TUR：一种在 UNIX 内核中实现的高性能路由器*

毕 军 吴建平

（清华大学计算机系 北京 100084）

摘要 本文在简单介绍 IP/X.25 路由器的功能和结构的基础上给出一种采用 STREAMS 机制的路由器 TUR 在 UNIX 中的实现. 本文还提出一种在面向连接子网服务上实现无连接网络互联的通用有限状态机模型和一种动态地址映射算法. 这些方法已应用在 TUR 中，以提高系统的性能. 本文最后给出性能、指标及在 TCP/IP 网络中的一个应用实例，并讨论这类通信软件普遍存在的问题及其解决思路；互操作测试.

关键词 计算机网络，网络互联，通信软件，路由器，UNIX 内核.

中图分类号 TP393

X.25 协议已被世界上许多国家广泛接受并普遍应用，成为国际公认的组建公用分组交换数据网的基础协议. 我国于 80 年代末开始组建 X.25 试验网，并于 1994 年初进行了扩容，形成了覆盖全国的中国公用分组交换数据网 CHINA PAC. 目前，Internet 已成为全球最大的计算机网络，被普遍接受为 GII 的雏形. TCP/IP 已成为事实上的工业标准，几乎所有的工作站和 UNIX 系统都采用 TCP /IP作为网络软件，而 PC 机上也出现了相应的 TCP /IP软件，这使得采用 TCP/IP 的计算机和异构网络互联成为可能.[1]我国目前比较成功的符合 Internet 标准的学术网络主要有中国教育和科研计算机网(CERNET)和邮电部的 CHINA NET 等.

考虑到我国 X.25 网作为 WAN 的骨干作用和 Internet 上丰富的资源，我们采用UNIX 提供的一种用于开发通信协议的 STREAMS 机制重新构造了 UNIX 内核，在 UNIX 核心中加入了 IP/X.25 路由器 TUR 的设计实现. 它不仅可使 PC 机作为主机加入公用数据网，而且可使 PC 机作为路由器实现 WAN 与 LAN 的互联.

本文第 1 节讨论 TUR 的功能和应用，第 2 节介绍 TUR 的设计与实现，并提出一种用于互联的状态机模型和地址映射的自适应算法，第 3 节给出该路由器的一些性能指标. 在结束语中，我们给出了一个应用实例.

---

* 本文研究得到国家自然科学基金和国家八五攻关项目基金资助. 作者毕军，1972 年生，博士研究生，主要研究领域为计算机网络协议测试，ATM 与高速网. 吴建平，1953 年生，教授，主要研究领域为计算机网络协议工程学，ATM 与高速网.

本文通讯联系人：毕军，北京 100084，清华大学计算机系

本文 1996-05-23 收到修改稿

1 TUR的功能与应用

X.25网有着范围广大的用户,同时许多单位也拥有了以UNIX为操作系统的装有TCP/IP软件包的PC机,并建立了局域网.因此,在X.25基础上实现对TCP/IP的支持,可以让范围广大的公用数据网用户使用TCP/IP协议族丰富的应用软件,并通过路由器实现X.25WAN与LAN(如以太网)的互联,这是一件很有意义的工作.[2]为此,Internet体系结构委员会先后制定了通过PSDN进行IP数据报传输和异构网互联的协议RFC877和RFC1356.[3]与此同时,国外也涌现了一些IP/X.25路由软、硬件产品(如美国的Sunlink软件和CISCO路由器).清华大学在参加国家"八五"攻关课题的研究中,在高档微机上实现了基于UNIX内核的IP/X.25路由器TUR,使X.25成为操作系统中的一个有机组成部分.目前,它已经过CERNET网络中心等单位的使用,并于1995年9月通过了"八五"攻关子专题鉴定.它不仅可以使PC机作为主机在X.25网上实现TCP/IP访问,还可以使PC机作为路由器,使本地局域网上的用户通过该路由器在X.25网上实现TCP/IP访问.用户通过该路由器,可以应用Telnet,FTP,e-mail,WWW等网络,并可以通过CERNET访问Internet.

