

层次式交换机设计与实现

游军玲^{1,2}, 何海燕¹

(1 北京理工大学 管理与经济学院, 北京 100081; 2 中国科学院计算机网络信息中心, 北京 100190)

摘 要: 介绍了如何创新性地利用可编程硬件, 设计和实现的层次式交换机, 层次式交换机支持数据层面、控制层面和管理层面分离, 体现层次式交换技术协议特性, 同时与现有网络兼容, 且灵活支持各种创新架构的实现, 满足未来网络设备可编程、灵活性等特性, 代表了未来网络领域关键设备的一种发展方向。

关键词: 未来互联网; 层次交换网络; 层次式交换机; IPv6; CNGI 部署

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2014)03-0150-04

Design and Implementation of Hierarchically Switch

YOU Jun-ling^{1,2}, HE Hai-yan¹

(1 School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China;

2 Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper will focus on the design and implementation of Hierarchically Switched Network switch (based on programmable network hardware). The HSNET has data plane, control plane, and management plane separated which reflects the characteristics of HSNET, meanwhile, HSNET is fully compatible with the legacy network and supports the implementation of more innovative architectures which provides the flexibility for the Future Internet and stands for the trend of the development of critical network devices in the Future Internet.

Key words: future internet; hierarchically switched network; hierarchically switch; IPv6; CNGI deployment

1 引言

未来网络体系结构思想来源于网络实践, 并需要在实际环境中验证和改进, 然而, 现有的商业设备系统封闭性过高, 接口开放性和可编程性支持不够。因此, 研究未来互联网体系结构必须同时考虑到未来网络核心设备——路由器和交换机。未来网络交换机作为承载未来互联网体系结构的基础设备, 其研制倍受重视, 当前的研究趋势是, 体系结构研究和设备研制并行, 设备体现新架构的特性, 新体系结构在设备上开展测试和验证。互联网协议将不断演变, 要求交换机必须灵活可编程, 以适应未来互联网的新型架构^[1-4]。

在未来网络关键设备研发方面, 国内已走在国际

前沿水平, 研发出多种先进的未来网络设备: 中科院计算所提出了一个支持未来网络创新的可编程虚拟化路由器平台, 采用通用计算机和专用数据包处理卡相结合的方式构建可编程虚拟化路由器; 国防科技大学采用 FPGA 和多核处理器相结合, 研制了可重构路由器; 解放军信息工程大学研制了一种能支持多种网络体系的可重构路由交换平台; 中科院网络中心基于网络处理器, 设计和实现层次式交换机, 并将新体系结构和设备部署于 CNGI 真实环境中。

2 层次式交换网络体系结构

以革命性思路构建的层次交换网络的基本思想是: 网络拓扑结构按层次结构构造, 地址空间按层次结构分配, 通过拓扑结构层次和地址结构层次的严

收稿日期: 2013-05-23; 修回日期: 2013-07-12

基金项目: 科技部“九七三”课题(2012CB315803); 感知中国先导专项课题(XDA06010200); 国家科技支撑计划课题(2012BAH01B03); 科技部“八六三”课题(2011AA01A101); 发改委 CNGI“未来互联网试验床建设”项目

格匹配,以层次交换代替传统路由.新体系结构具有如下优点:避免了频繁的路由信息交换和复杂、费时的路由表查询;网络传输路径固定,网络行为确定,网络可管理;提供有效的服务质量保证;提供 ISP 设备地址空间与用户地址空间的分离,防止对核心网络设备的攻击;防止源 IP 地址的仿冒,有利于对不良网络行为的追踪定位;网络能力、信道容量和节点交换能力可扩展等.

3 层次式交换机硬件系统

层次网络交换机机架提供了一个基于 ATCA 标准的模块化硬件平台,板卡逻辑关系如图 1 所示.

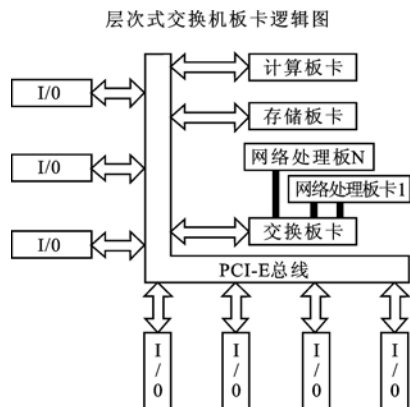


图 1 层次式交换机板卡逻辑图

交换机的主要板卡介绍如下:

- (1) 存储板卡:数据可以在其他的板卡和存储板卡之间相互传递,有效地扩展系统的数据存储能力.
- (2) 交换板卡:交换板卡用于将各个网络处理

板卡连接在一起,可以将从一个输入端口进入的数据高速地传输到一个或者多个输出端口.

