



一种分布式 ACR 中 PPPoE 快速转发的实现

王高伟¹, 张吉兴², 肖志辉^{1,3}, 陈一山³

(1. 西南交通大学信息科学与技术学院 成都 610031;

2. 电子科技大学通信与信息工程学院 成都 611731;

3. 迈普通信技术股份有限公司 成都 610041)

摘要: 阐述了大规模接入汇聚路由器 (ACR) 面向三网融合应用多接入方式的重要意义。在三网融合的大背景下, 分析了 ACR 在网络中所起的作用。基于分布式系统主控与转发分离的思想, 分析了 PPPoE 的原理及流程, 并在 ACR 上实现 PPPoE。测试结果表明, 基于分布式系统实现的 PPPoE 能满足大规模接入应用场景, 使 ACR 支持高速宽带接入技术, 实现上千台整机并发。

关键词: 大规模接入汇聚路由器; 分布式系统; 以太网上的点对点协议

doi: 10.3969/j.issn.1000-0801.2014.03.016

Realization of PPPoE Fast Forwarding in Distributed ACR

Wang Gaowei¹, Zhang Jixing², Xiao Zhihui^{1,3}, Chen Yishan³

(1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China;

2. School of Communication and Information Engineering, University Science and Technology, Chengdu 611731, China;

3. Maipu Communication Technology Co., Ltd., Chengdu 610041, China)

Abstract: The large-scale involvement of the large-scale access convergence router was described for the significance of application of multiple access networks. In the background of network convergence, the role of ACR in the network was analyzed. Based on the thought of forwarding and control element separation (ForCES) of distributed system, the PPPoE protocol principle and process were analyzed, and the PPPoE access technology was achieved in ACR. The test results show that, the implementation of PPPoE based on a distributed system can meet the requirement of large-scale access application scene, and ACR can support high bandwidth for thousands of concurrent.

Key words: ACR, distributed system, PPPoE

1 引言

随着高清在线视频业务为代表的高带宽业务的不断发展, 高带宽业务成为当前互联网发展过程中的一个重要特点。为了吸引用户并提升用户黏性, 网络业务提供者都致力于不断提升用户的感官体验, 因此从长期趋势来看, 用户对网络带宽的需求将继续呈现爆发式增长趋势。当前网络带宽的提供能力、服务质量保障能力等面临巨大挑

战, 网络基础设施的承载能力必须能够与实际相匹配, 并且具有足够的弹性和自适应性才能满足需求环境的快速发展。在此背景下, 许多的新技术得以提出和发展, CDN (content distribution network, 内容分发网络)、网络存储等技术的实施一定程度上缓解了现实压力, 但未能从根本上解决问题。三网融合的提出及其实施, 使得 IP 网的承载能力面临着更大的挑战, 这一点在接入层面体现得更为突出。在融合网络环境中, 以高带宽视频业务为代表的新型



业务将以更快的速度出现,新的更加严峻的带宽矛盾将很快呈现。然而,传统的接入网长期存在接入层级多、设备种类多、宽带提供能力差、维护管理复杂等突出问题,使得“带宽不宽”的局面长期存在。网络用户不得不面对上网体验差、网络故障频发的无奈局面,而网络运营商面临着设备更新换代频繁、用户投诉多且故障处理效率低的困境。

在此背景下,源自 3Tnet 的大规模接入汇聚路由器 (large-scale access convergence router, ACR)^[1]具备高带宽提供能力等特点和优势,成为摆脱上述困扰的优先选择,为三网融合的深度发展提供了基础的设备保障。本文从接入层面上考虑,研究实现 ACR 中的 PPPoE (PPP over Ethernet, 以太网上的点对点协议)^[2]功能,以大规模大容量接入为目标,在分布式路由器上实现 PPPoE 接入认证。ACR 的研究与实现能有效降低网络的异构复杂度,使其用最少的互联级数,达到降低时延和抖动的目的,用带有远端用户接口单元,分合路器组成的树形分叉地域分布式系统架构,保证大规模的用户接入。在大规模汇聚接入条件下的交换节点具有更好的资源共享性和经济性、更低的用户平均响应时间,大规模接入有利于简化管理复杂度,降低运营成本,使边缘网络到用户接口的管理一体化。ACR 的网络位置如图 1 所示。