2 IP/X.25路由器的设计

2.1 TUR在UNIX核心中的结构

STREAMS是UNIX操作系统为开发通信服务而提供的一套通用、灵活的开发工具.它通过把各层协议实现为STREAMS模块或多路器,完成网络协议体系结构的实现或UNIX设备驱动程序的开发.STREAMS机制定义了用于系统内核的数据I/O和核心与用户的接口,包括一整套系统调用、内核数据结构和内核例程.这些标准接口和机制使高性能的网络服务及其构件的开发能够实现模块化,不仅易于系统集成,软件可移植性、可复用性也很好.一个流由流首、模块和驱动程序(或多路器)组成.数据在流内以消息的形式在流首、模块和驱动程序间通信.消息是一组数据结构,用于在用户进程、模块和驱动程序之间传递数据、状态和控制信息.[4]

由于STREAMS具有上述优点,我们在TUR中采用了这种机制.在这个结构中,IP/X.25路由子层作为STREAMS多路器实现.它在核心内的上层是IP多路器,下层是X.25 PLP多路器,通过X.25LAPB模块和物理驱动程序实现底层通信支持.一方面,IP/X.25多路器(以下简称IXM)接受来自IP多路器的数据报,通过与用户空间的虚电路管理进程(以下简称VCM)的同步,获取相应的X.25逻辑信道(呼叫、维护、拆除),将IP数据报拆卸为X.25数据分组,并在该逻辑信道上通过X.25PLP多路器和底层链路模块、物理驱动发送;另一方面,IXM从X.25网接收X.25分组,并根据分组的内容向VCM或IP多路器发送原语和数据.其他模块或驱动程序间的通信类似,不再赘述.

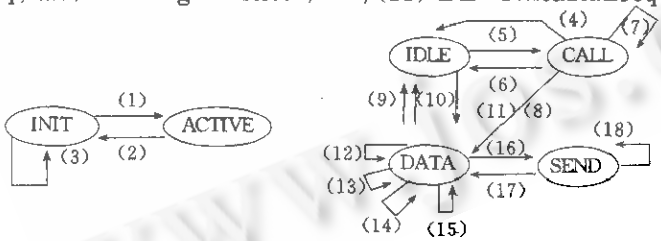
2.2 抽象服务原语的定义

多路器和模块之间通过抽象服务原语实现通信.我们利用STREAMS的M_PROTO类型作为原语的数据结构.X.25PLP层对上采用CCITT X.213定义的网络层原语作为与

上层的数据接口:N_Con (req/ind/rsp/cfm);N_Data(req/ind);N_Dis (req/ind/ind/rsp/cfm)等. IP 层对下采用LLI 逻辑链路接口:DL_Info (req/ack);DL_Bind (req/ack);DL_Unitdata (req/ ind);DL_Unbind_req;DL_Ok_ack. 我们在 UNIX 用户空间的 VCM 和UNIX核心空间的 IXM 之间定义了一套新的基于 STREAMS 的原语,实现用户空间的管理进程与内核协议之间的协调和同步:Link_req 由 IXM 发向 VCM,申请新连接;Link_ind 由 VCM 发向 IXM,指示新的连接;Link_cfm 由 VCM 发向 IXM,作为对 Link_req 申请的虚电路的确认;Unlink_req 由 IXM 发向 VCM,申请拆除一个已连接的虚电路;Unbind_req 由IXM 发向VCM,指示拆除所有已连接的流;Addr_chg 是当IXM 发现地址映射关系由内核中的多路器发向 VCM 时,申请拆除一个已连接的虚电路;State_chg 是当 VCM 不能从地址映射表中找到IP 地址时,通过向 IXM 发送这个原语,使核心状态从 CALL 转向IDLE.

