扩展服务路由器操作维护系统的研究*

江 勇, 吴建平, 徐 恪, 喻中超

(清华大学 计算机科学与技术系,北京 100084)

E-mail: jyong@csnet1.tsinghua.edu.cn

http://www.tsinghua.edu.cn

摘要: 路由器操作维护(operating and maintenance,简称OAM)系统负责对路由器进行操作和管理.它是路由器正常运行的保证,是路由器中的重要模块.随着路由器技术的发展,对路由器软件动态升级的要求越来越受到人们的重视.对扩展服务路由器操作管理进行了深入的研究,首先介绍了扩展服务路由器操作管理的设计要求和研究现状,然后介绍了清华大学研制的扩展服务路由器原型系统的软、硬件体系结构及其对操作维护系统的功能要求,设计并实现了可实时动态加载扩展服务组件的操作维护管理系统.最后指出了进一步的研究方向.

关键词: 操作维护管理;扩展服务;路由器

中图法分类号: TP393 文献标识码: A

路由器管理的功能包括性能管理、配置管理、网络流量统计管理、错误恢复管理和安全管理这5大功能,在参考文献[1,2]中具体定义了这些要求.

随着Internet的发展,不断出现新的网络协议和对原有协议的扩展.除了转发分组之外,路由器还应该提供一些新的服务:

- 集成服务和区分服务;
- 增强的路由能力(包括第3层和第3层以上的路由及其交换技术);
- 安全功能(例如,能够实现虚拟私有网络VPN)和包过滤(即防火墙);
- 对现有协议的增强(例如,类似于随机早检测RED的拥塞控制算法);
- 新的核心协议(例如,IPv6).

我们把提供了这些扩展功能的路由器称为扩展服务路由器ESR(extended services router),而把传统的路由器称为尽力发送路由器(best effort router).

清华大学在承担的国家"九五"重点攻关项目——"高性能多协议路由器"[3]的基础上结合当前网络互连设备技术的发展,设计和开发了模块化可扩展的ESR原型机系统.本文将详细介绍该扩展服务路由器原型系统中的操作维护管理子系统的设计与实现.该子系统处于路由器软件体系结构的顶层.由于扩展服务路由器对OAM(operating and maintenance)子系统的软件体系结构和功能有特殊的要求,如对模块化可扩展组件的支持、扩展组件的动态加载、系统可扩展性和实时配置管理等等,在综合目前国际上一些这方面的技术的基础上,我们提出了自己的设计思路,并对具体实现进行了讨论.

收稿日期: 2000-01-27; **修改日期**: 2000-03-24

基金项目: 国家自然科学基金资助项目(69822002,69822003);国家863高科技发展计划资助项目(863-306-2D-07-01)

作者简介: 江勇(1975—),男,重庆人,博士生,主要研究领域为计算机网络体系结构,高性能交换结构,调度算法;吴建平(1953—),男,山西太原人,博士,教授,博士生导师,主要研究领域为计算机网络体系结构,计算机网络协议测试,形式化技术;徐恪(1974—),男,江苏淮阴人,博士,讲师,主要研究领域为计算机网络体系结构,计算机系统性能评价;喻中超(1977—),男,湖北武汉人,硕士生,主要研究领域为分组分类与调度算法.