(3) 计算板卡:该板卡有一个或者多个高性能 CPU,具有较强的计算处理能力,用于执行处理器密集型的任务.

(4) 网络处理板卡:又称为业务板卡,用于执行层次式交换机的核心转发功能.

(5) I/O 板卡:用于执行密集型数据转发任务.

其中,最为关键的是网络处理板卡的选择,层次式交换机选择基于 Intel IXP2800 网络处理器开发平台^[5]. Intel IXP2800 网络处理器的体系结构主要包括以下几个部分:

(1) Intel® XScale™ 核:32 位的精简指令集通用处理器.用来初始化和整个网络处理器.

(2) 微引擎:专门针对于网络应用设计的 16 个专用可编程处理器.微引擎负责数据平面对每个数据包执行的主要操作,运行频率高达 1 400 MHz,可以配置为 4 或者 8 线程模式.微引擎内部结构主要包括微码控制存储器、本地存储器、通用寄存器、传输寄存器、相邻寄存器等功能单元.

(3) DRAM 控制器:三个独立的 RDRAM 控制器.

(4) SRAM 控制器:四个独立的 QDR SRAM 控制器.

此外,Intel IXP2800 网络处理器还包括 Intel® XScale™ 外围设备、多功能时钟、SHA-1 单元、媒体和交换矩阵接口、PCI 控制器等,如图 2 所示.

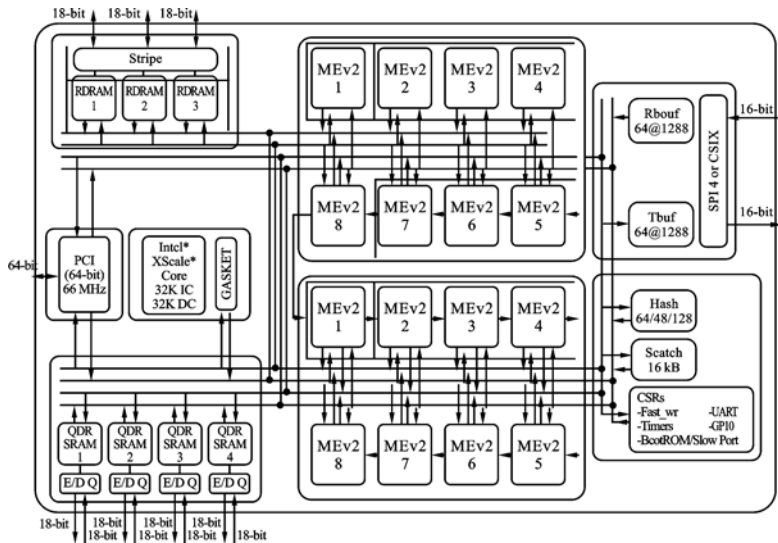


图 2 Intel IXP2800 网络处理器硬件结构图

基于 IXP 2800 开发的层次式交换机硬件结构,兼备高性能和灵活性特点,能够满足层次交换网络

对高速数据处理能力的要求,其特点包括:通过对执行环境、内存、硬件加速器、总线结构以及网络应用

开发接口的综合优化,具有灵活的软件体系,并提供硬件级处理性能,能够面向高速数据流的高层细化分组处理;开发方便,具有 WorkBench 集成开发环境,支持 C 语言和微代码,具有图形化界面,能够进行硬件平台软件模拟调试,同时提供丰富的 IP 协议 API、子程序、应用范例。

4 层次式交换机软件系统

4.1 软件系统层面划分

层次式交换机软件系统设计具体包括数据层面、控制层面和管理层面^[7-8]三部分,各层面彼此隔离。

数据层面负责线速处理和转发数据包,实现数据平面协议,数据层面的实现遵循层次交换网络标准,利用网络处理器的多微引擎、多线程的微码模块技术完成数据包接收,包分类、隧道封装/解封装、逻辑端口计算、负载均衡、层次封装/解封装、数据包内部转发、队列管理、队列调度、组播转发、区分服务以及数据包发送功能。

控制层面处理协议消息,建立、配置和更新转发表,实现控制平面协议,主要完成层次式交换机的自动配置,以及与其他交换机交互生成各种交换表用于数据转发平面。控制层面是一个分布式系统,主要功能运行在控制卡上,部分功能运行在其他板卡,通过搭建网络处理器 Xscale 核开发测试环境,在嵌入式操作系统 Montavista OS 下进行控制层面模块的设计与开发,实现数据包转发表生成、边界端口表生成、系统自动故障恢复、网络链路状态管理、节点域内事件通知、泛洪法转发数据包、定时器、数据包封装、事件分类等功能。

