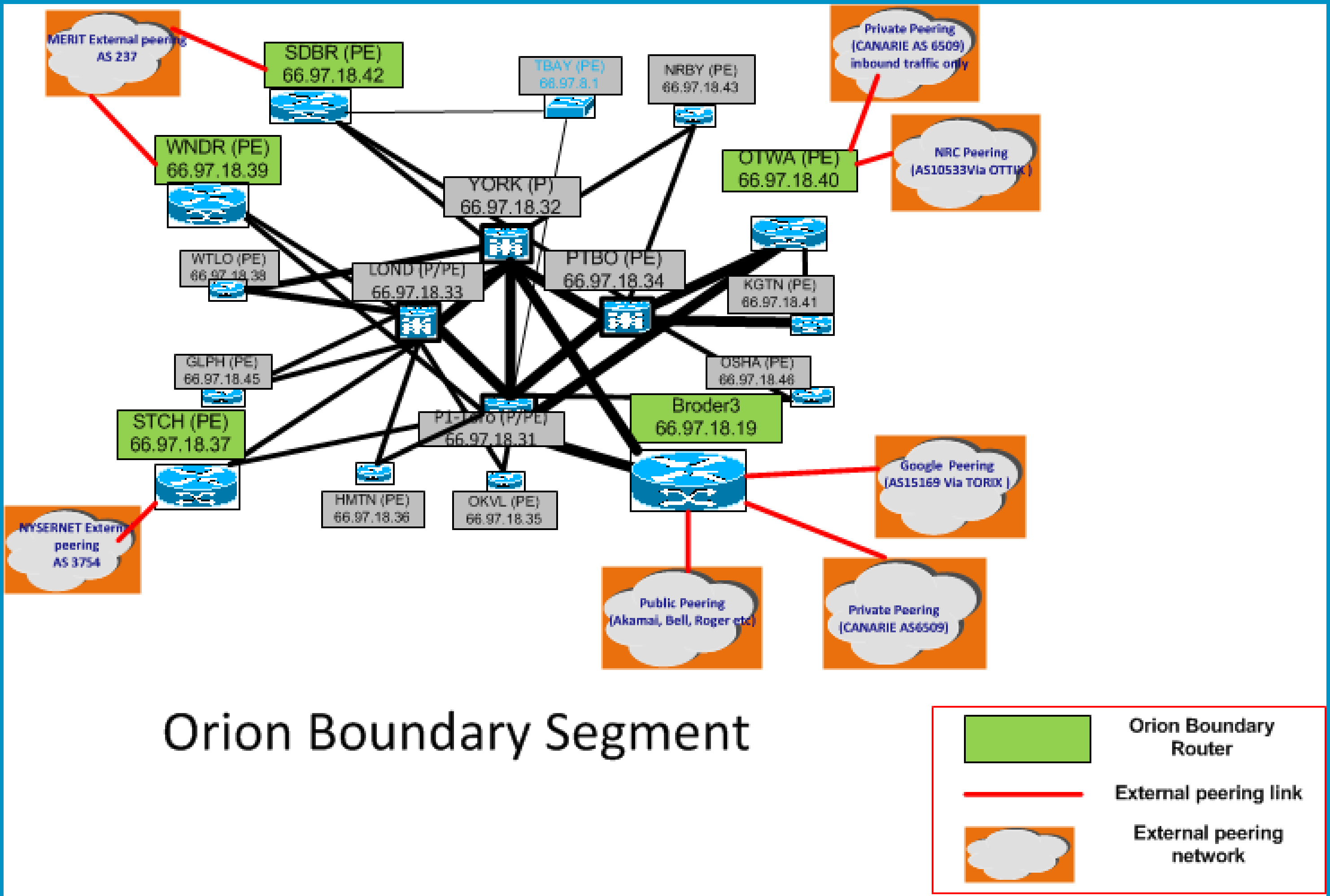
Setup on Access segment

- ON P-P/PE
- Interface Tengi/gi XXX
- MTU 9192
- Verification:
- Show int Tengi/giXXX
- R&E service :
- Show int Tengi/giXXX.YYY

