应用于射电天文数字终端的网络传输系统的设计

于 威，张秀忠，吴亚军，郭绍光，甘江英

（中国科学院 上海天文台，上海 200030）

摘 要：射电天文数字终端对各类天体发出的无线电信号进行处理后，需把数据传送到数据记录设备存储起来，以进行更进一步的分析和处理。为了满足数据传输需要，本文设计了一种千兆网网络传输系统把射电天文数字终端产生的数据传送到数据记录设备，测试结果表明，该网络系统在数据速率低于 640 Mb/s 的情况下工作良好，数据丢失率非常低。目前该系统已初步应用于 DBBC、连续谱终端等射电天数字终端中。

关 键 词：射电天文；数字化终端；千兆网

中图分类号：TN919.6^{+}4；TP319

1 引 言

现场可编程门阵列 (field programmable gate array，FPGA) 是当今数字电路设计的主要硬件平台，其主要特点是用户通过在软件上编程来设计硬件电路。修改和升级 FPGA 数字电路时，只需要在计算机上修改和更新程序，不需要改变印刷电路板 (printed circuit board，PCB)，这大大缩短了硬件电路的设计周期，提高了灵活性并降低了成本[1]。由于 FPGA 强大的性能以及设计数字电路的方便性，目前射电天文数字终端的主要部分一般都是基于 FPGA 开发的。射电天文数字终端对各类天体发出的无线电信号进行处理后，需把数据传送到数据记录设备存储起来，以进行更进一步的分析和处理。常见的数据传输接口有外围部件互连总线 (peripheral component interconnect，PCI) 接口、串口、通用串行总线 (universal serial bus，USB) 接口、网络接口等多种，其中网络接口具有传送距离远、传输速率高、操作灵活等优点，非常适合 FPGA 设备与数据记录设备 [如高性能个人计算机 (personal computer，PC)、工作站、服务器等] 之间进行数据传输。中国科学院上海天文台甚长基线干涉测量 (very long baseline interferometry，VLBI) 技术实验室开发的数字基带转换器 (digital base-band converter，DBBC) 电路板上目前为千兆网网口，所以基于此电路板开发的中国 VLBI 数据采集系统 (Chinese VLBI data acquisition system，CDAS)、连续谱终端、硬件处理机等设备均采用千兆网传输系统。

传输控制协议/网际协议 (transmission control protocol/internet protocol，TCP/IP) 栈

收稿日期：2012-06-18； **修回日期**：2012-08-23

资助项目：国家自然科学基金资助项目 (11173051)；国家自然科学基金资助项目 (10978024)

包括应用层、传输层、网络层、数据链路层。在本设计中，应用层即为射电天文数字终端，其产生的数据为原始数据，没有任何协议封转；网络层采用 IP 协议；数据链路层采用以太网协议。传输层可以采用的协议有 TCP 协议和用户数据报协议 (user datagram protocol，UDP)，TCP 是一种面向连接的协议，工作过程比较复杂，数据传输可靠性高，但实时性和传输效率不高。由于其“滑动窗口”和“重传确认”等复杂的工作过程，一般要求数据发送端有很大的缓存，比较适合对实时性要求不高但对数据可靠性要求较高的设计中使用。UDP 面向无连接、无应答机制，工作过程比较简单，实时性和传输效率较高，但不保证数据的可靠性，不需要发送端有很大的缓存。因为本设计中数据发送端为 FPGA 设备，不可能有很大的缓存，而且射电天文数字终端不是根据接收端的数据接收情况间歇性地产生数据，而是不断恒定地产生数据，所以对数据传输的实时性要求较高，因此本设计采用 UDP 协议。为了保证数据传输尽可能可靠，本文设计了一种有应答机制的 UDP 协议。应答方式为：发送端发送完一组数据包后 (在实际设计中一组数据为 20 个 UDP 数据包，数据包包长为 1518 个字节)，等待接收端回送应答帧 (acknowledgement，ACK)，收到 ACK 后继续发送下一组数据。

本文设计的网络传输系统分为两部分：数据发送端和数据接收端。数据发送端用 FPGA 硬件实现，作为一个独立的知识产权核 (intellectual property core，IP core) 嵌入到射电天文数字终端中。数据接收端是在 Linux 操作系统下基于原始套接字编写的软件。

