



双栈环境下 IP 城域网 MTU 研究

颜永明

(中国电信股份有限公司上海分公司 上海 200030)

摘要:通过研究城域网双栈环境下各类 MTU 处理机制的特点,对已有技术存在的不足和适用场景进行了分析,提出了现网部署时需关注的问题和现阶段建议方案,并对理想方案进行了探讨。

关键词:双栈;IPv6;城域网;分段;最大传输单元;最大接收单元;最大分段尺寸

doi: 10.3969/j.issn.1000-0801.2014.10.025

Research on MTU in Dual Stack Environment in Telecom IPMAN

Yan Yongming

(Shanghai Branch of China Telecom Co., Ltd., Shanghai 200030, China)

Abstract: Currently, there are various MTU processing mechanisms in dual stack environment in telecom IPMAN, but they have respective shortages and appropriate usage scenarios. Based on the analyses of above problems, the relatively feasible solution was provided, and the expected ideal target was discussed.

Key words: dual stack, IPv6, IPMAN, segment, MTU, MRU, MSS

1 引言

中国电信股份有限公司上海分公司(以下简称上海电信)IP城域网在全国同类网络中规模最大,网络结构最为复杂。网络分为A平面“城域网通道”和B平面“城域网优化通道”两大部分。A平面主要承载互联网上网业务,B平面主要承载IPTV、软交换、VPN、IPRAN等业务。上海电信在中国电信集团公司下一代互联网示范商用首批试点项目中,主要采用IPv6/IPv4公网双栈过渡技术,于2013年实施了城域网A平面的IPv6改造,首批实现了数十万IPv6公客PPPoE拨号用户的双栈接入。在实践中发现,因不同网络和设备的MTU(max transfer unit,最大传输单元)配置存在差异,IPv6和IPv4在分段和MTU处理机制上有着各自特点,可能会造成用户在有些情况下上网业务的使用出现问题。本文就双栈环境下分段和MTU技术原理、部署中需注意的问题和解决方案进行讨论。

2 数据分组分段相关技术

2.1 MTU 相关定义和解析

MTU指某一通信协议层面上所能通过的最大数据分组字节数大小,在IP城域网环境下通常将IP网络层的MTU简称为MTU。当网络设备转发数据分组时,会检查其数据域大小是否小于等于出接口的MTU设置值。以Ethernet II帧为例,如图1所示,数据域的大小=帧长-DMAC(目的MAC)-SMAC(源MAC)-type(类型字段)-CRC(检验字段),计算所得通常称为数据分组大小。

DMAC	SMAC	type	data	CRC
------	------	------	------	-----

图1 Ethernet II帧结构

一个IP数据分组从源主机到达目的主机所经路径上的各段链路MTU值不尽相同,要完成端到端的传送,IP网络层数据分组尺寸必须小于由各段链路构成的整条路径上的最小MTU。路径上每一段链路所能传送的最大IP分

组尺寸,称为 link MTU(链路 MTU),源主机到目的主机整条路径上的最小 MTU 则称为 path MTU(路径 MTU),其大小等于数据分组途经所有链路 MTU 的最小值。MTU 值可在网络设备的接口上配置,用来确保从该接口发出的数据分组尺寸小于或等于该值。如果源主机发送的 IP 网络层数据分组尺寸大于路径 MTU,该数据分组会被分段或丢弃。

2.2 MRU 定义及解析

PPP(point-to-point protocol,点对点协议)定义了 LCP(link control protocol,链路控制协议)配置选项,该选项用于协商修改 PPP 点对点链路的默认特性,其中包含称为 MRU(maximum receive unit,最大接收单元)的字段,发送给 PPP 对端设备以告知本地设备能接收更大的数据分组,或者通知对端发送更小的数据分组。上海电信公客拨号用户采用 PPPoE 拨号认证方式,PPPoE 基于 PPP,同样定义了 MRU 字段,协商建立 PPPoE 拨号会话时,拨号客户端设备会将拨号接口所配 MTU 值作为 PPPoE MRU 字段发给拨号服务器(城域网中为 BRAS),拨号服务器将收到的 MRU 值与自身接口配置的 MTU 值相比较,选取两者中的较小值作为其后续发送数据分组的最大尺寸。