Setup on Access segment

- L2vpn xconnect service:
- On both endpoints
- Show int Tengi/gi XXX.ZZZ
- Show l2vpn xconnect group name detail
- L2vpn bridge group service:
- On all endpoints
- Show int Tengi/gi XXX.ZZZ
- L2vpn bridge-group group name detail

C.Orion Boundary Segment



Orion Boundary Segment

orion

C. Orion Boundary Segment

Peering	location	device	interface
TORIX public peering	TFS	BRDR3	TE 0/1/0/0
Google	TFS	BRDR3	PO111
CARANIE	TFS	BRDR3	TE0/1/0/2
	OTWA	PE-OTWA	BE204
PEERING-FEDERAL-GIGPOP-NRC	OTWA	PE-OTWA	Gi0/1/0/7
NYSERNET-PEERING	STCH	PE-STCH	Gi0/1/0/9
PEERING-MERIT-AS237	WNDR	PE-WNDR	Gi0/1/0/5
	SDBR	PE-SDBR	Gi0/2/0/19.311

Implementation Steps

- Backbone

core end	core device	site	remote device	limiltion type	action plan
York	P-York	BARI	ME-BARI(3600)/CAT1-BARI(2950)	2950 platform	replace with ASR9001 customer link directly
PTBO	P-PTBO				
SDBR	PE-SDBR	SSMR	CAT1-SDBR(2950)/CAT1-SSMR(3650)/BAYSTack	platform	replace with ASR9001, transit device support Jumbo frame
WNDR	PE-WNDR		Radio link	Merit network	work out with Merit network
PTBO	P-PTBO	BLVL	CAT1-BLVL(3560)	no	replace with ASR9001 customer link directly
KGTN	PE-KGTN		ME-BLVL(3600)		

core end	core device	site	remote device	limiltion type	action plan
NRBY	PE-NRBY	TIMM	CAT1-TIMM(3560) /CAT1-NLKD(2950)	platform	TIMM replace with ASR9001, CAT1-NLKD (3560) support
SDBR	P-SDBR	TBAY	DIST-TABY(7600)	no	replace with ASR9006 customer link directly
Toro	P-Toro				
LOND	P-LOND	SARN	CAT1-SARN(3560) /CAT1-CHHM(3550)/CAT1-LOND(3550)	platform	SARN replace with ASR9001, transit device support Jumbo