管理层面的功能是配置系统、收集统计信息等。管理层面包括 6 种操作类型: GetRequest, GetNextrequest, GetBulkRequest, SetRequest, Trap 和 InformRequest, 完成网络管理服务器与被管对象之间的交互、网络管理服务器与层次网内部主机间的交互以及外网主机与网管服务器的交互。具体实现是将 TCP 协议移植到层次网数据层面慢速通道上,实现简单网络管理协议(SNMP)的接口。管理层面的实现,使网络管理人员能够方便地使用 SNMP 协议远程管理和配置层次网交换机。

4.2 数据流程设计

在层次交换新架构网络中,共有三种类型数据包: IPv4, IPv6 和标记交换数据包。当未来网络中有其他类型数据包到来时,通过层次标记交换,简单修

改交换机的微代码,即可支持未来网络中的其他类型数据包。如下为交换机对数据包的处理过程^[8]。

(1)当数据包经 I/O 板卡进入层次交换机的网络处理板卡之后,数据包接收模块首先将数据包从媒体与交换接口中的接收缓存中转移到 DRAM 存储器中,对数据包进行重组,并在 SRAM 中构造数据包的描述符。然后将包描述符的句柄通过 Scratch 队列传递到分类和标记模块。

(2)层次式交换分类和标记模块对数据包先后进行区分服务代码点分类和 6 元组分类,然后交给测量和整形模块进行处理。如果是组播数据包,分类和标记模块则会将数据包交给组播转发模块进行处理。

(3)组播转发模块通过查找组播转发信息库确定从哪几个端口将数据包转发出去,并生成组播包复制请求,通过该请求来要求组播包复制模块进行数据包复制,复制完成后,数据包交给测量和整形模块以完成对该组播包的区分服务处理过程。

(4)数据包从测量和整形模块出来以后,进入层次网络数据包转发模块,该模块是实现的核心模块,负责按照层次交换网络交换控制过程的规定转发数据包,包括逻辑端口号计算、负载均衡和内部转发等处理。转发模块处理完毕之后将数据包交给队列管理和调度模块。

(5)队列管理模块负责完成对硬件队列的管理,调度模块通过调度算法确定数据包的发送顺序。调度模块将数据包出队请求发送给队列管理模块,队列管理模块将相应数据包出队之后向数据包发送模块提交发送请求。

(6)发送模块首先将完整的数据包切分成为一个个的元数据包,将元数据包逐一发送;在将完整的数据包发送完成之后,通过网络处理器提供的一种特殊的通信机制——反射写操作——通知调度模块数据包已完成发送。

数据流程模块如图 3 所示。

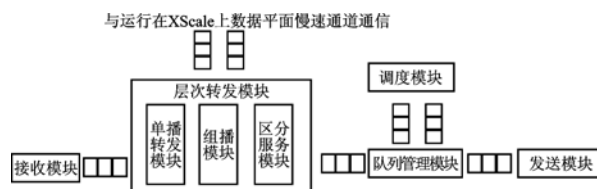


图 3 数据流程模块图

4.3 微引擎任务分配及任务优化

微引擎是层次交换机上最重要的硬件单元之一,对微引擎的分配和合理利用关系到设备的性能。IXP2800 网络处理器 16 个微引擎,分为 2 组,每组

有专用的命令和数据总线;每个微引擎可支持4或者8个硬件线程,线程之间零开销切换。

基于IXP 2800网络处理器的层次交换机软件系统实现,首先要分配每个模块占用的微引擎,由于在一个微引擎上运行的接收模块最高支持4 Gb/s的线速处理速率,设计使用两个微引擎来提高接收模块的数据处理能力。基于同样的原因,用两个微引擎来实现发送模块。但接收模块和发送模块在微引擎的组织方式不同。接收模块的两个微引擎采用的是功能流水线的方式,第一个微引擎执行数据包接收的任务,第二个微引擎执行数据包重组的任务,二者合起来完成整个数据包的接收功能;发送模块的两个微引擎采用并行运行的方式,每个都单独地完成整个数据包的发送功能。队列管理模块和调度模块则分别运行在单独的微引擎上。