2.3 MSS 定义及解析

MSS(maximum segment size,最大分段尺寸)用于 TCP 报文,规定了每个 TCP 报文中数据域的最大长度。MSS 定义的数据域不包含 TCP 头部。MSS 与 MTU 的关系为: $MSS = MTU - IP \text{ 头部} - TCP \text{ 头部}$ 。当建立 TCP 连接 3 次握手时,协商的双方会互相通告 MSS 大小。通过设定合适的 MSS 值可确保数据域被 TCP、IP 头部封装后,尺寸小于或等于 path MTU,避免 TCP 报文在传送过程中被分段。MSS 只定义于 TCP,在 UDP 及其他传输层协议中无定义。

2.4 数据分组分段原理解析

IPv4 数据分组头部包含 DF(don't fragment,不分段)位,路由器据此判断能否对数据分组分段。若 IPv4 数据分组大小超过出接口 MTU 且设置了 DF 位,则被丢弃。若未设置 DF 位则由路由器分段后转发。

IPv6 头部与 IPv4 协议不同,包含如下基本字段:版本、业务流类别、流标签、净荷长度、下一分组头、跳数限制、源地址、目的地址,共 40 byte。除基本分组头外,还定义了多种可选 IPv6 扩展分组头,fragment(分段)分组头就属于其中的一种,用于源节点对长度

超过其和目的节点间路径 MTU 的分组实施分段。根据 RFC 规定,除了 hop-by-hop(逐跳)扩展分组头外,其他扩展分组头都不会被传送路径上的节点检查或处理,只有当数据分组到达了目的地址标识的节点时这些扩展分组头才被处理,且必须严格按照它们在数据分组内出现的顺序依次进行。因此,对 IPv6 数据分组而言,分段动作只能在源节点实施,不能被分组转发路径上的路由器执行。当 IPv6 数据分组大小超过路径 MTU 时,会出现分组丢失现象。

3 双栈环境下 IP 城域网 MTU 研究

上海电信公客拨号用户上网路径可划分为 3 个段落,如图 2 所示:第一段落为城域网接入侧,包含用户终端、家庭网关、EPON、汇聚交换机和 BRAS 下联接口;第二段落为城域网络侧,包含 BRAS 上联接口和上海电信城域网路由器;第三段落为 Internet 侧,包括骨干网、异地/异网运营商、ICP 内网及应用服务器。公客拨号用户终端通过家庭网关上联 EPON,接入 IP 城域网 BRAS,路由器等设备,最终实现 Internet 访问。

因城域网、互联网、异地/异网运营商、ISP、ICP 网络情况纷繁复杂,MTU 的配置存在差异,开放的网络环境中数据分组经过的路径又存在不确定性,这些都可能对数据分组因 MTU 问题无法顺利到达目的端。尤其对于 IPv6 数据分组,因其分组头比 IPv4 长 20 byte,若有设备和网络沿用了 IPv4 的 MTU 配置参数值,很可能会引发分组丢失和性能下降等问题。如何避免和解决 MTU 问题,业界提出了很多方案,结合城域网双栈环境,具体分析如下。

3.1 中间设备分段方案

以上海电信城域网为例,BRAS 负责终结 PPPoE 拨号请求,家庭网关可配置成桥接模式或路由模式。桥接模式下 PPPoE 由用户终端发起,路由模式下则由家庭网关发起。考虑到 PPPoE 封装需额外增加 6 byte PPPoE 和

