服务元网络体系结构和微通信元系统构架

曾家智, 徐 洁, 吴 跃, 李毅超, 胥 能
(电子科技大学计算机科学和工程学院, 四川成都, 610054)

摘 要: 本文通过对现有网络中服务类别的分析、归纳, 针对现有的分层网络体系结构存在的层间功能重叠和复杂的分层处理过程所带来的网络服务效率低下的问题, 提出了一种基于服务元的网络体系结构. 服务元只提供服务, 不接受服务, 所以避免了层间交互和服务传递的开销. 服务元不仅能为本节点应用提供服务, 而且不同节点的服务元可以合作向某一节点或整个网络提供服务. 本文给出了微通信元系统构架, 它是一种易于从 TCP/IP 过渡的服务元网络体系结构的构架. 作为微通信元的服务元被组织成微通信系统, 大量的微通信系统被组织成网络系统. 微通信元系统构架具有简洁、可扩展和容易实现的特点.

关键词: 网络体系结构; 层次结构; 面向对象; 服务元; 微通信元
中图分类号: TP393 文献标识码: A 文章编号: 0372-2112 (2004) 05-0745-05

Service Unit Based Network Architecture and Its Micro-Communication Element System

ZENG Jia-zhi, XU Jie, WU Yue, LI Yi-chao, XU Neng
(School of Computer Science and Engineering UESTC, Chengdu, Sichuan 610054, China)

Abstract: Based on the analysis of network services and the deficiency in current layered network architectures caused by multi-layer functional duplication and complicated processing, this paper proposes a Service Unit based Network Architecture (SUNA). Service Unit (SU) only provides services, and does not receive any services, thus it avoids inter-layer interactions and costs of delivering services. SU provides services for local node applications, and SUs of different nodes can also collaborate to provide services for the entire network or a certain node. Furthermore, the Micro-communication Element System (MCES) is presented for SUNA, which is easy to be transitted from TCP/IP. Micro-communication System is constructed of organized Micro-communication Elements which are equal to SUs, and network system consists of many Micro-communication Systems. MCES is easy to design and implement and has greater flexibility.

Key words: network architecture; layered architecture; object-oriented; service unit; micro-communication element

1 引言

从 20 世纪 90 年代开始, 在国际网络界就进行了许多关于高性能的网络体系结构的研究. 例如, D Clark 和 D Tennenhouse 在 1990 年提出了面向网络协议处理性能优化的应用级组帧(ALF)的网络体系结构思想[1], 试图消除传统 OSI 参考模型中由于高层协议分层过多而造成协议软件处理性能较低的不足; D Tennenhouse 等人在 1996 年提出了可以在单个分组上进行资源分配和调度的高性能网络模型))) 主动网络技术[2], 试图消除传统 Internet 对所有分组采用单一资源分配和调度的模式; 1997 年 A Lazar 提出可以根据应用需要, 定制网络服务的可编程网络模型, 试图改变传统网络对所有应用只能提供固定服务的静态模式[3]. 上述研究都是基于传统的层次结构网络, 对网络的性能进行改善, 但难以解决层次结构自身存在的问题.

Stefan Boecking 提出的 MCS(Modular communication system)构架[10~14]是一种具有代表性的面向对象的网络体系结构. 其出发点在于: 满足不断涌现的应用对于网络的不同性能和服务质量的需求.

2002 年 10 月, Braden 等人为了解决现有网络层间交互和难于扩展新的服务的问题, 提出了一种无层次的基于角色的网络体系结构[15], 并给出了角色的模型.

2 现有网络层次结构存在的问题

提出层次网络的出发点在于简化协议设计的复杂性. 层由实体(硬件和/或软件)构成, 能够接受下层提供的服务, 并能向上层提供增值服务. 层具有封装性、隐蔽性和抽象性.