为了使层次交换机各级存储系统均衡,按照对存储空间和访问性能要求的不同进行分配。微引擎内部的CAM和本地存储器存储容量小而访问次数高,用于隧道封装信息的缓存;数据包头缓存数据结构位于本地存储器中,将对DRAM的访问操作降低为两次;逻辑链路表占用524 280字节,访问次数较低,存储在容量大的DRAM存储器上。多级存储体系有效降低对存储器的访问延迟。

硬件加速器是层次交换机的重要硬件资源。利用硬件加速器,可以有效地将计算任务从微引擎或者XScale上分摊到负载较小的其他硬件单元中,降低微引擎和XScale的计算负载,如负载均衡模块使用哈希单元,避免使用微引擎进行复杂的哈希计算;队列管理模块使用SRAM控制所提供的硬件队列有效地加速对队列的管理和调度过程;各个微引擎之间使用Scratchpad所提供的硬件队列使相互之间的通信更为便捷。

5 仿真测试

Developer Workbench是一个带有源码编辑、编译、调试、仿真等功能的集成开发环境,提供强大的调试和仿真功能。试验平台构建如图4所示,层次式网络节点域由4台交换机组成。每台交换机具有4个物理端口。编号为1的交换机具有两条下行物理链路和两条内部链路;编号为2的交换机具有一条下行链路、两条内部链路和一条短接链路;编号为3的交换机具有一条短接链路、一条上行链路和两条内部链路;编号为4的交换机具有一条下行链路、两条内部链路和一条上行链路。节点域地址前缀用

二进制表示为:001(FP) 00000000000001(TLA) 00000000(RES) 0001(SF1) 00000000(SF2) 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 0000。节点域前缀长度为28,S为28,L为8。在编号为1的交换机上进行原型试验。

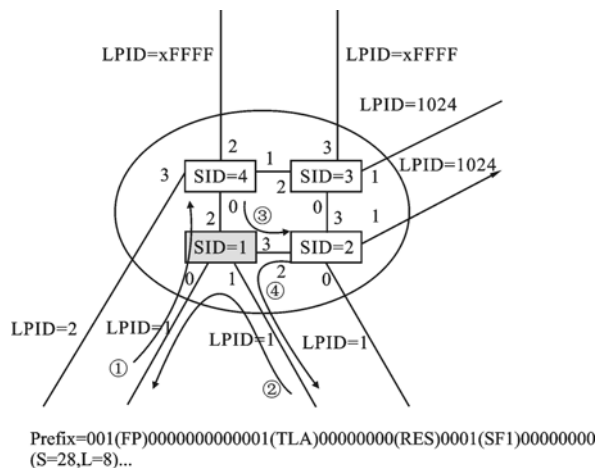


图4 层次交换网络仿真模型

当端口速率为10 Gb/s时,层次交换机的各个功能单元的利用率最大没有超过70%,内部总线利用率远低于30%,层次交换机微引擎利用率最大值没有超过80%,丢包率为0,支持线速转发。

又利用安捷伦N2X对层次式交换机进行数据包转发的性能测试,测试依据为RFC2544和RFC2460,测试中主要选择了64字节、512字节、1024字节以及1518字节等常用的以太网帧长度,测试结果表明:层次式交换机支持线速处理,丢包率为0,平均延时稳定。

6 层次式交换机部署

层次交换网络体系结构易于部署,不要求对现有IPv4或者IPv6网络作任何协议的修改,支持多架构共存,协同工作。新架构和设备目前部署在中国下一代互联网示范工程CNGI的网络环境中,如图5所示,包括两级层次结构,顶级前缀域负责汇接功能,由四台层次式交换机组成,通过GE/POS接口与CNGI核心IPv6骨干网或者IPv4骨干网相连。每个次级前缀域配置一台层次式交换机,用户网络通过GE接口接入。层次网络内部不需部署路由协议,次级前缀域下联的用户接入网,可以为纯IPv6用户接入网或者IPv4/IPv6双栈用户接入网。

(下转第158页)

主机之间提供了一种通用的通信策略, 监控主机的控制台软件和目标机的代理程序通过中间层的通信协议完成彼此之间的通信, 能够较好地屏蔽不同通信接口之间的差异性, 并且很方便的实现对底层通信接口和上层应用功能的扩展。

参考文献:

- [1] 杨永生, 施笑安. 基于串口嵌入式操作系统的远程监控系统[J]. 西安科技大学学报, 2005, 25 (2): 204-207.
- [2] 底素然, 张伟功, 巴峰, 等. 面向星载数据管理系统的 SpaceWire 总线仿真系统开发[J]. 微电子学与计算机, 2012, 29(1): 164-167.
- [3] 叶晓国, 孙力娟, 王汝传. 适应性的混合卫星通信网络 MAC 协议[J]. 计算机工程, 2010, 36 (1): 7-11.
- [4] 何鹏举, 戴冠中, 傅磊, 等. 远程网络化测控系统研究与实现. [J]. 微电子学与计算机, 2006, 23 (7): 196-202.
- [5] 胡毅, 于东, 李培楠, 等. 基于现场总线的开放式数据系统的设计与实现. [J]. 小型微型计算机系统, 2008, 29

(9): 1745-1749.