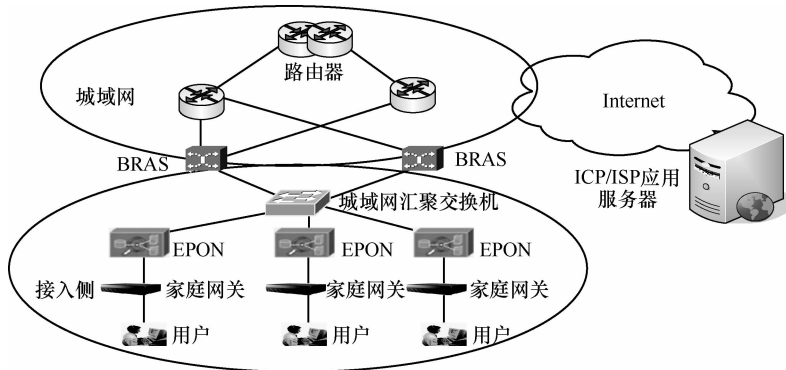


图 2 上海电信公客拨号用户上网路径示意



2 byte PPP 头部开销,上海电信 BRAS PPPoE 拨号接口配置的 MTU 值为 1 492 byte(在标准的 1 500 byte 基础上减去 6 byte PPPoE 头部和 2 byte PPP 头部计算所得),因此 BRAS 用于 PPPoE 协商的 MRU 值也为 1 492 byte。而绝大多数用户终端的 IPv4 和 IPv6 协议栈 MTU 值默认设为 1 500 byte。若家庭网关采用桥接模式,PPPoE 拨号在用户终端和 BRAS 之间完成,经协商后,用户终端和 BRAS 均会选用彼此 MRU/MTU 中较小值 1 492 byte 来发送数据分组,IPv4 和 IPv6 数据分组都不会引起中间设备对其分段。但如果家庭网关采用路由模式,PPPoE 拨号在家庭网关 WAN 口和 BRAS 之间完成,经协商后,它们之间也会选用 MRU/MTU 中较小的值 1 492 byte 来发送数据分组,而此时的用户终端并不知道 PPPoE 协议的存在,最大仍会使用 1 500 byte 发送数据分组,在纯 IPv4 环境下,这些大于 1 492 byte 的数据分组从用户终端发送到家庭网关 WAN 接口时,由于 IPv4 支持中间设备对分组进行分段,所以没有设置 DF 位的 IPv4 数据分组会被家庭网关分段成两个数据分组后发送,数据分组的重组则由目的端负责,这就是中间设备分段方案,最终可实现 IPv4 分组的成功收发。

然而,此方案存在两点不足。

- 家庭网关对大量 IPv4 数据分组的分段操作会影响性能,而目的端对分组重组同样会消耗自身存储、计算资源,进而可能造成时延和分组丢失等问题。
- 从第 2.4 节分析可知,IPv6 并不支持中间设备对分组进行分段操作,因而该方案无法解决 IPv6 数据分组遇到的 MTU 问题。由此可知,在城域网环境中并不适合采用本方案。

3.2 path MTU discover(路径 MTU 发现)方案

path MTU discover 是指当数据分组大小超过某中间设备出接口设置的 MTU 时,该设备将丢弃超长数据分组并发送 ICMP packet too big 消息给发起此超长分组的主机,ICMP 消息中包括了分组丢失设备的 MTU 值。源主机收到此 ICMP 消息后,将重发分组的尺寸设为 ICMP 消息中要求的 MTU 值,如此反复,直到分组的尺寸不大于端到端路径上所有设备出接口的 MTU 值,并顺利达到通信对端主机或服务器为止。path MTU 由数据分组发送方向上途径的各设备出接口 MTU 配置决定,所以该机制是单向的,在来去两个方向上的路径 MTU 发现结果可能会不一致(例如来去方向上途经的路径不对称或者同一设备不同接口上的 MTU 配置值不同)。IPv4 和 IPv6 都支持此方案,但在

具体实现上有差异。对 IPv4 数据分组而言,该机制只对设置了 DF 位的数据分组有效。而 IPv6 协议没有定义 DF 字段,但天然地支持了 path MTU discover 机制。