目前, 实用的网络体系结构都是层次结构, 例如 OSI、TCP/IP、SPX/IPX 和 ATM 等. 其中, OSI 参考模型被公认是最严格的. 尽管 TCP/IP 比 OSI 等协议具有较高的效率, 但是功能

收稿日期: 2003-04-04; 修回日期: 2003-12-16
基金项目: 国家自然科学基金(No. 69871005)

冗余重复[4~9], 影响了它在宽带网络中的应用.

n 层向 n+ 1 提供的服务 $S_{n/n+1}$是 n21 层向本层提供的服务 $S_{n-1/n}$和本层自身提供的服务 S_n 的并集, 记为:

$$S_{n/n+1} = S_n \cup S_{n-1/n} \quad (1)$$

进行叠代后有:

$$S_{n/n+1} = S_n \cup S_{n-1} \cup S_{n-2}, \ S_2 \cup S_1 \quad (2)$$

由式(2) 可知, n 层网络协议所提供的服务是第 1 层到第 n 层各层自身提供服务的并集. 为了提高效率, 各层服务能力不应该具有交集, 因为交集表示各层服务功能重复. 交集是不提供增值服务的.

TCP/ IP 存在的低效率问题在于各层的服务能力具有大量的交集. 分别讨论如下:

(1) 多层检错和一层检错相比, 并不能提高检错能力.

检错能力定义为能检出错误的集合. 如果多层检错, 则总检错能力为各层检错能力之并集. 例如, 最常见的 TCP/ IP 以太网中, MAC 层采用 32 位 CRC 检错; IP 层和传输层各自采用 16 位校验和检错, 其总的检错能力仅仅等于 MAC 这一层的检错能力. 正是因为/ IP 校验和的计算是 Ipv4 的一个主要开销0[10], 所以经过激烈的争论 IPv6 取消了校验和. 遗憾的是在对峙争论中获胜的一方仅仅解决了/ IP 校验和多余0的问题, 而不可能解决/ TCP 校验和也多余0的问题, 因为他们的任务仅仅是制定网络层的 IPv6. 此外, 采用多层检错时, 各层的包头检错的次数少于包的内容. 显然, 包头的错误比包内容的错误危害更大. TCP 和 UDP 的伪头似乎弥补了这个缺陷, 但关键在于并未增强检错能力.

(2) 地址重复降低效率

例如 IP 地址和 MAC 地址都是一个节点的地址. 由于两者不一样才出现了 ARP(Address resolution protocol) 和 RARP(Reverse address resolution protocol) 等协议及其开销. 组播也存在着 IP 地址到 MAC 地址的映射问题.

(3) 分片(segment) 问题

链路层、网络层和传输层都要处理, 不仅多次开销, 而且无法避免各层之间的交互.

(4) TCP/ IP 协议栈本来就是为了窄带文本数据而开发的, 所以仅仅网络层服务类型具有分级分类标志, 而且早期并未使用. 后来随着宽带网络和多媒体技术的发展, 为了保证 QoS, 在各层(甚至在层间) 打了一系列的补丁: 资源预留协议 RSVP、实时传输协议 RTP、实时传输控制协议 RTCP、IEEE802. 1D 协议、区分服务 DiffServ 和多协议标签交换 MPLS 等. 这些技术相互重复且不一致, 例如各层的优先级位数和类型就不一致. 关键在于优先级和类型本身就不适合于层次结构, 如果各层处理不一致, 则会造成混乱; 如果一致, 则会重复处理效率低. 如果不分层, 只需一次处理端对端的 QoS 即可.

(5) 上述功能重复还造成包头(首部) 增长, 传输效率降低.

3 层的地址和端地址

现有的网络体系结构, 无论是层次结构还是 MCS 构架都认为层间服务访问点 SAP 由层的地址来标识. 问题是层的地址只有两种: 节点(主机和路由器) 地址(网号和网内节点号) 和端口号. 所以如果存在着二层以上的网络模型, 则标识 SAP 的地址就会重复. 例如 TCP/ IP 中, 传输层和应用层的地址都是端口号, 网络接口层的 MAC 地址和网络层的 IP 地址都是节点地址.

在新的网络体系结构中, 将抛弃层的地址的概念, 而采用端到端地址. 端到端地址就是由节点地址和端口号构成的序偶. 即

$$\text{端到端地址} = (\text{节点地址}, \text{端口号}) \quad (3)$$

使用地址时, 可以使用端到端地址整体, 也可以使用其一个部分(节点地址或端口号).