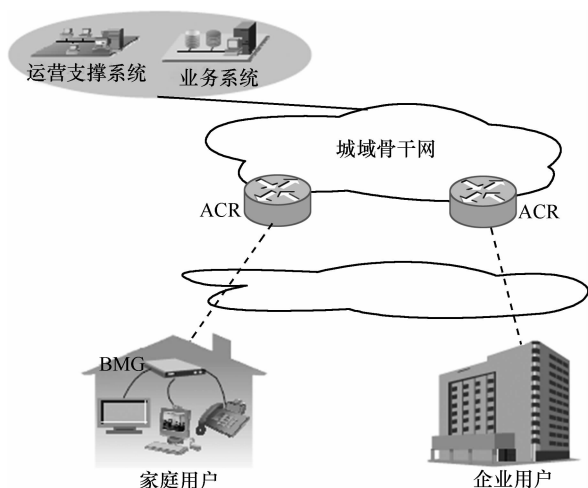


图 1 ACR 的网络位置

2 PPPoE

在众多的接入技术中,把多个主机连接到接入设备的最经济的方法就是以太网,而 PPP(point to point protocol, 点对点协议)^[3]可以提供良好的访问控制和计费功能,为了在以太网上利用 PPP 既有的访问控制和计费功能,于是就产生了在以太网上传输 PPP 报文的 PPPoE。