- [6] 张志军, 郭渊博, 刘伟, 等. 基于 m 叉树与 DH 协议的组密钥协商协议[J]. 计算机工程, 2010, 36 (1): 161-163.
- [7] 黄智, 王科, 杨沿. 平车身舒适系统总线协议研究[J]. 计算机工程, 2009, 31 (4): 331-335.
- [8] 张伟功, 朱晓燕, 关永, 等. 基于微包协议的三模冗余容错计算机无缝重构算法[J]. 计算机科学, 2009, 36 (6): 286-290.
- [9] 张伟功, 辛明瑞, 张永祥, 等. SPARC V8 监控软件中总线协议无关通信方法: 中国 CN102215230A [P]. 2011.

作者简介:

李 静 女, (1990—), 研究方向为嵌入式计算机系统结构与容错技术.

周继芹 女, (1978—), 博士研究生. 研究方向为嵌入式计算机系统结构与容错技术.

朱晓燕 女, (1967—), 高级工程师. 研究方向为嵌入式计算机系统结构与可靠性技术.

(上接 153 页)

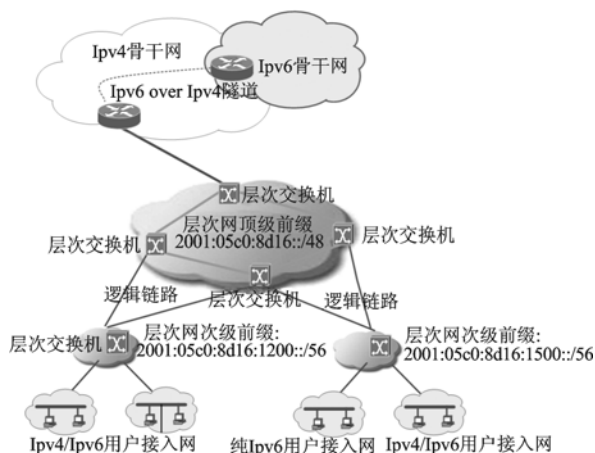


图 5 基于层次交换技术的 CNGI 试点部署图

7 结束语

本文设计并实现基于网络处理器的层次式交换机, 给出交换机的硬件实现方案和软件模块工作流程. 层次式交换机支持数据层面、控制层面和管理层面相分离, 全面测试和验证层次交换体系结构特性, 无需路由交换实现数据包转发, 且方便部署于现有网络中, 兼容 IPv4/IPv6 网络, 层次式交换机的可编程、灵活性等特性, 为未来网络设备研制提供了一种参考。

参考文献:

- [1] 张宏科, 苏伟. 新网络体系基础研究——一体化网络与普适服务[J]. 电子学报, 2007, 35(4): 593-598.
- [2] 吴建平, 任罡, 李星. 构建基于真实 IPv6 源地址验证体系结构的下一代互联网[J]. 中国科学, 2008, 38(10): 1583-1593.
- [3] 刘强, 汪斌强, 徐格. 基于构件的层次化可重构网络构建及重构方法[J]. 计算机学报, 2010, 33 (9): 1557-1568.
- [4] 钱华林, 葛敬国, 李俊. 层次交换网络体系结构[M]. 北京: 清华大学出版社, 2008.
- [5] Inter Corporation. Inter® IXP2800 Network Processor Hardware Reference Manual[R]. 2008.
- [6] Inter. Inter® IXP2800 and IXP2850 Network Processors Datasheet[EB/OL]. [2013-04-20]. <http://www.intel.com/design/network/>, 2005.
- [7] Inter Corporation. Intel® IXDP28X1 Advanced Development Platforms System User's Manual[R]. 2005.
- [8] Intel. Intel® Internet Exchange Architecture Software Building Blocks Applications Design Guide[M/CD]. USA: Intel Corporation, 2005.

作者简介:

游军玲 女, (1981—), 硕士. 研究方向为网络体系结构.

何海燕 女, (1963—), 博士, 教授. 研究方向为系统工程、战略管理.