4 服务功能元素和服务

每一个基本网络服务功能称为服务功能元素, 而服务功能则定义为服务功能元素的集合. 纵观各种网络, 其服务功能元素归纳如表 1 第一列所示. 某些服务元素只在路由器中完成, 例如 OSPF、RIP、EGP、BGP、DVMRP 和各种路由递交等功能. 某些服务功能元素只能由主机完成, 例如分片功能, 对于宽带网络而言, 路由器和路由交换机等网络设备是不应参与分片的. 应该采用类似 Ipv6 的方法, 当源主机所发包的尺寸大于某网络的允许值时, 网络设备向源主机返回允许的包尺寸, 源主机重发允许尺寸的包.

表 1 网络的各种服务功能元素

类型	服务功能元素	对应包的域	对应服务元
网络接口类	位的编码和解码 验错 成包和拆包 MAC 介质分配	基本硬头和硬校验	NIC 服务元
实时 QoS 类	* 资源监控 (类 RTCP)	广义 ICMP 专用域	广义差错控制服务元
	资源预留 (类 RSVP)	类 RSVP 专用域	资源预留服务元
	* 优先级处理	仅基本软头	优先级处理服务元
	防抖动(类 RTP)	类 RTP 专用域	类 RTP 服务元
无连接服务类	发送数据	基本硬头含校验	无连接发数据服务元
	接收数据	同上	无连接收数据服务元
有连接服务类	建立连接	连接专用域 (含校验)	建立连接服务元无
	释放连接		释放连接服务元
	保序、流控、拥塞控制(卡纠错)		有连接发数据服务元 有连接收数据服务元
安全类	加解密 数字认证	ESP 专用域	安全净荷服务元
		AH 专用域	认证服务元
路由类	OSPF RIP EGP BGP DVMRP 单播递交 组播递交 源选径递交	报文在数据域	OSPF 服务元
		报文在数据域	RIP 服务元
		报文在数据域	EGP 服务元
		报文在数据域	BGP 服务元
		报文在数据域	DVMRP 服务元
		仅基本首部	单播递交服务元
		仅基本首部	组播递交服务元
		源选径专用域	源选径递交服务元
其他类	ICMP, IGMP, 分片, RTCP	报文在数据域	各种广义差错 控制服务元

注: 服务元一般都要使用基本首部, 所以表中一般不再列出.

5　服务元的模型

服务元网络体系结构也是模块化结构,模块是服务元.服务元是能够提供服务而又隐藏内部细节的最小实体(硬软件).服务元不接受服务,只提供服务.

服务元提供服务是通过服务数据单元 SDU 完成的. SDU 又称为包 Packet.服务元是 SDU 的发送者(源)、接收者(目的)、转发者(递交)或变换者.按照启动服务的方式和与 SDU 的关系,服务元可以分为五类,对应模型如图 1 所示.其中,第一、二、三和四类服务元是 SDU 的源和目的,用矩形表示.矩形下方的下箭头和上箭头表示服务元按规定顺序发送或接收的一系列 SDU.

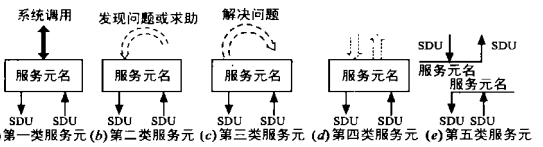

图 1　五类服务元模型

第一类服务元由于执行系统调用而启动服务.矩形上方粗箭头表示执行系统调用.粗箭头的方向表示应用的信息流向:上箭头表示接收;下箭头表示发送.例如应用执行系统调用 write(),启动有连接发送数据服务元,它把应用发送的数据分成小块组成一系列的包发送,还要接收一系列的确认包.粗箭头的方向向下.又例如应用执行系统调用 connect(),启动(主动)建立连接服务元,通过三次握手建立连接.下方的上、下箭头表示先发出第一次握手的包,再接收第二次握手的包,后发出的第三次握手的包.如果应用为了接收而执行 connect()则粗箭头的方向向上.第一类服务元是为本节点(应用)提供服务的.对于没有 OS 的节点,系统调用将被 API 函数取代.