2.3 状态机与虚电路管理

面向连接子网服务上实现无连接网络互联有 2 种机制:静态和动态. 静态模式是底层连接,在通信过程中一直建立;动态模式是底层连接只在上层有数据传输时才存在,当传输空闲超过某一用户规定的超时值时,底层连接被拆除. 静态模式比较容易实现但是效率不高. 我们用一个有限状态机(FSM)来抽象地描述动态机制,它不仅用于我们的软件,而且可以作为一种通用的机制适用于其它在面向连接的服务上开发无连接网络互联的应用,如 IP over ATM. 这个FSM 模型如图 1 所示. 它的各标号变迁定义如下:(1) DL_Info_req / DL_Info_ack;(2) DL_Unbind_req / Unbind_req;(3)DL_Bind_req / DL_Bind_ack;(4)I_Unlink/ ～;(5) DL_Unitdata_req / Link_req;(6)Sate_chg /～;(7) DL_Unitdate_req / N_con_req; (8) Link_cfm / ～;(9) N_Dis_req /～;(10)I_Unlink /～;(11)Link_ind /N_Con_rep;(12)N_Data_ind / DL_Unitdata_req;(13) Address changed / Addr_chg;(14) Timeout / Unlink_req;(15) N_Dis_req / Unlink_req;(16) DL_Unitdata_req / N_Data_req;(17) Sending finished / ～;(18) DL_Unitdata_req / ～.



(a) Main-FSM in UNIX kernel (b) Sub-FSM in UNIX kernel

图1 一个通用的动态互连FSM模型

IXM 接收到一个要转发的IP 数据报时将查询地址映射表. 如果此时与该地址没有连接,IXM 将向 VCM 发原语来指示 VCM 建立一条流并且发送一个 X. 25 呼叫请求分组. 如果此呼叫被接受,即可在此连接上进行数据传输. 我们使用超时机制来动态地拆除连接,这样可以大大提高系统的执行效率,减少通信费用. 接收的规程与此类似. 我们可以把这种机制准形式化地描述为如下算法:

```
do while IP datagram comes
    look up routing table
    if the connection exists
        begin
            forward datagram as packet
            reset timer
            if timeout
              disc the connection
```

```
            end-if
        end
    else
        look up address mapping table
        setup connection
        start timer
        forward datagram as packet
    end-if
end-do
```

2.4 协议数据单元的转换、路由与寻址

协议数据单元被通信的对等实体用来传输数据.在TUR中,我们把IP数据报(包括报头和报文数据)作为X.25数据分组中的数据部分并使用M_bit对数据报分段,同时要求对方按M_bit重组完整的数据报.

TUR将为它连接的每一个子网及其上的计算机建立路由表.它包含下列信息:(a)远端网络的路由器的地址;(b)每个网络的主机地址或地址域.当TUR收到一个从远端网络发来的数据报时,它从该数据报中分解出网络地址并且根据这个地址查找路由表.在判断是否能够抵达目的地后,TUR将转发这个数据报至目的主机或下一个路由器.

在网际互联地址与物理地址的映射中,经常用到3种方法.第1种方法是查询法,比如大家比较熟知的ARP.第2种方法是静态地址映射表法.在这种方法中,地址映射的关系是固定的并且不能hot-fix.也就是说,当我们在逻辑上改变一个机器的IP地址与物理地址的映射关系时,必须通知其他路由器的管理员来相应修改有关的地址配置文件以保持地址映射关系的一致性.这种方法被广泛地应用到国外各种路由器产品的IP/X.25路由模块中,例如CISCO 4700和Sunlink 8.0.我们在TUR的开发中提出并实现了一种动态的地址映射算法.采用这一算法,路由器可以学习地址映射的改变并自适应地修改地址映射表.当远方某路由器的地址映射关系改变时,它不需通知其他路由器,而本地的路由器将在通信中发现这一情况并自动地修改本地的地址表并通知系统管理员.上述过程对管理员都是透明的.为了实现对通信双方的IP地址和X.121地址的转换,我们在UNIX用户空间维护一张地址映射表并且用ioctl( )系统调用来将其送往UNIX核心空间,通过我们所定义的原语来保持两者的通信.这个地址表如表1所示,含有地址信息和与虚电路有关的状态与时钟信息,这样便于与图2所示的FSM和算法相结合.这种机制的思想虽然比较简单,但可以在一定程度上减轻管理人员的工作.