这种类型的操作维护管理系统在国内还没有见到相关的报道.而国际上关于同类型的相关报道也正处于研究阶段,如微软用于 Windows NT 4.0 和 5.0 服务器的路由和远程访问服务 RRAS (routing and remote access service)[4],该系统的优点是易于使用并且易于配置,但是 RRAS 的可扩展性很有限,位于内核中的网络子系统不能扩展;Sony 公司开发的 ALTQ[5]是用于分组调度的体系结构,它是基于 FreeBSD 内核实现的,有用于带宽预留的 CBQ(class based queuing[6])和 DRR(deficit round robin)模块,但不允许动态加载模块;普林斯顿大学的可扩展路由器项目利用基于路径(path)的 Scout 操作系统[7]实现了可扩展的路由器体系结构[8],它实现了 DRR[9],Virtual Clock[10]和 WFQ[11]等算法,但由于实现一个好的操作系统需要做的工作太多,Scout 在系统稳定性和完备性方面还有待加强;华盛顿大学的研究小组提出了路由器插入程序(plugin)体系结构[12],用于对集成服务路由器进行功能扩展,该体系结构是基于 NetBSD 实现的,只支持对流进行配置,不支持分布式路由器的端口配置,并且它的分类器的性能不高,难以用于高性能路由器.

本文第 1 节介绍 ESR 的软、硬件体系结构.第 2 节介绍 ESR 中 OAM 子系统的功能要求及其总体设计.第 3 节详细介绍支持模块化和可扩展的 OAM 子系统的设计与实现.第 4 节总结全文并指出进一步的研究方向.

1 扩展服务路由器(ESR)的软、硬件体系结构

1.1 ESR 的硬件体系结构

ESR 采用了与我们所研制完成的高性能多协议路由器相一致的基于分布式路由的多处理器 3 级体系结构.这种 3 级体系结构如图 1 所示,各级功能如下:

(1) 中央处理器模块(CPM)

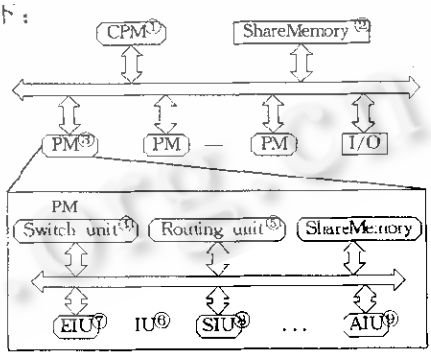
- 处理路由协议包
- 维护全局路由表
- 保持局部路由表与全局路由表的同步
- 完成对路由器的操作配置与网络管理功能
- 维护全局安全数据库

(2) 扩展处理器模块(PM)

- 使用局部路由表完成 IP 包的转发
- 使用安全数据库完成 IP 包的过滤
- 支持多种网络接口协议

(3) 接口单元(IU)

- 完成对物理通信端口(PU)的驱动
- 支持高密度的通信端口



①中央处理器模块,②共享内存,③扩展处理器模块,④交换单元,⑤路由单元,⑥接口单元,⑦以太网络接口单元,⑧串行接口单元,⑨并行接口单元.

Fig. 1 High performance multi-protocol router hardware architecture

图1 高性能多协议路由器硬件体系结构

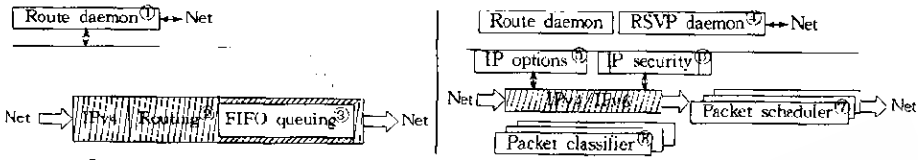
1.2 ESR 的软件体系结构

ESR 是实现支持多种路由和网络管理协议,并提供服务质量以及 IP 安全等多种扩展服务功能的高性能主干网络路由器.

提供扩展功能的扩展服务路由器 ESR 和传统的尽力发送路由器在软件体系结构上有很大的不同.如图 2 所示为这两种路由器的比较.

与尽力发送型路由器相比,ESR 内核增加了如下的组件:包调度器、包分类器、安全机制、基于

服务质量的路由和交换、防火墙以及拥塞控制.



①路由服务,②路由,③FIFO队列,④RSVP服务,⑤IP选项,⑥IP安全,⑦调度器,⑧分类器.