第二类服务元因网络发生不正常事件或请求帮助而启动服务,并主动向某节点发警告或求助信息.第三类服务元由于收到此警告或求助信息而启动服务,进行内部处理. I 节点的第二类服务元和 J 节点的第三类服务元协作向 I 节点或 J 节点提供服务.例如路由器的第二类服务元向源主机的第三类服务元发数据格式错的信息.又例如 I 节点 PING 服务元求助 J 节点 PING 服务元进行可通性测量.

第四类服务元周期性地启动或收到相关包启动.其服务通过按规定顺序发送和接收的一系列的包完成.并且通常是通过包的组播方式进行收发.例如路由选择协议服务元,动态地为网络各自路由器填写路由表.路由递交类型服务元接收的到包按路由表递交到相应端口.第四类服务元用于各相关节点协作为整个网络系统提供服务.

第五类服务元由于 SDU 的到来而启动服务并对 SDU 进行变换后输出.三角形上方的下箭头表示源于本节点 SDU 的到来,输出的 SDU 由三角形下方的下箭头表示;三角形下方的上箭头表示源于其它节点 SDU 的到来,产生的 SDU 由三角形上方的上箭头表示.例如压缩解压服务元、身份验证服务元、安全净荷服务元和 NIC 服务元等.由于第五类服务元功能是两两互逆的,我们将互逆的服务元的两个三角形画在一起.

第五类服务元中,谁的输出作为谁的输入并不是固定的,只要收、发方匹配即可.特例是 NIC 服务元,它总是发送包所经过的本节点的最后一个服务元.它一方面通过介质访问控制把本节点的包转换为 Bit 流放在网络介质上,另一方面从网络介质抓取 Bit 流到本节点转换成包.考虑到:(1)网络接口层功能完全由网卡完成,而服务团队其它服务元基本由主 CPU 完成,二者可以并行;(2)由于网络接口层只提供服务,并不接受服务,所以网卡既能作为层次结构的最下层,又能作为新结构中的一个 NIC 服务元.

6　服务元网络体系结构的节点模型

服务元网络体系结构的节点模型如图 2 所示.

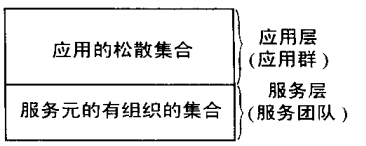

图 2　服务元网络体系结构节点模型

节点模型分为两部分:应用层和服务层.应用层只接受服务,服务层只提供服务.由于它们都不是典型的层,所以我们分别称之为应用群和服务团队.应用群包括应用基础(网络管理和域名解析)、典型应用(WWW、E2mail 和 FTP 等)和一般应用.请注意应用群包含了所有的应用,而不只是共性的应用.应用群是各种应用的松散集合.服务团队是服务元的有组织的集合,它除了向本节点应用层提供服务外,还能和其它节点服务元合作向整个网络系统提供服务或向某一节点提供服务.

7　微通信元系统构架

因为服务元是 SDU 的发送者、接收者、转发者或变换者(和网络介质一起组成有源信道),又因为一个节点包含许多服务元,所以将它们称为微通信元.相关节点的服务团队将微通信元组织成微通信系统,再将大量微通信元系统组织成网络系统.这就是我们把服务元网络体系结构的第一个网络系统称为微通信元系统 MCES(Micro communication element system)构架的原因.

微通信元系统构架的构建原则是容易从 TCP/IP 过渡而来:(1);包格式尽可能靠近 TCP/IP(但要删除其冗余重复部分),以便简化包转换器;(2)大量吸收 TCP/IP 的成功经验,例如服务功能元素的定义、套接字机制、三次握手建立和释放连接、TCP 的状态迁徙图、滑动窗口技术,等等;(3)沿用 TCP/IP 的系统调用格式;(4)可扩展.

7.1　微通信元系统构架的参考模型

微通信元系统构架的参考模型如图 3 所示.图 3 中矩形和三角形中的 $S_{i,j}$ 表示它是第 i 类服务元中的第 j 种服务元.换言之 i,j 是服务元的标识.