本方案有如下限制和安全问题。

- 路径 MTU 发现机制依赖于 ICMP packet too big 消息的传递,若有中间设备被禁止产生该消息或传递路径上有中间设备拒绝该 ICMP 消息通过,会使源主机无法侦测路径上的最小 MTU 值,最终导致发现机制失效并产生大量分组丢失。一旦出现此问题,因互联网的复杂性,各种因素互相交织,排障将变得非常困难。
- 中间设备产生并回复 ICMP packet too big 消息一般需由其 CPU 处理,若大量用户触发该类消息或存在恶意攻击行为,设备性能遭受重大影响。
- 若有攻击者利用大量“肉鸡”假冒源主机地址发送大分组至网络,将有大量 ICMP packet too big 数据分组返回攻击无辜的源主机。虽然很多厂商设备能对 ICMP 分组的产生进行限速,但不区分对错的限速会影响到正常使用的 MTU 发现机制,因此多数主流设备厂商并不建议在城域网实际部署中开启该功能。鉴于以上安全问题对于城域网而言可能带来毁灭性的后果,因此本方案在城域网中的可实施性并不强。

3.3 router advertisement messages (路由器通告消息)方案

IPv6 在 RFC2461 中定义了 neighbor discovery (邻居发现)协议,该协议包含 5 种 ICMP 分组类型,分别是重定向消息、成对使用的路由器请求和路由器通告消息、成对使用的邻居请求和邻居通告消息。其中,路由器通告消息中含有 MTU 选项,可以实现在同一链路段上的所有节点使用相同的 MTU 值。基于该协议,通过对家庭网关(受电信管控)进行开发,可实现向连接在家庭网关 LAN 口上的用户终端通告运营商所希望的 MTU 值大小,用户终端接受网关通告并修改 MTU 后,可避免其使用默认却又不合适的 MTU 配置发出数据分组,进而在城域网和互联网上遭遇 path MTU 问题。

此方案的限制在于:

- 因 MTU 选项是 IPv6 neighbor discover 协议中新定义的,IPv4 中无此选项,因而该方案只适用于 IPv6,无法用于 IPv4 协议栈;

- 因位于 Internet 上的应用服务器端网关一般由用户自行部署,其 MTU 设置并不受运营商控制,本方案只能影响公客拨号用户去往服务器数据分组的 MTU 大小,无法解决返回数据分组的 MTU 问题。

若采用此方案,并结合用户服务器端发出的 MTU 大小符合城域网要求,就有可能解决 IPv6 的 MTU 问题。因此,建议在有条件的情况下实施该方案。

3.4 TCP MSS 调整方案

根据第 2.3 节描述,若能确保以 MSS 大小发送的 TCP 类型数据分组的数据域在封装 TCP、IP 分组头后的尺寸小于 path MTU,则可避免数据分组在传送过程中被分段。MSS 大小可由建立 TCP 连接的两端设备直接进行协商(例如 PC 和 Web 服务器直接协商),也可以由中间路由设备参与协商,中间设备参与协商的一个复杂场景流程如图 3 所示。由图可知中间的路由设备可在两个方向上分别控制 MSS 的大小。因 TCP 两端的会话协商过程不受控,很难依靠用户自身协商来确保 TCP 数据分组的 MSS 大小符合 path MTU 的要求,此时配置中间设备参与 MSS 的协商成为运营商的一种可选方案。

基于以上思路,上海电信在使用路由模式的家庭网关上开启了类似 `ip tcp adjust-mss`、`ipv6 tcp adjust-mss` 的命令,

可实现对 TCP 应用客户端和服务端端的 IPv4 和 IPv6 MSS 大小控制。考虑到上海电信城域网上最小 MTU 为 1 492 byte (PPPoE 封装段), $IPv4MSS=pathMTU(1\,492\text{ byte})-TCP\text{ 报头}(20\text{ byte})-IPv4\text{ 报头}(20\text{ byte})=1\,452\text{ byte}$, 所以应将 IPv4 `tcp adjust-mss` 值设置小于或等于 1 452 byte; IPv6 $MSS=path\ MTU(1\,492\text{ byte})-TCP\text{ 报头}(20\text{ byte})-IPv6\text{ 报头}(40\text{ byte})=1\,432\text{ byte}$, 应当将 IPv6 的 `tcp adjust-mss` 值设置小于或等于 1 432 byte。若考虑到用户有自建额外隧道的配置需求,需将 `adjust-mss` 值设置得更小,为额外封装分组头保留余量。