Fig. 2 Best effort vs ESR

图2 尽力发送和扩展服务路由器

在路由器中,不同的算法和组件的不同实现在性能、系统开销等方面都存在着不同.而大部分算法都在不断发展,需要经常替代和升级,因此,这样的路由器系统组件具有相对的"流动"性,它们应该和其他保持相对稳定的子系统加以区分.我们把这些相对稳定的部分叫做固定模块或核心模块.扩展服务路由器的模块化组件,即使它们实现的是同样的功能(比如用于分组调度的不同调度算法),也需要经常在系统中共存,例如可能需要在一个端口采用一种调度算法,而在另一个端口采用另一种算法.

在本文中,我们提出了一种 ESR 软件体系结构来满足下列要求:

- 模块化:以模块形式实现特定的算法,我们称为可扩展组件 EC(extensible component).
- 可扩展性:新的模块应能实时动态加载.
- 灵活性:EC 句柄可被创建、配置和绑定特定流.
- 高性能:系统应该提供非常有效的数据通路,无数据拷贝,不附加中断过程,增加的系统开销不会严重影响系统性能.

2 操作维护管理子系统的功能要求和总体设计

简单地说,操作维护管理子系统的基本功能是对路由器进行维护管理.

在 ESR 中,操作维护管理子系统处于最高一层,也是系统最先启动的任务,由它再启动加载路由器的支撑子系统.路由子系统及接口单元等模块,并初始配置各模块和分配系统资源.对于路由器之外的管理者而言,它是网络监控、管理和配置的惟一途径.相对于路由器内部的各个模块而言,它是系统的总控模块,必须进行各层协议实体的配置.控制和管理,因而这个模块的功能相当复杂,对于路由器及其网络运行的安全、可靠和稳定起着重要的作用[13~16].

2.1 操作维护管理子系统的功能要求

考虑到我们的目标是建立一个模块化可扩展的路由器系统,它必须具有网络流的概念,并能实现基于流的组件选择.因此,我们认为,OAM 子系统除了包括文献[1,2]所定义的基本功能以外,还应该具有以下的增强功能:

(1) 在路由器系统中实时动态加载和卸载可扩展组件 EC.EC 是一个实现特定扩展服务功能的代码模块.

(2) 创建可扩展组件 EC 的灵活而独立的实例.一个实例是一个独立的 EC 运行实体.在内核中经常存在一个 EC 的多个实例.为了保证不同 EC 之间的互操作并提供简单的配置接口,需要定义标准消息机制.

(3) 在数据分组和流之间可有效地进行映射,绑定流和 EC 组件.过滤规则用来区分流,它用 6 个域来区分:

〈源地址,目的地址,协议,源端口,目的端口,进入接口〉

每个域都可以是通配符.另外,网络地址可以用掩码部分通配相应的区域.

(4) 保证高性能.我们通过尽量少的数据拷贝、快速而有效的分类策略、数据通路流水作业等多种策略保证增加的系统开销尽量小.分类算法采用了 RFC(recursive flow calssification)算法[17],并结合了高速缓存机制,分组流的数据传输只在第 1 个分组绑定 EC,而其后的分组均可快速流水传送.

2.2 操作维护管理子系统的总体设计

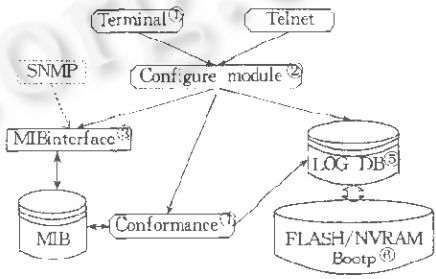
OAM 的总体设计如图 3 所示.

下面,我们从控制信息通路的角度介绍扩展服务路由器 OAM 模块的总体设计.

不同组件间的控制信息通路如图 4 所示.

• IPv4 核心模块.IPv4 核心模块包括处理分组的几个可以流水工作的组件,这些组件不能动态加载,它们完成与网络设备的交互以及将分组分派到可扩展组件 EC 进行处理.