服务元的功能如表 1 所示.其中,有连接类服务元和无连接类服务元为第一类服务元;ICMP、IGMP、RTCP 和分片的主动服务元为第二类服务元,其被动服务元为第三类服务元;路由类服务元为第四类服务元;加解密服务元、认证(和数字签名)服务元和 NIC 服务元是第五类服务元.

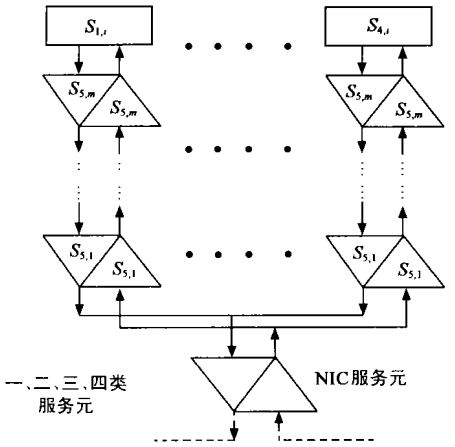

图 3 微通信元系统架构的参考模型

同一节点的第五类服务元序列可以相同,也可以不同.这样就可以实现一台主机可以和不同节点同时进行不同的通信,例如加密通信和一般通信.第五类服务元序列的结构在初始化时设置,也可以通过节点间协商进行再设置.

允许用户定义服务元用以扩展网络功能.例如用户定义新的路由递交服务元,可以实现主动网络的功能.

712 微通信元系统构架的包格式

微通信元系统构架的包设计为如图 4 所示.包由首部和数据构成,其中,基本首部由基本硬头和基本软头组成.基本硬头和 CRC 校验域由网卡处理.

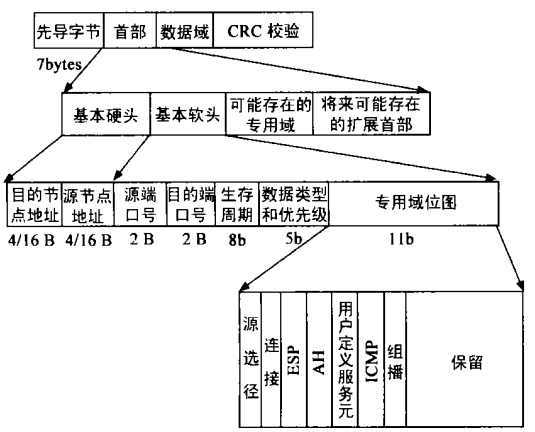

图 4 微通信元系统架构的包

基本首部是一定存在的,并由服务团队各服务元共用.各个专用域由位图表示是否存在,1 表示存在;0 表示不存在.每个专用域可以分为多个字段.例如连接专用域分为:序号、确认号、窗口尺寸和紧急指针等四个字段.每一个专用域还可以为多个服务元共用.例如连接专用域为建立连接、发送数据、接收数据和释放连接等服务元共用.

广义 ICMP、OSPF、RIP、EGP、BGP、DVMRP 和源选径目录等使用数据域(当然应该有类型标志来区别)而不使用首部;路由递交服务元仅仅转发包到相应端口的第五类服务元序列去;第五类服务元变换包的数据域、首部或整个包.例如压缩解压服务元只变换数据域,NIC 服务元变换整个包.

8 向服务元网络体系结构的过渡

基本沿用原有的网卡,仅仅稍作改变(将主机地址设为网卡地址,主机地址可以由用户来设置,主机地址采用 Ipv4 或 Ipv6 的 IP 地址)就能使用.第一步保持系统调用不变,使 TCP/IP 的浩如烟海的应用程序和建立在 TCP、UDP 上的所有软件都可以照常使用;第二步通过增添服务元,扩充系统调用以便直接支持话音和视频信号的实时传送.服务元体系结构中,由于省去了帧的处理,路由器甚至/第四层0交换机的硬件结构和包转发率和价格都将和现在的第二层交换机相近.廉价的新网络系统通过包转换器和现有因特网相连使过渡可以逐渐展开.