虽然此方案的不足之处在于只对 IPv4 和 IPv6 的 TCP 数据分组有效,而对互联网应用广泛使用的 UDP 等其他传输层协议不起作用,但在实际应用中,设计基于 UDP 的应用时,编程者通常会有意识地将 UDP 数据分组的大小设定为较小值,以适应复杂的网络环境,而在真实环境中能否成功调控 TCP 数据分组大小更具有现实意义,因此,在城域网中应优先应用本方案。

4 结束语

IPv6 技术虽已发展多年,但在大型城域网内规模部署的案例仍然不多。双栈环境下遇到的 MTU 和分段问题较

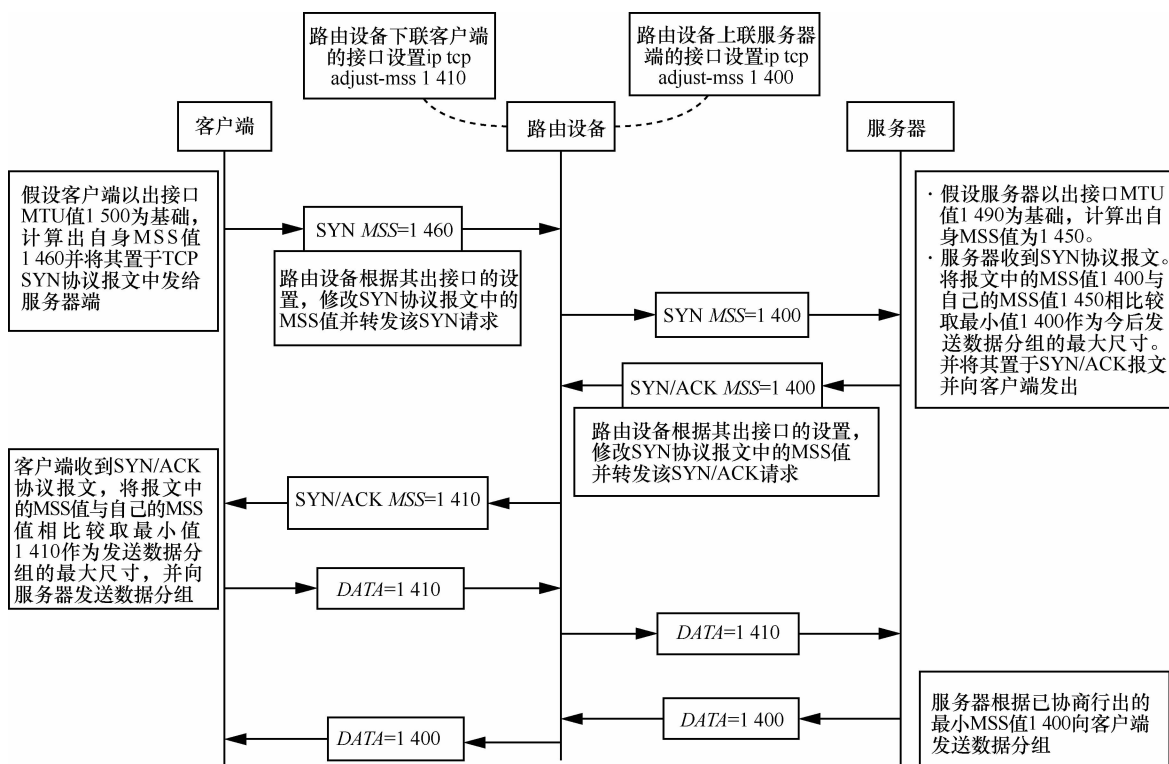


图 3 中间路由设备参与 MSS 协商流程



为复杂,需通盘考虑 IPv4 和 IPv6 的协议特点,运营商、互联网内容提供商专业人员若囿于 IPv4 配置经验,往往会引起 IPv6 的 MTU 问题。根据本文分析,因存在一定技术和应用限制,单纯依靠现有的某个方案并不能彻底解决问题。就城域网运营商而言,现阶段可在综合考虑网络设备功能、性能、安全因素的前提下,结合多种方案,多管齐下,尽可能避免因 MTU 问题导致分组丢失和网络转发性能下降问题。在一个城域网内可考虑适当放大相关设备的 MTU 设置值,优先确保本地网络不会成为 MTU 瓶颈。从长远来看,标准化组织、研发机构和相关厂商应考虑对相关 MTU 处理机制作进一步开发完善,理想中的 MTU 处理机制应该同时满足以下要求:

- 应能同时解决 IPv4、IPv6 的 MTU 问题;
- 应能同时作用于 TCP 和 UDP 等各类协议数据分组;
- 不应采用分段方式增加路由器的处理负担;
- 解决 MTU 问题不能引发新的不可控的安全问题。

就第 3.2 节的方案来看,当特定 ICMP 数据分组能顺利穿越整个网络的情况下,若能彻底解决引发的各类安全问题,也许改良后的路径 MTU 发现方案就能成为一种较为理想的方案。另一方面,ICP、操作系统、终

端、应用软件厂商应充分考量双栈环境下 MTU 问题的复杂性,在彼此负责的段落内为 IPv4 和 IPv6 设置更为合理的 MTU 值,既减少问题出现的几率又确保数据分组的转发效率。

参考文献

- 1 RFC2460. Internet Protocol, Version 6 (IPv6) Specification, 1998
- 2 RFC4459. MTU and Fragmentation Issues with in-the-Network Tunnel, 2006
- 3 RFC4638. Accommodating a Maximum Transit Unit/Maximum Receive, 2006
- 4 RFC2461. Neighbor Discovery for IP Version 6 (IPv6), 1998