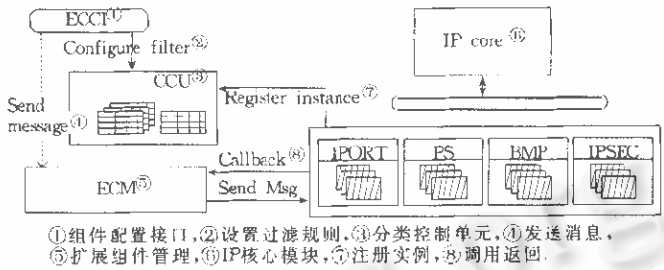
• EC 组件.图中显示了 4 种不同的 EC——IP 选项(IPOPT)、分组调度组件(PS)、进行最长匹配的组件(BMP)和 IP 安全组件(IPSEC).



①串口终端,②管理配置模块,③MIB接口,④一致性控制,⑤日志,⑥后援数据库.

Fig. 3 OAM sub system architecture

图3 操作维护管理子系统结构



①组件配置接口,②设置过滤规则,③分类控制单元,④发送消息,⑤扩展组件管理,⑥IP核心模块,⑦注册实例,⑧调用返回.

Fig. 4 Control path in ESR

图4 扩展路由器的控制信息通路

• 可扩展组件管理器 ECM(extensible component manager).ECM 管理 EC 组件,向 EC 实时可靠发送其他核心模块和子系统的消息.

• 分类控制单元 CCU(classify control unit).分类控制单元实现分组分类,并在流和 EC 之间建立相应的关系.

• EC 配置接口 ECCI(EC configure interface).EC 配置接口是用户对系统进行配置的接口单元.ECCI 通过命令行和类似于 Shell 的描述文件对路由器 EC 配置信息库(CIB)进行控制配置,而系统根据 EC 配置库(CIB)中的数据对每流要绑定的 EC 组件进行配置.

同时,扩展服务路由器采用的是基于分布式路由的多处理器 3 级体系结构.为此,我们在操作维护管理子系统中设计了一种分布式的主从管理结构(如图 5 所示),在中心处理模块(CPM)中运行 OAM 子系统的主管理进程,而在各个扩展处理模块(PM)中均有一个服务管理进程.主管理进程和服务管理进程之间通过内部总线或者内部交换网络进行通信.

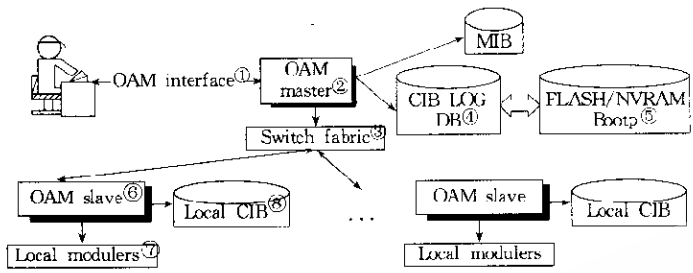


Fig. 5 OAM distributed architecture
图5 OAM分布式结构示意图

3 ESR 中操作维护子系统的设计与实现

3.1 EC 组件和组件管理器 ECM

我们使用名称和类型标识一个 EC 组件,由于组件不同,其实现可能很简单(如 IP 选项组件只有十几行代码),也可能很复杂(如分组调度算法). 目前的原型系统支持 4 种类型的 EC 组件:IP 选项、IP 安全、分组调度和最长匹配(作为 CCU 的分类器). 将来我们计划支持 QoS 路由、网管统计、拥塞控制和防火墙等扩展服务. 我们在扩展服务路由器中实现了两种分组调度器:用于公平排队的不充分轮转法(deficit round robin,简称 DRR[9])和基于类的分组调度的层级公平服务曲线(hierarchical fair service curves,简称 H-FSC[18]). 由于篇幅关系,这些组件的具体实现方法在本文中略去.