由于设计微通信元系统构架时,考虑了包的格式和收发规则与 TCP/IP 的对应关系.包转换器的功能就是将微通信元系统构架的包的格式和收发规则转换成和 TCP/IP 一样.

9 结论

服务元网络体系结构的 MCES 大量吸收了 TCP/IP 的成功经验和成果,其中包括 TCP 的状态迁徙图,因此可以直接使用其建模结果.由于 MCES 避免了功能重复和服务传递开销等问题,其性能参数将有所提高.MCES 的参考模型易于映射到计算机程序空间,从而容易实现.MCES 可扩展性好,可以不断改善其性能,因此适合作为下一代宽带多媒体网络的体系结构.

参考文献:

[1] D Clark, D Tennenhouse. Architectural considerations for a new generation of protocols[A]. Proceedings of Sigcomm'90[C]. New York, NY, USA: ACM Press, 1990. 200-208.

[2] D Tennenhouse, D Wetherall. Towards an active network architecture [J]. Computer Communication Review, 1996, 26(2): 79-83.

[3] A Lazar. Programming telecommunication networks[A]. Proc. of 5th International Workshop on Quality of Service 1997[C]. USA, 1997. 3-24.

[4] V Paxson, M Allman, S Dawson, et al. RFC2525, Known TCP Implementation Problems[S]. 1999203.

[5] A Mustafa, M Hassan, S Jha. Design and performance of a rate control feedback architecture for TCP/IP network[A]. Proc. of 21st IEEE International Performance Computing and Communications Conference [C]. Phoenix, Arizona, 2002.

[6] M Hassan, H Sirisena. Optimal control of queues in computer networks [A]. Proc. of IEEE International Conference on Communications[C]. Finland: ICC, 2001. 637-641.

[7] J Kay, J Pasqucde. Profiling and reducing processing overheads in TCP/IP[J]. IEEE/ACM Trans. Net, 1996, 4(6): 817-828.

[8] R Stewart, C Metz. SCTP: New transport protocol for TCP/IP[J]. IEEE Internet Computing, 2001, 5(6): 64-69.

[9] R Engel, D Kandlur, A Mehra, D Saha. Exploring the performance impact of QoS support in TCP/IP protocol stacks[A]. Proceedings of IEEE INFOCOM'98[C]. Francisco, USA: IEEE INFOCOM. 98, 1998. 3. 883-892.

[10] Stefan Boecking. 面向对象的网络协议[M]. 北京: 机械工业出版社, 2000, ISBN 7 111 08076 9/TP.

[11] Stefan Boecking, et al. A run2time system for multimedia protocols[A]. Proc. of Fourth International Conference on Computer Communications and Networks(ICCCN 1995)[C]. Las Vegas, USA: ICCCN 95, 1995. 9. 178- 185.

[12] Siemens AG. Performance and software evaluation of the modular TIP communication system[A]. Proc. of 5th International Conference on Computer Communications and Networks[C]. USA, 1996. 10.

[13] S Burkhard. Configuration of Protocols in TIP[R]. University of Cam2bridge Computer Laboratory. s Technical Report, 1995, 6. 368- 373.

[14] 金诚, 杜勇, 曾家智. 基于对象的动态协议配置[J]. 计算机应用, 2001, 6.

[15] B Braden, T Faber, M Handley. From Protocol Stack to Protocol Heap 2 Role2Based Architecture[R]. First Workshop on Hot Topics in Net2working, 2002, 10.

作者简介:

曾家智　男, 1939 年 8 月生于四川成都, 1962 年毕业于西北工业大学, 现为电子科技大学计算机学院教授、博士生导师, 西南网络与信息系统专委会副主任, 长期从事计算机网络的教学和科研. 有专著四本, 一部译著, 论文 60 余篇. 主持完成科研项目十余个. Email: jzzeng@uestc. edu. cn.

徐　洁　女, 1963 年 1 月生于四川成都, 1982 年毕业于电子科技大学计算机系, 1988 年获硕士学位, 现为电子科技大学计算机学院副教授, 目前的研究方向为计算机网络、网络体系结构和协议分析等. Email: xujie@uestc. edu. cn