PPPoE 可以分为 3 个阶段,发现阶段、会话阶段和终止阶段。PPPoE 链路建立的前两个阶段如图 2 所示。

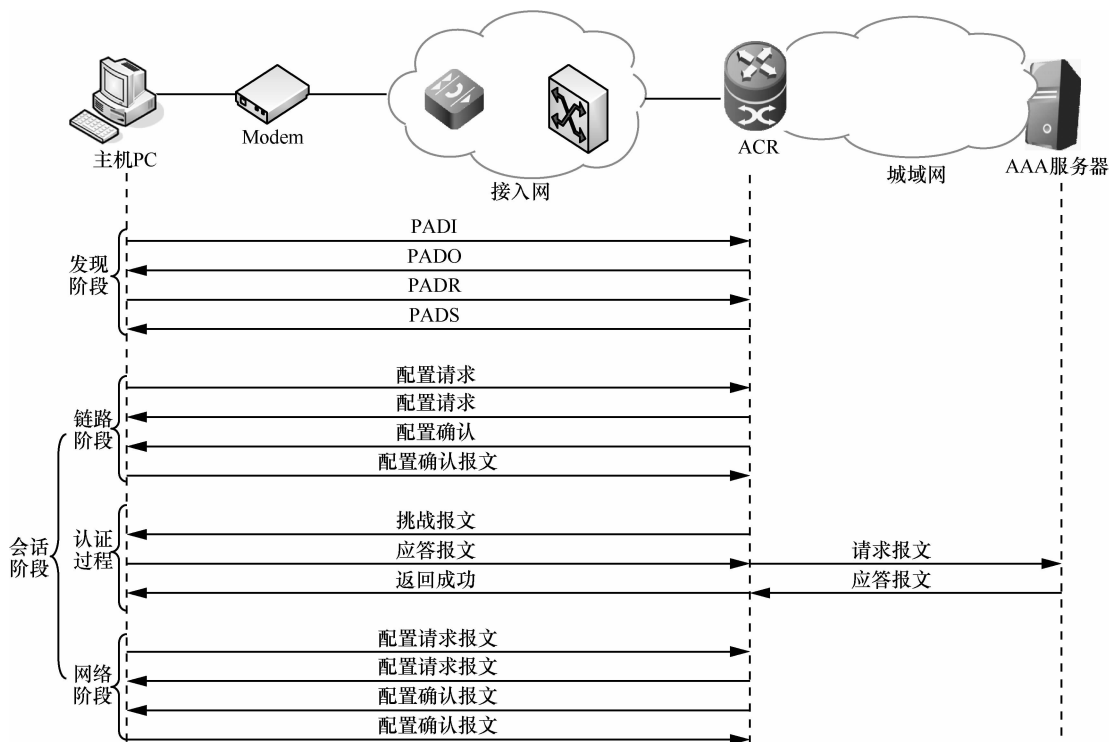


图 2 PPPoE 链路建立报文交互过程

2.1 PPPoE 的发现阶段

PPPoE 发现阶段共有 4 个部分组成。它们是由 4 条交互报文完成。首先,客户端 PC 利用广播发送一个 PADI 报文(PPPoE active discovery initiation packet),报文信息包含了告诉对端其想要获得的服务类型。这时,所有的访问集中设备都会收到 PC 端发来的 PADI 报文,并回复一个单播报文,告知 PC 端其能够提供它所要的服务。复杂的网络拓扑结构决定了 PC 会收到不同访问集中设备发来的 PADO 报文(PPPoE active discovery offer packet),此时 PC 把最先收到的 PADO 报文对应的访问集中设备作为自己的通信方,并单播发送一个 PADR 报文(PPPoE active discovery request packet)。访问集中设备收到报文并确定一个唯一的会话 ID,通过发送一个 PADS 报文(PPPoE active discovery session-confirmation packet)把会话 ID 发送给 PC,如果没有错误,就进入 PPPoE 的会话阶段。

2.2 PPPoE 会话阶段

PPPoE 会话阶段进行 PPP 的流程,如图 2 所示,下文不再赘述。PC 与访问集中设备依据 PPP 传送 PPP 数据报文,进行 PPP 的各项协商和数据传输。在会话阶段传送的数据分组中必须包含发现阶段所确定的会话 ID 并保持不变,然后开始后续通信。通常情况下,PPP 完成结束会话的控制,但在 RFC 文档中,定义了一个 PADT 报文(PPPoE active discovery terminate packet)用来终止 PPPoE 会话,PC 或访问集中设备在 PPP 会话开始后的任何时候发送该报文都可以结束会话。

2.3 PPPoE 终止阶段

PPPoE 通信的双方 PC 与访问集中设备会使用 PPP 自身提供的 PPP 终结报文来结束 PPPoE 会话。若不能使用 PPP 结束会话时,可以由 PC 或访问集中设备的一方发送 PADT 报文来实现。

3 ACR 总体架构

ACR 是指同时具有常规大容量高性能双协议栈核心路由器的宽带、IPv4/IPv6 组网功能和支持业务精细化管理、支持多种接入方式提供分级可靠 QoS 保障,实现极小汇聚层级的大规模接入汇聚功能的新型路由交换设备。因此,功能和性能双重因素使得它的设计非常复杂。

在复杂繁多的路由器体系标准下,IETF 提出标准的 ForCES^[4](转发与控制线分离协议)具有独到的观点,倍受许多专家和学者推崇。本文研究的 ACR 便是基于 ForCES

体系实现转发与控制分离的思想。分布式 ACR 体系结构如图 3 所示。这种体系具备高可靠性和高扩展性,利用网络解耦实现控制与转发分离,在系统中各单元网络互操作性得到了充分的体现。