扩展组件管理器 ECM 管理一个存有 EC 组件类型、名称和调用函数等信息的表. EC 一旦加载到系统中,首先在 ECM 中注册调用函数,其后的所有对 EC 的控制管理都要经过 ECM. 它必须负责向 EC 组件发送消息和处理异常情况.

3.2 分类控制单元 CCU

CCU 在操作维护管理子系统中是实现模块化扩展服务最重要的部分. 它实现了分组分类器、快速流检测和提供实例与过滤规则之间的绑定. 它管理着过滤规则表(filter table)和快速流表(flow table)这两个重要的数据结构.

快速流表的实现基于哈希表. 我们用分组报文头的 5 个域〈源地址,目的地址,协议,源端口,目的端口〉来计算哈希表的索引. 哈希表的数组在系统启动时就分配好了,它的大小依赖于路由器的运行环境,我们的缺省值是 32 768.

下面,我们来介绍过滤规则表的 RFC 算法实现. 过滤规则表用来对分组进行分类,有几种通用的分组过滤算法[19~21],它们都是灵活并且功能强大的,但其实现非常复杂,在高速情况下性能较差. 而我们需要的是能够匹配 6 个域〈源地址,目的地址,协议,源端口,目的端口,进入接口〉的快速查找算法.

最后,我们用 RFC 算法来实现查找最好匹配的过滤规则.

RFC 包括预处理(preprocessing)和查找(look-up)两个部分. 预处理通过软件方式实现. 查找既可以通过软件实现,也可以通过硬件实现,当然,二者的性能有很大的差别. 为了保证通用性并考虑到软件实现也已经能够保证我们需要的性能,我们选择了全部用软件实现 RFC.

在以软件方式实现分类算法时,在最坏的情况下(采用三阶段查找),整个查找过程的代码执行时间为($140clks+9t_m$),其中 $clks$ 为时钟周期,t_m 为访存时间.

3.3 数据分组的处理流程

在我们的 ESR 原型系统中,数据分组被送到实现特定功能的 EC 组件实例进行处理,因为数据分组在数据通路中的处理流程直接影响到系统性能,下面我们叙述一个 IP 分组的处理步骤.若该流不在快速流表里:首先,硬件接口收到分组,送到 IP 核心模块;IP 核心模块找到相关的 EC 组件句柄并调用 CCU,参数为分组指针和 EC 组件名称标识;CCU 先在快速流表里查找,失败后查找过滤规则表(由于查找步骤是系统开销较大的地方,需要快速而有效的查找实现机制);找到后,CCU 将句柄指针存入快速流表,并返回给 IP 核心模块一个句柄;调用 EC 组件实例,处理分组;重复以上步骤,继续处理分组.

处理了第 1 个非高速缓存流的分组后,由于有了快速流表,则可以通过一条更快捷的通路进行处理:分组直接从快速流表里获取一个记载有要处理该分组的相关 EC 组件句柄的流索引,然后从快速流表里取得相应的 EC 句柄对分组进行处理.若快速流表里的某条表项在一段相当长的时间里一直空闲(没有收到新的分组),则流表里的该表项将被删除或被替代.这种实现机制具有高度模块化和最小系统开销的特点.另外,由于系统开销只与第 1 个分组的查找有关,故可扩展性好.

3.4 配置信息库 CIB 及后援数据库的管理

在扩展服务路由器系统启动时,需要从配置信息库 CIB 中读入配置数据以配置各模块的初始运行参数,如底层网络接口的配置参数,同时对过滤规则集和扩展组件 EC 进行配置管理.在系统运行过程中,用户对路由器配置及网络拓扑结构等进行配置后需要将数据存入一个专有的数据结构.因为这部分数据是保证路由器正常工作的关键数据,而且必须要存储在 NVRAM 上,所以这部分数据由操作维护管理子系统专门控制管理,以提供其他模块协调一致运行的数据结构.