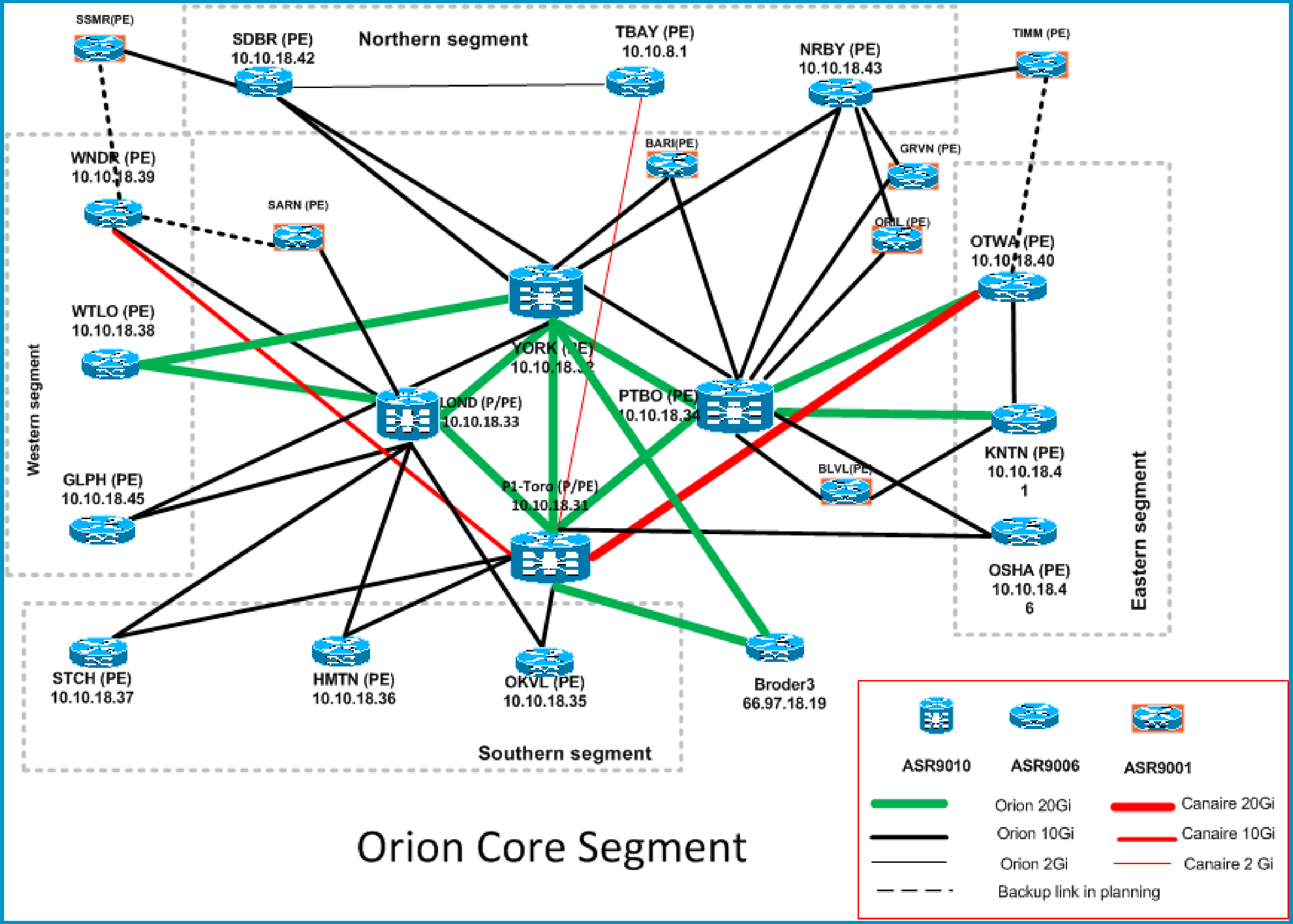
Implementation Steps

- Backbone –L2VPN
- ON CAT4-Toro
- System MTU 9198 and reload the device
- Backbone –IPv6
- ON BRDR3
- Enable MTU size to 9192 on Gi0/0/0/3

Implementation Steps

Segment	Objective	Change NO.
Core segment	Move customer connection endpoint and Replace network platform in order to Orion network could support Jumbo frame end by Access points that facing to customer end. Optional: enable TCP Path-MTU-Discovery on our P/PE router	10
Access Segment	enable MTU Jumbo frame size under each of interface connected to our customer and including end to end interfaces for their L2vpn service of Coegnt connection and IPv6 connection	5
Boundary Segment	enable Jumbo frame size under our Peering interface level, IPv4 BGP MSS will be changed while enable TCP path-mtu-discovery feature under IPv6 segment	1
IPv6 Segment	enable IPv6 Jumbo frame size from our Peering end with our customer end; IPV6 our custmer end size setup will completed with Access segment L2VPN service setup; including 2 steps: on border3 Router, enable Jumbo frame size on interface level; enable IPv6 BGP MSS size by enable TCP path-mtu-discovery	1

Implementation Steps



Orion Core Segment

Q & A

The Ontario Research and Innovation Optical Network

360 Bay Street
7th Floor
Toronto, Ontario M5H 2V6

T. 416-507-9860
F. 416-507-9862

www.orion.on.ca