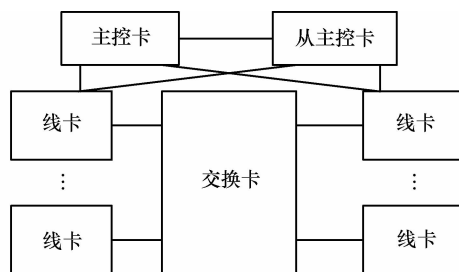


图 3 ACR 体系结构

在 ACR 采用的分布式结构中,交换和主控分离设计,交换背板带宽为万兆级,线卡报文处理器采用多核 CPU。因此 ACR 系统架构简单来说就是多核分布式高端路由器。主控卡采用高性能的单核 CPU 做主处理器,使用吉比特级以太网线路做各张线卡、交换卡的控制通道和主从备份通道,并配备大容量的内存做各种表项存储,负责控制平面的大部分工作,包括 PPPoE 控制报文的处理和转发信息的管理并利用分布式数据库,经适配层的处理后下发到各张线卡,从主控卡负责备份,并实时同步。若因外界原因发生故障时,ACR 可以快速切换到从主控卡,保证 ACR 的可靠性。交换卡则负责完成在两张线卡之间进行报文的交换功能。线卡由两部分组成,一部分是母板,一部分是子卡。母板上的主要部件是多核 CPU,它作为数据平面的报文处理器对接收到的数据分组进行转发和控制处理,并出万兆和吉比特以太网信号到两个子卡。子卡的主要功能是对外出各种 WAN 和 LAN 的接口,将多核报文处理器出来的以太网信号进行扩展和转换的部件,比如 FPGA 或者以太交换芯片。FPGA 将以太网信号转换为各种 WAN 接口;以太交换芯片将一个万兆级接口扩展为多个吉比特级接口。因此,线卡负责各种数据报文的快速转发、控制等功能。

在主控卡上,报文处理器负责控制报文的处理、转发表的管理并通过分布式数据库下发到各个线卡上。同时,主控卡应实时对从主控卡和线卡上表项同步、包括删除、插入、更新等操作,实现从主控卡和线卡表项信息的一致。在线卡上,转发表主要由通道号、接口索引及标志信息构成,线卡中的转发引擎通过查找转发表找到对应通道和端口号,并根据通道号和端口号来确定快速转发流程。