3.5 扩展组件配置接口 ECCI 的实现

用户和网络管理员面对的 ECCI 界面是一个命令行格式的命令解释器.它可以接收用户的命令,对命令进行分析,也可以接收一个配置文件.这个分析过程实际上是一个编译过程,编译的输出结果是网络管理者可以识别的网络管理请求消息,扩展组件管理模块 ECM 对此命令请求信息进行相应的路由器管理操作.

4 总结与展望

本文的主要贡献是基于扩展服务路由器的模块化体系结构,对操作维护管理的要求设计和实现了对模块化扩展组件的支持、模块组件的动态加载、系统可扩展性和实时配置管理等操作维护管理系统,该系统可扩展性和灵活性好,实现了分组的快速分类,在过滤规则数较多的情况下,一个 IPv4 分组的分类只需要最多 9 次内存访问,在过滤规则数较少时需要的内存访问次数更少.它能够对扩展服务路由器进行全面的管理和维护,充分保证了路由器的动态可扩展性和可靠运行.

路由器操作维护管理系统对路由器的运行和维护至关重要;同时,网络互连设备新技术的产生和应用必然会对操作维护管理提出新的要求.基于以上考虑,操作维护管理在以下几个方面可以进行进一步的研究:

(1) 进一步适应新的网络协议的要求.随着新的网络技术和网络协议的提出,出现了一系列对操作维护管理的新要求.即使对于目前已经存在并且已经比较成熟的协议,也有待于进一步研究如

何使用 MIB 信息库描述协议的状态和控制协议的行为,从而降低冗余度.

(2) 进一步提高网络的安全性能.网络的发展所带来的问题已经不仅仅是一个技术问题,而逐渐成为一个社会问题.改进的方法通常是利用通信密码学的研究成果,这就要求必须能够对密钥进行维护和管理.另外,也需要加强认证机制,以实现安全的操作维护管理.

(3) 把操作系统的管理功能和路由器的操作维护管理统一起来.

(4) 把主动网络机制[22,25]引入操作维护管理.

References:

[1] Baker, F. Requirements for IP version 4 routers. RFC1812, 1995.

[2] Senie, D. Changing the default for directed broadcasts in routers. RFC2644, 1999.

[3] Fan, Xiao-bo, Lin, Chuang, Wu, Jian-ping, *et al*. Performance model and analysis of a distributed router. Chinese Journal of Computers, 1999,22(11):1223~1227 (in Chinese).

[4] Microsoft Corporation. Routing and remote access service for Windows NT server. White Paper, 1997. http://www.microsoft.com/ntserver/zipdocs/rras.exe.

[5] Cho, K. A framework for alternate queueing: towards traffic management by PC-UNIX based routers. In: Douglis, F., ed. Proceedings of the USENIX 1998. New York: IEEE Computer Society Press, 1998. 137~145.

[6] Floyd, S., Jacobson, V. Link-Sharing and resource management models for packet networks. IEEE/ACM Transactions on Networking, 1995,3(4):365~386.

[7] Mosberger, D. Scout: a path-based operating system [Ph.D. Thesis]. Department of Computer Science, University of Arizona, 1997.

[8] Peterson, L.L., Karlin, S.C., Li, Kai. OS support for general-purpose routers. In: Druschel, P., ed. Proceedings of the HotOS Workshop. New York: IEEE Computer Society Press, 1999. 38~43.

[9] Shreedar, M., Varghese, G. Efficient fair queuing using deficit round robin. In: Wecker, S., ed. Proceedings of the SIGCOMM'95. New York: ACM Press, 1995. 231~242.

[10] Suri, S., Varghese, G., Chandranmenon, G. Leap forward virtual clock. In: Hasegawa, T., Pickholtz, R., eds. Proceedings of the INFOCOM'97. New York: ACM Press, 1997. 557~565.