2 网络传输系统的具体实现

2.1 数据发送端

数据发送端除物理层芯片外全部集成在一片 FPGA 内，采用全硬件的方式实现。在本设计中 FPGA 采用 XILINX 公司的 virtex4 FX60 芯片。

发送端硬件框图如图 1 所示，图中 TX 表示发送信号，RX 表示接受信号。原始数据是射电天文数字终端产生的数据，也是需要传送的数据。先入先出队列 (first input first output，FIFO) 起缓冲的作用，以防止数据溢出，在本设计中 FIFO 的容量为 65 536×32，即深度为 65 536，数据宽度为 32 位。如果射电天文终端产生的数据不是 32 位，则需要先通过并串或者串并转换，把数据转换为 32 位再进入 FIFO。

协议封装模块是本文设计的重点，其主要功能包括：计算 UDP 和 IP 校验和、把原始数据封装成标准的网络数据包、添加序列号、按照本地连接 (LOCALLINK) 格式发送以太网数据包、接收 ACK 信号、设置超时定时器等。其具体工作流程将在后面详细介绍。

LOCALLINK 模块由两个具有 LOCALLINK 接口的 FIFO，即本地连接先入先出队列 (locallink first input first output，LLFIFO) 构成，其中一个 FIFO 接收协议封装模块发来的网络数据包，并传送给三态以太网媒体访问控制 (tri-mode ethernet medium access control，TEMAC) 模块，另一个 FIFO 接收 TEMAC 发送过来的 ACK 信号，并传送给协议封装模块。LOCALLINK 接口标准传输数据的时序图如图 2 所示。图中只画出了传输 8 个数据的样图。图中 clock 为时钟信号，data[7:0] 为要传输的数据，sof_n 为每帧数据的起始标志，eof_n 为每帧数据的结束标志，src_rdy_n 为源端是否准备就绪的标志，dst_rdy_n 为目的

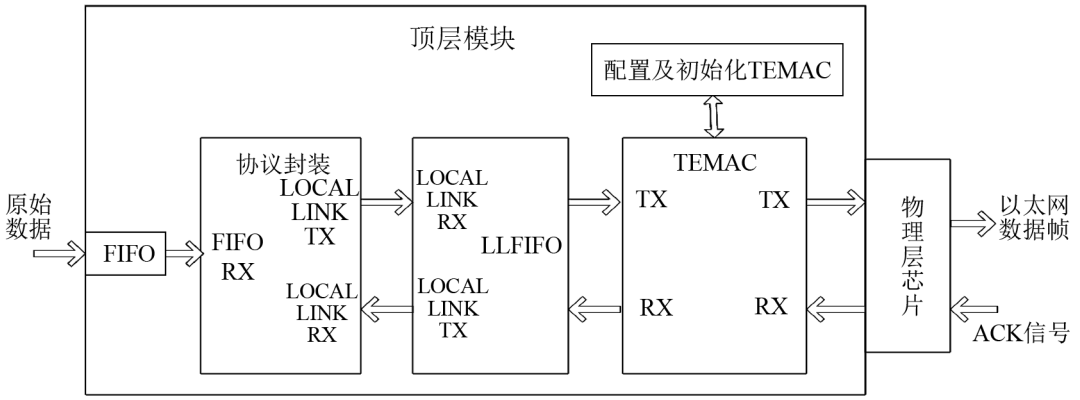

图 1　发送端硬件框图

端是否准备就绪的标志。在传输第一个数据时，sof_n 信号电平拉低一个时钟周期，此后的传输过程中 sof_n 信号恢复高电平，在传输最后一个数据时，enf_n 信号电平拉低一个时钟周期，在整个传输过程中 dst_rdy_n 和 src_rdy_n 要都处于低电平[2]。