表1　地址映射表的结构

No.	IP Addr.	X.121 addr.	State	Timer	error	control
1						
2						

这个算法可以抽象地表述如下:

```
remote router:
    A system manager modifies the address table file in hard disk
    Manager sends a signal to VCM
    VCM modifies the address mapping relation in UNIX user space
    VCM indicate IXM to modifies the address table in UNIX kernel
local router:
    IXM receives a X.25 call request packet
```

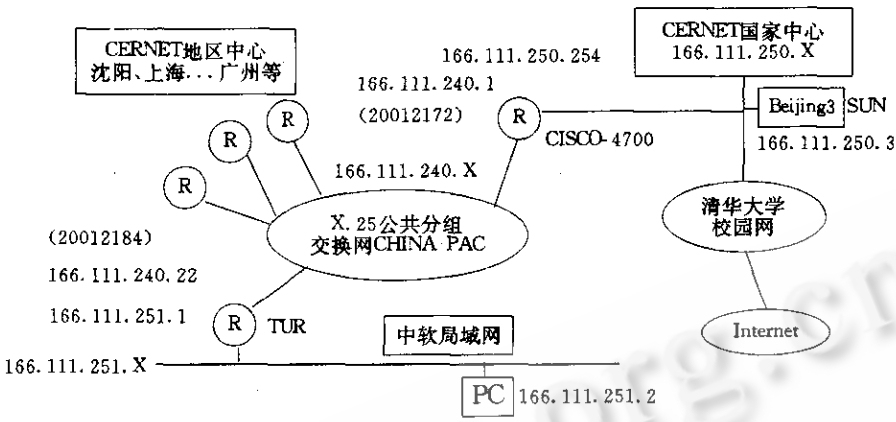


图2 高性能路由器TUR的网络应用

```
IXM record the X.121 address
IXM send a X.25 call response packet
IXM received a IP datagram
IXM look up address mapping table
if changed
    IXM modifies the address table in UNIX kernel
    IXM sends a Addr_chg primitive to VCM
    VCM modifies address mapping relation in UNIX user space
    VCM modifies the table file in hard disk and send a notice to the system manager
end-if
```

3 性能与指标

TUR 的软硬件共同作用,实现了如下主要功能和技术指标:支持协议:分组层 CCITT X.25PLP;链路层 CCITT X.25LAPB(SLP, MLP);物理层 CCITT X.21bis, RS-232 等国际标准;可扩充 X.32 拨号和 Frame Relay. 采用 RFC791、RFC877 和 RFC1356 等工业标准,支持 TCP/IP,可以实现异网互联,使用高层应:Telnet、FTP、SMTP、NFS,WWW 等. 同步串行最高通信速率可达到 256Kbps(自环测试时可达 500Kbps). 通道数:2;虚电路:64;分组尺寸:16～2048;滑动窗口尺寸:2～127. 提供开放的程序员接口:符合 CCITT X.213、CCITT X.212、CCITT X.211 服务原语. 线路方式:专线;传输方式:全双工.

表 2 给出了我们在这种高性能的路由器中设计的 X.25 适配器(X.25SCC)与国内外其它产品的性能比较. 其中前 2 种是目前国内市场上使用较多的,性能比较优越的网卡. 第 3 种是美国 TITN 公司 90 年代生产的,目前在国际上比较先进的网卡. 表 2 中的最大速率一项是在微机上通过自环测试得到的. 进行路由器的性能评价时,传输速率、内存容量和与主机通信的方式是 3 个重要的指标. 传输速率作为性能评价的重要指标是显而易见的,而内存大小往往被用户忽略. 实际上,传输速率是与所需内存大小成正比的,即线路速率越高,所需的缓冲区就越大,一旦出现缓冲容量不够时,将导致数据丢失. 所以如果没有大内存,支持高速率传输将是一句空话. 另外,分组大小、虚电路数目等参数都取决于内存容量. 从表 2 中可以看出,只有 X.25SCC 可以提供 500Kbps 的高速传输速率,这足以满足 Frame Relay 的 56～256Kbps 的要求,能适应我国公用数据网的发展. 此外,我们采用双端口存储器的共享机制实现与主机的通信,还具有程序下载(downlocad)能力,避免修改程序后需要重写

EPROM的缺点,同时也有利于用户根据自己需要开发应用程序,使其成为一种高性能的专用通信I/O前端机.