[11] Demers, Keshav, Shenker. Analysis and simulation of a fair queuing algorithm. In: Chanson, S.T., ed. Proceedings of the SIG-COMM'89. New York: ACM Press, 1989. 1~12.

[12] Decasper, D., Dittia, Z., Parulkar, G., *et al*. Router plugins: a software architecture for next generation routers. In: Neufeld, G., ed. Proceedings of the ACM SIGCOMM'98. New York: ACM Press, 1998. 133~140.

[13] Baker, F., Watt, J. Definitions of managed objects for the DS1 and E1 interface types. RFC1406, 1993.

[14] Kastenholz, F. Definitions of managed objects for the ethernet-like interface types. RFC1643, 1994.

[15] Bierman, A. Iddon, R. Remote network monitoring MIB protocol identifiers. RFC2074, 1997.

[16] Waterman, R., Lahaye, B., Romascanu, D., *et al*. Remote network monitoring MIB extensions for switched networks (Version 1.0). RFC2613, 1999.

[17] Pankaj, Gupta, Nick, Mckeown. Packet classification on multiple fidels. In: Chapin, L., ed. Proceedings of the SIGCOMM99. New York: ACM Press, 1999. 47~53.

[18] Stoica, I., Zhang, H., Ng, T.S.E. A hierarchical fair service curve algorithm for link-sharing, real-time and priority services. Computer Communication Review, 1997,27(4):249~262.

[19] Bailey, M., Gopal, B., Pagels, M., *et al*. Pathfinder: a pattern-based packet classifier. In: Lepreau, Jay, ed. Proceedings of the 1st Symposium on Operating Systems Design and Implementation. Monterey, CA: IEEE Piscataway, 1994. 115~123.

[20] Engler, D., Kaashoek, M. DPF: fast, flexible message demultiplexing using dynamic code generation. In: Estrin, D., Floyd, S., eds. Proceedings of the SIGCOMM'96. New York: ACM Press, 1996. 53~59.

[21] Mogul, J.C., Rashid, R.F., Accetta, M.J. The packet filter: an efficient mechanism for user-level network code. In:

Menees, S., ed. Proceedings of the 11th ACM Symposium on Operating Systems Principles. New York: ACM Press, 1987. 39~51.

[22] Tennenhouse, D., Smith, J. M., Sincoskie, W. D., *et al*. Survey of active network research. IEEE Communication Magazine, 1997,35(1):80~89.

[23] Alexander, D., Arbaugh, W., Hicks, M., *et al*. The switchware active network architecture. IEEE Network Magazine, 1998,12(3):29~36.

附中文参考文献:

[3] 范晓勃,林闯,吴建平,等.分布式路口器的性能模型与分析.计算机学报,1999,2:(11):1223~1227.

Research of the Operating and Maintenance System in Extended Services Router*

JIANG Yong, WU Jian-ping, XU Ke, YU Zhong-chao

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

E-mail: jycng@csnet1.cs.tsinghua.edu.cn

http://www.tsinghua.edu.cn

Abstract: The OAM (operating and maintenance) module of a router implements the management and manipulation of the router. The OAM is one of the most important sub systems in a router, guaranteeing it to work in going order. With the rapid rate of router technic development, it is becoming increasingly important to dynamically upgrade router software in an incremental fashion. In this paper, the authors make a thorough research on router operating and maintenance system. First, the ESR-OAM designed requirements and evolutions are introduced. Then the extended services router architecture and ESR-OAM function requirements are presented, which is developed by Tsinghua University. And the ESR-OAM is designed and implemented in detail, which can dynamically add and configure extensible component at run time. At last, research directions and open problems in this area are discussed.

Key words: OAM (operating and maintenance); extended service; router

* Received January 27, 2000; accepted March 24, 2000

Supported by the National Natural Science Foundation of China under Grant Nos. 69822002, 69822003; the National High Technology Development Program of China under Grant No. 863-306-2D-07-01