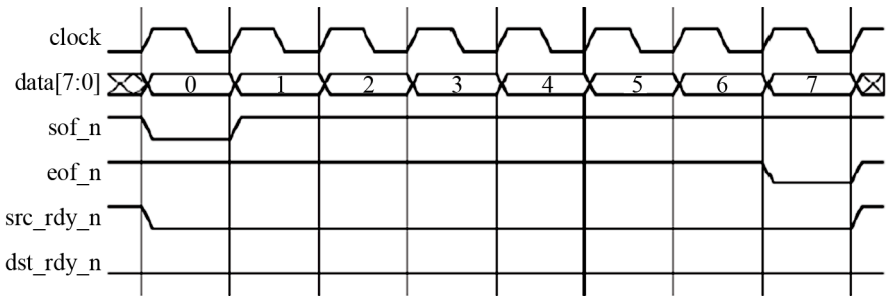

图 2　LOCALLINK 接口时序图

TEMAC 模块调用 XILINX 公司的免费 IP core，其功能较多，在本设计中只用到部分功能，主要如下：计算以太网的校验和，并把校验和 (4 字节) 添加到数据帧的最后，然后把以太网数据帧转化为物理层芯片能够识别的数据格式，并通过千兆媒体独立接口 (gigabit medium independent interface，GMII) 发送到物理层芯片。如果数据帧少于 64 字节，则需加入填充数据。通过管理数据输入输出 (management data input/output，MDIO) 接口检测和配置物理层芯片。设置地址过滤功能，阻止媒体访问控制 (medium access control，MAC) 地址非法的数据帧进入上层。TEMAC 在工作之前要对其进行配置。配置 TEMAC 是通过一个状态机实现的。

物理层主要包括物理编码子层 (physical coding sublayer，PCS)，物理媒介适配层 (physical media adaptation layer，PMA) 两部分，其功能主要包括：安装或卸载前导码、完成 4B-5B 的编解码、串并/并串转换、非归零码 (no return to zero，NRZ) 和非归零反相编码 (no return to zero inverse，NRZI) 之间的转换等[3]。本设计中物理层芯片采用的是 marvell 公司的 88E1111 芯片，此芯片有 1000BASE-T、100BASE-TX 和 10BASE-T 三种工作模式[4]

(即 1 Gb，100 Mb，10 Mb 三种速率)，本设计采用 1000BAE-T 工作模式。

协议封装模块的工作流程框图如图 3 所示。流程步骤如下：

1) 从 FIFO 中取出原始数据，并存入随机存取存储器 (random access memory，RAM) 中 (此 RAM 深度为 367，位宽为 32，用来存放一帧原始数据)。在取数据的同时计算 IP 以及 UDP 校验和 (IP 校验和只计算 IP 报头的数据，UDP 校验和要计算全部 UDP 数据包)。

2) 当 RAM 中数据满时，为此帧数据添加各层协议的包头 (从外到内依次为以太网帧头、IP 报头、UDP 包头) 和序列号，构造成一个标准的网络数据包。序列号为 32 位，紧跟在 UDP 包头后面，序列号的高 24 位表示此帧的组号，低 8 位表示此帧的组内帧号，每组有 20 帧数据。帧号从 0 递增到 19；组号从 0 递增到 4 294 967 295，然后再从 0 开始递增。

3) 用发送状态机发送已经构造好的数据包。这里的发送状态机是指按照 LOCALLINK 接口标准把一帧数据发送到 LOCALLINK FIFO 中。

4) 发送完一帧数据后，判断此组数据是否发送完成。判断方式为：如果此帧数据组内帧号为 19，说明此组数据发送完成，否则说明此组数据还没有发完。如果发完一组数据则进入步骤 5，否则进入步骤 1。

5) 等待发送端回送 ACK，并设置定时器。如果在定时器未超时时收到 ACK，则进入步骤 1，并把定时器清零，继续接收下一组数据。如果发生超时，则进入步骤 3，重发此组数据。