[作者简介]



颜永明,男,中国电信股份有限公司上海分公司总工程师室高级工程师、数据专业主管,主要研究方向为数据通信。

(收稿日期:2014-08-27)

(上接第 155 页)

参考文献

- 1 胡荣贻,彭木根,张洪岩等. LTE-Advanced 系统中的移动顽健性优化技术研究. 电信科学, 2013, 29(5):84~89
- 2 祝炜凯,赵竹岩,康剑锋等. 基于实际城区场景的 TD-LTE 异构网络性能研究. 电信科学, 2012, 28(9):47~51
- 3 余翔,张丽,王蓉. LTE-A 中异构网络的切换分析. 电讯技术, 2014, 54(1):89~96
- 4 3GPP TR 36.814. Further Advancements for E-UTRA Physical layer Aspects (Release 9), 2011
- 5 Triantafylllopoulou D, Passos N, Kaloxylas A, et al. Coordinated handover initiation and cross-layer adaptation for mobile multimedia systems. IEEE Transactions on Multimedia, 2009, 11(6): 1131~1139
- 6 3GPP TS 36.331. E-UTRA Radio Resource Control (RRC); Protocol Specification (Release 9), 2010
- 7 Hui G, Legg P. Soft metric assisted mobility robustness optimization in LTE networks. Proceedings of International Symposium on Wireless Communication Systems, Paris, France, 2012: 1~5
- 8 Barbera S, Michaelsen P, Säily M, et al. Mobility performance of

LTE co-channel deployment of macro and picocells. Proceedings of IEEE Wireless Communications and Networking Conference, Shanghai, China, 2012: 2863~2868

[作者简介]



陈国平,男,南京信息职业技术学院讲师,中兴通讯学院兼职高级培训师,主要研究方向为移动通信、通信与信息系统。



杜庆波,男,南京信息职业技术学院教授,主要研究方向为移动通信技术、网络通信。

(收稿日期:2014-04-30)