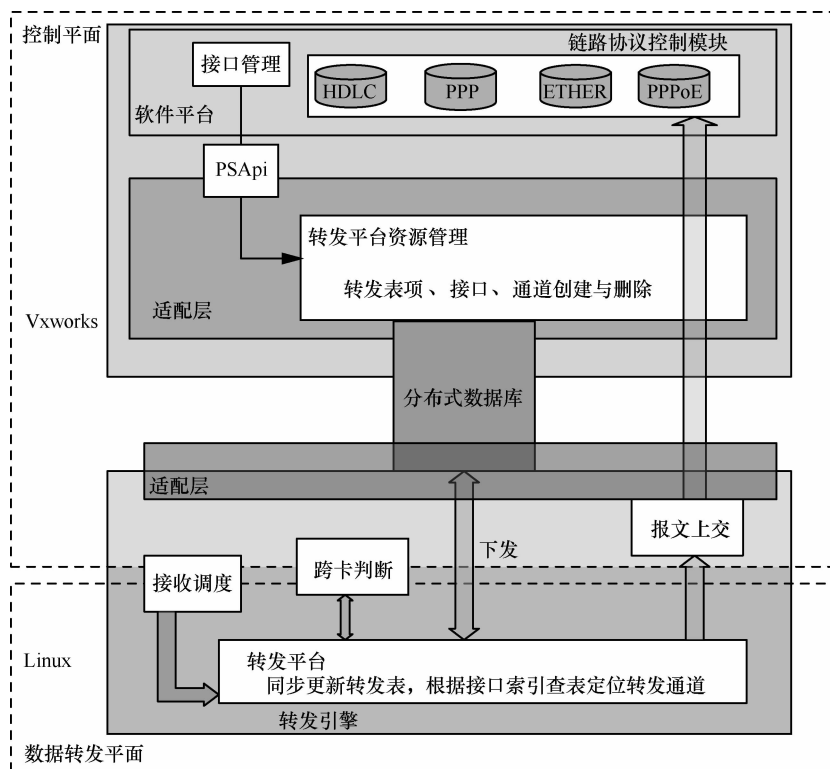


图4 分布式 ACR 转发控制结构

4 ACR 快速转发系统架构

4.1 转发控制架构

分布式 ACR 转发控制架构如图 4 所示。

控制平面上,链路转发平台为各链路协议提供通信接口来实现链路管理。链路协议控制报文(PPP、HDLC、以太网、PPPoE 及 MPPP)经过握手协商阶段学习到相关转发信息,创建并记录到转发表项中的信息通过转发表项信息管理模块进行处理。适配层既分布在主控卡也分布在线卡上,主控卡的适配层模块维护数据表,并通过分布式数据库,将信息分布到线卡上。线卡的适配层模块不再维护数据表,收到来自主控卡上的数据信息后,直接调用 PPPoE 对应的产品接口函数,更新产品软件接口的数据。例如,链路管理模块在主控卡上维护一张转发表,在线卡上链路协议模块收到分布的表项信息,直接调用链路管理模块提供的接口函数,更新转发表的信息。

如图 5 所示,在适配层上创建转发表后,通过分布式数据库组织转发表并下发到各个线卡,以此同时同步从主控卡上的转发信息。线卡上的转发表伴随主控卡上的下发信息同步更新。

在控制平面,转发信息主要有接口索引、通道号、槽位

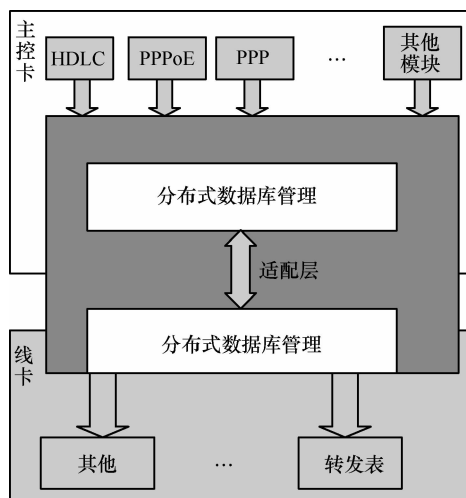


图5 转发表通过分布式数据库下发

号等信息组成转发表,并且以各个协议对应的转发表项以列句柄的方式组织。在转发层面,转发引擎利用查找转发表作为快速转发的依据,通过对接口索引、通道号实现快速查找匹配,找到相应的通道实现快速转发的目的。

4.2 分布式数据库下发方式

PPPoE 和转发相关的信息包括接口、槽位号、转发通道号、协议配置、协议状态及链路信息。主控卡上适配层转发表中的信息需要包含以上的信息,每个转发表项对应一

个接口。根据需要,转发表中还可以包含其他一些和链路转发过程相关的信息,如链路层的统计信息等。链路协议为分布式数据库提供转发表项信息,在适配层上创建转发表,由于分布式数据库作为数据库具备插入、删除和更新等功能,可以对转发表做实时的修改。只要转发表信息发生改变,链路管理模块就会把变更信息通过分布式数据库下发到线卡。

链路管理模块发出对转发表项的插入、删除和更新操作完成后,由分布式数据库把操作结果下发到线卡上,并查找对应线卡上的相关操作函数表实现同步更新,各个转发表项的参数与主控卡上传递的参数接口一致。对于链路协议模块而言,主控卡的链路模块只负责具体数据下发,线卡的对应模块只根据下发数据做出相应的处理。即当线卡收到主控卡下发的数据信息时,调用相关函数更新转发表项。数据平面就通过查找转发表项的数据状态,找到对应的端口和通道直接转发数据报文,而不再统一交给主控制卡去实现转发动作。这样便是由线卡的转发引擎实现快速转发,这种优势在多线卡的分布式路由器上体现得更加明显,多张线卡同步并行处理数据分组有效地分担了上层控制面 CPU 的工作,达到负载分担,相比集中式系统大大提高了转发性能。

5 快速转发的实现

5.1 主控卡 PPPoE 实现

主控卡上 PPPoE 模块在平台上注册相关的应用程序接口函数,并统一形成函数表。完成 PPPoE 接口产品软件 API 的创建、插入、删除和更新函数,在 PPPoE 数据转发阶段用到的虚拟接口(virtual access,VA)也是在此时完成创建。这是因为实际物理端口数目的限制,从而需要创建一个 VA 用来和对端交换数据。此时,系统会按照用户的配置,选择一个虚拟模板(virtual template,VT),根据 VT 的配置参数,动态地创建一个 VA。