表2　性能比较表

	CPU	Mem.	SCC	DMA	S. W.	Speed
FX X. 25	none	none	6M	none	none	9. 6K
X. 25PCB	8086/8M	192K	4M	4M	none	64K
PC-COM	80188/10M	1M	5M	5M	16K	256K
X. 25SCC	80188/12M	1M	6M	6M	16K	500K

4　结束语

本文论述了高性能路由器TUR的功能、结构、设计和实现,并给出了一些算法. TUR在AST/486高档微机和AT&T UNIX SVR4.0系统上实现,已经通过邮电部的测试,并经过CERNET网络中心等单位使用,于1995年9月在中软总公司通过了由电子部主持的专题鉴定. 图2所示是它的一种网络应用环境,TUR实现了中软局域网与CERNET和Internet的互联. 以太网上的一台PC机(地址166. 111. 251. 2)上的用户通过该路由器和CHINA PAC与CERNET国内各地区网络中心互联并经CERNET进入Internet,访问了美国麻省理工学院的文件服务器. TUR的设计与实现十分符合我国的国情(PC机多、局域网多、X. 25的覆盖广阔和TCP/IP商业应用软件丰富),因此应用前景比较广阔. 我们已在其他版本的UNIX系统(如SCO, Linux)上移植. 在国家"九五"科技攻关中,我们正进行一种基于Power Pc的高性能多处理器体系结构的路由器研制,以期随着我国计算机网络的发展适应更广泛的用户,为我国信息化建设作出更大贡献.

RFC 877和RFC 1356定义的协议功能比较"薄",这既给协议实现者以较大自由空间(每个厂商的实现版本实际上都定义了独有的附加协议功能),同时又给用户带来了不同实现版本之间(路由器)之间的互通和互操作问题(如对IP数据报分段的基准长度、地址映射方法、虚电路维护规程的不统一等). 这是任何IP/X. 25路由软硬件产品固有的问题,推而广之,也是异构网互联领域存在的一个重要问题. 解决这一难题的办法就是进行互操作测试. 互操作测试旨在检测同一种协议的不同实现版本之间的互通能力和互操作能力. 目前这项研究已经引起了国际网络协议工程学领域的重视. 我们正在尝试使用一种形式化的数学模型对互操作测试的概念、测试方法和测试过程进行定义与分析,并用于TCP/IP协议的测试实验,这种方法可以明确地定义互操作的内涵是什么、什么样的测试属于互操作测试、哪种测试结果能表明2个系统之间具有互操作能力. 这项研究具有很深远的理论和现实意义,希望这个思想能为进行相关研究的学者提供参考.

参考文献

1　Comer D G *et al*. Internetworking with TCP/IP. 3th ed., Prentice Hall Inc, 1994.

2　Schepers H *et al*. LAN/WAN interworking in the OSI envirment. Computer Network & ISDN System, 1992, **23** (**4**).

3　RFC 1356. Multiprotocol interconnect to X. 25 and ISDN in the packet mode. 1992.

4　Bi Jun, Shi Meilin, Wu Jianping. UNIX kernel_based Reference Implementation in PCTS. In: Cao Z G ed., 5th

International Conference on Communication Technology, Beijing: CIE/CIC/IEEE, 1996. 248～250.

TUR: A HIGH PERFORMANCE ROUTER IN UNIX KERNEL

BI Jun WU Jianping

(*Department of Computer Science Tsinghua University Beijing* 100084)

Abstract After a brief introduction of the function of IP/X. 25 router, this paper presents the structure of a router TUR, which is designed and implemented in UNIX based on STREAMS mechanism. A general finite state machine for the connectless internetworking over connection oriented subnetwork service and a dynamic address mapping algorithm will be described. Finally, this paper introduces its performance and gives the application an example. The interoprability testing will be discussed too.

Key words Computer network, internetworking, communication software, router, UNIX kernel.

Class number TP393