PPPoE 在适配层上为转发平台查询、维护一张转发表,该表需要进行同步和下发。转发表项的具体信息包括接口索引、槽位号、通道号和同步标志等其他报文首部信息。PPPoE 模块所有适配层对平台提供的接口函数都采用异步的方式实现,即对适配层表项的添加、删除、更新操作,分布式数据库对表项的分发返回结果处理以及批量定时器超时处理都以事务的形式添加到适配层的任务中执行。

PPPoE 在适配层上需要对事务进行接口创建,反映到

分布式数据库上就是对转发表项的添加,此时要调用分布式数据库插入函数添加适配层表项。添加表项时,首先检查要添加的表项是否已经存在,如果存在则表示是从主从切换后执行,这种情况需要将同步标记更新为 false,并向分布式数据库返回插入成功。分布式数据库会根据该返回值进行表项的同步和分布操作。如果添加的表项不存在,则创建表项,并进行初始化各列的值。根据列句柄设置同步标记的值,如果是分布列句柄,则设置同步标记为 true,否则不设置。

PPPoE 适配层的任务调度执行 PPPoE 接口删除事务,执行接口删除。PPPoE 接口删除反映到数据库即为转发表项的删除,此时调用分布式数据库删除函数,删除适配层表项。删除表项时,首先检查要删除的表项是否存在:如果存在,则删除表项,向分布式数据库返回 OK,使分布式数据库继续后续的同步和下发;如果不存在,则可能是主从切换后的主控卡上执行脚本或者程序逻辑上存在问题。本文对这两种情况都统一向分布式数据库返回成功,但返回结果不会同步到从主控卡也不会下发到线卡上。

PPPoE 适配层平滑处理,发生主从切换后,新的主控卡会先执行脚本。PPPoE 模块根据脚本重新设置转发平台表项,并更新到适配层转发表中。如果更新的表项不存在则添加,如果存在则只更新同步标记为 false。系统通知平滑处理时,调用 PPPoE 模块的平滑处理函数进行适配层的平滑,这时只需要检查适配层表项中的同步标记为 true 的表项,并删除该表项。

5.2 线卡 PPPoE 实现

PPPoE 模块线卡的适配层上也是维护一张转发表,该转发表是主控卡上转发表的子集,它是由分布式数据库下发而来的,该转发表比较简单,只包含接口索引、槽位号和通道号的信息。

线卡上的适配层存储实现与主控卡上的适配层类似,由插入、删除、更新等接口函数,调用插入函数就可以完成线卡适配层转发表项的添加,此时没有同步标记,所以新增的表项需要设置转发表项的同步标记为 true。添加表项之前需要先检查新增表项的索引是否已经存在,当主从切换执行脚本或者主控卡执行批量时,会出现添加的表项已经存在的情况。如果表项存在,只需要更新同步标记为 true。线卡上转发表通过同步标记来进行平滑。

PPPoE 在数据转发流程中,不会使用到线卡转发表中的信息。因为设置了转发表项与转发通道关联,但实际不



会使用。

同理,调用删除接口函数,可以完成线卡适配层转发表项的删除。在线卡上适配层转发表项还关联批量下发事务,当新插入一张板卡时,线卡上适配层调用批量事务开始函数,遍历所有的转发表项,设置同步标记为 false(默认是 true)。当批量下发结束时,线卡适配层调用批量事务结束函数,因为在主控卡批量过程中,重新下发所有转发表项,线卡将下发下来的转发表项全部标记值更新为 true。批量结束时,检查所有转发表项。如果同步标记值仍然为 false,则表明主控卡上已经不存在该表项,此时线卡在批量结束时需要将同步标记值为 false 的表项删除。

PPPoE 在线卡上的产品软件接口子模块(PPPoESpi)主要完成数据平面报文的接收处理,对接收到的报文进行合法性检查以及修改报文描述符等工作。

5.3 PPPoE 数据报文转发

在快速转发平台的多核 CPU 转发引擎中,以太接口收到报文进行链路处理时,如果是 PPPoE 发现阶段的报文,则上交到主控卡进行处理,此时进入控制平面的数据接收流程。在输入处理的链路控制阶段的业务控制流程中,对 PPPoE 数据分组,去掉外层的 PPPoE 和以太分组头,换入接口为 VA,调用 VA 的链路处理(PPP 封装)和输入解析函数,开始完整的 PPP 处理流程。

PPPoE 数据报文快速转发流程如图 6 所示。发送阶段,在正常的路由阶段后,出口是 VA,判断跨卡,输出在相应卡上处理,再经 QoS 队列处理后进 VA 的逻辑发送,发送中封装 PPPoE 和以太分组头,进入以太接口的链路输出控制和发送处理。

VA 一般能确定绑定的物理口,故 VA 通道生成后,判

断是否跨卡。若跨卡,直接跨到对应绑定的物理接口所在的板卡上进行输出处理和封装、发送处理。当 VA 绑定的物理口变化时,转发平台应该响应并更新在通道结构中对应的值。

5.4 PPPoE 软件模块设计

整个 PPPoE 模块在分布式系统下主要软件实现函数从上到下由 PPPoESapi、PPPoEMpuPsal、PPPoEDdbMpu、PPPoEDdbLpu、PPPoEDfp、PPPoELpuPsal、PPPoESpi 模块组成,如图 7 所示。

PPPoESapi
PPPoEMpuPsal
PPPoEDdbMpu
PPPoEDdbLpu
PPPoEDfp
PPPoELpuPsal
PPPoESpi

图 7 PPPoE 分布式系统软件模块

PPPoESapi 模块:对转发平台的 PPPoE 业务模块提供操作产品软件的接口,通过类似函数表的方式挂接各个接口来实现。

PPPoEMpuPsal 模块:PPPoE 业务模块适配负责类似参数转换、资源操作等处理,以注册的方式挂接到 Psapi。为控制平面提供 API,提供查询接口 API,完成自身的数据表项同步和下发。

PPPoEDdbMpu 模块:对简单的转发表操作,包括创建、插入、删除、查询,实现数据的分发、包括从主控卡分发到

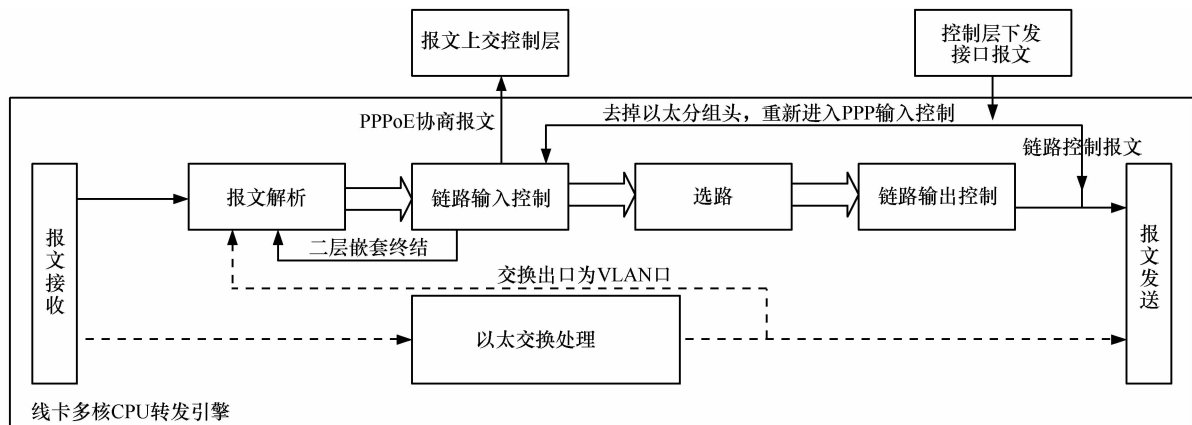


图 6 PPPoE 数据报文快速转发流程

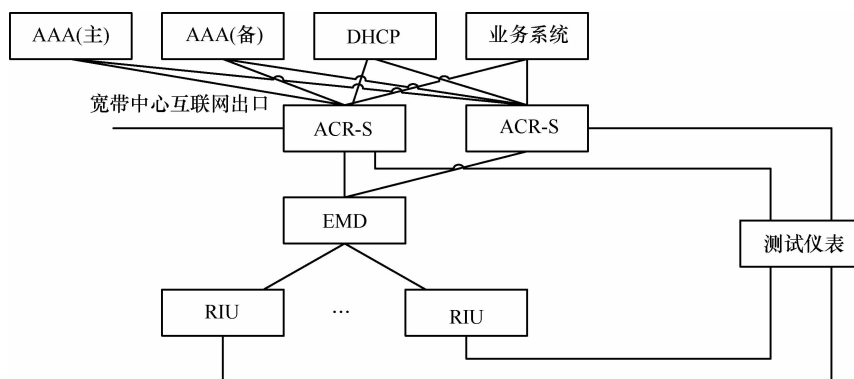


图 8 测试拓扑结构

从主控和线卡。

PPPoEDdbLpu 模块:实现分布式数据库存储接管函数。

PPPoEDfp 模块:提供报文的接收调度、入队和出队的接口及队列报文的调度。

PPPoELpuPsal 适配层模块:不再维护数据表,收到来自控制平面的数据信息后,直接调用 PSI 的函数,更新 PSI 的数据。

PPPoEPsi 模块:完成线卡转发表项的添加删除操作,完成对数据转发平面报文接收处理。

6 测试报告

本次测试在上海工程技术大学宽带中心一楼实验室内完成,包括大规模接入汇聚路由器相关测试、ACR 主机测试、ACR 远端模块测试和 ACR 宽带接入服务及业务控制测试,具体测试拓扑如图 8 所示。

测试配合设备采用机顶盒与电视机,测试仪器采用思博伦公司的 Spirent Test Center。本文涉及关于以太网端口上 PPPoE 接入测试时,验证 ACR 支持的最大链接数(PPPoE 会话连接),测试结果符合要求。

7 结束语

本文首先描述了分布式路由器架构,重点分析了 ACR 的转发控制架构,详细介绍分布式数据库数据下发的关键实现。主控卡上实现链路协议模块相关接口的注册,适配层上完成转发表项的创建,并利用分布式数据库实现同步和下发。主控卡和网络层交下来的控制报文送到线卡上去发送,PPPoE 数据报文经过链路处理、输入控制、选路输出控制后,实现快速转发。测试结果显示,ACR 完全可以满足大规模接入场景应用,而且通过后期升级,可以达到更高的报文处理速度。

参考文献

- 1 孟赞,刘九评. ACR 设备在广电网络内的应用部署设计. 有线电视技术,2013(6):47~52
- 2 RFC2516. A Method for Transmitting PPP over Ethernet, 1999
- 3 RFC1661. The Point-to-Point, 1994
- 4 孙宗峰,肖志辉,孙健. 基于分布式路由器的 IPv4-IPv6 转发控制架构研究. 电信科学,2012(6):42~47
- 5 RFC894. A Standar for the Transmission of IP-Data Grams over Ethernet Networks, 1984
- 6 Zeng X Y, Cheng C Q. Discuss of user identification technology in DSL network. Proceedings of 2009 International Symposium on Intelligent Ubiquitous Computing and Education, Chengdu, China, 2009

[作者简介]



王高伟,男,西南交通大学信息科学与技术学院硕士生,主要研究方向为电子与通信工程。

张吉兴,男,电子科技大学通信与信息工程学院硕士生,主要研究方向为通信与信息系统。

肖志辉,男,迈普通信技术股份有限公司首席执行官、高级工程师,西南交通大学硕士生导师,中国通信标准化协会(CCSA)理事,主要研究方向为路由器系统及下一代网络架构。

陈一山,男,迈普通信技术股份有限公司高级工程师,主要研究方向为路由器系统架构及网络。

(收稿日期:2014-02-28)