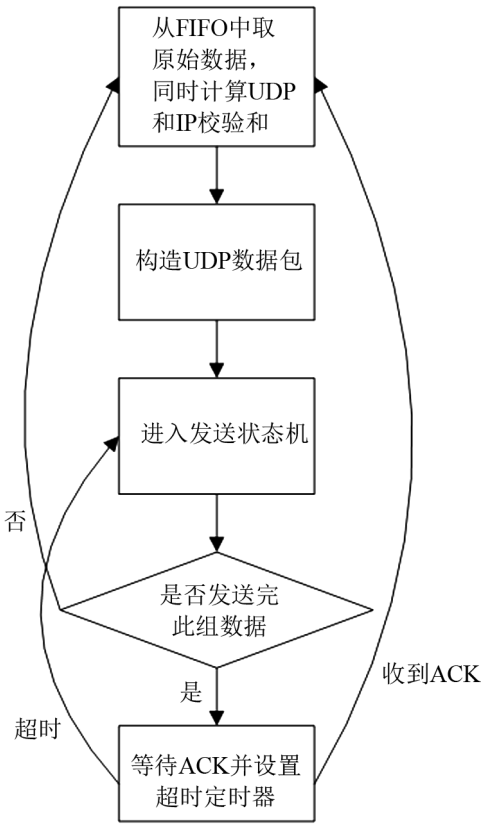

图 3 协议封装模块工作流程图

本设计中定时器超时时间设定为 2 ms。超时定时器的大小要根据发送端发送完一组数据到接收到 ACK 信号这段时间的时长 (在此表示为 t_1) 来设定。为了测量 t_1 我们在 FPGA 设计了一个计数器，当发送端发送完一组数据的最后一帧后，计数器开始计时 (计数器工作的时钟频率为 125 MHz)，当收到 ACK 信号时，先用一个寄存器记下计数器的值，再把计数器清零，然后通过 ChipScope 软件来抓取寄存器的值。经过多次测量发现计数器清零前最大值大约为 7000~9000，但偶尔也有出现最大值到 70 000 的情况。计数器最大值为 7000~9000 对应的 t_1 的大小范围为 56 μs~72 μs，计数器最大值为 70 000 对应的 t_1 的大小为 560 μs。在确定超时定时器超时时间时要综合考虑两方面因素：① 定时器超时时间不能过短，这是为了防止接收端因意外情况而推迟发送 ACK(接收端为一通用计算机，计算机的 CPU 可能会忙于处理其他进程而推迟发送 ACK)，导致 t_1 大大超过 56 μs~72 μs，如上述 t_1 为 560 μs 的情况；② 定时器超时时间不能过长，如果某次接收端没有捕获到某组数据的最后一帧，则会一直等待此帧数据而不发送 ACK，这时如果定时器超时时间过长，会使射电天文终端产生的数据溢出发送端的 FIFO 而导致数据在发送端直接丢失。所以综合考虑上述两种情况，在设计时把超时定时器设定为 1 ms、2 ms、10 ms 三种情况进行测试。经过测试发现在 2 ms 情况下数据丢失率最小。

2.2　数据接收端

接收端接收 UDP 数据有多种方式：① 采用 Linux 操作系统自带的基于 TCP/IP 协议栈的套接字 (socket)；② 采用 Linux 操作系统自带的原始套接字；③ 采用第三方网络数据捕获函数库如 BPF、libpcap、PF_ring 等。对于第一种方式，网络数据首先进入网卡缓存，网卡剥去数据包最外面的以太网帧头。然后数据包进入 Linux 操作系统内核缓存，由标准的 TCP/IP 协议栈程序去除 IP 报头和 UDP 包头 (在这个过程中数据包还会发生几次内存拷贝)，最后原始数据才进入用户内存区域。这种方式下网络数据经过多次内存拷贝，数据捕获效率严重降低。第二种方式的工作过程为：网络数据进入网卡缓存后，网卡不对数据进行任何处理 (网卡工作在混杂模式)，用户空间从网卡缓存中直接取出数据，然后把以太网帧头、IP 报头、UDP 包头一起去除。这种方式减少了内存拷贝的次数，大大提高了数据捕获效率。第三种方式中的网络数据捕获函数库其实是调用了第二种方式中的原始套接字，然后基于原始套接字开发了各种上层函数，方便用户使用。其工作过程与第二种方式一样，但多了一次内存拷贝，因此降低了数据捕获效率。因此第二种方式数据捕获效率最高，本设计采用第二种方式。需要说明的是，后两种方法中网卡工作在混杂模式，程序需要指定网卡的端口名称，因此对于不同的计算机要根据计算机中网卡的端口名称对程序做相应的修改。

接收端工作流程图如图 4 所示。流程步骤如下所示：

1) 建立及初始化原始套接字。

2) 建立两个缓冲区 buff A 和 buff B，用乒乓操作的方式交替缓存数据包。缓冲区的大小都是 1468×20 080 字节 (每个数据包前 42 个字节为以太网、IP、UDP 的包头，接下来四个字节为序列号，随后的 1486 个字节为要存储的有用数据。实际上一个缓冲区只存放 20 000 个数据帧，缓冲区开得稍大是为了防止出现意外而溢出)。

3) 生成两个线程 (th1 和 th2)，th1 主要用来捕获数据包，th2 主要用来把缓冲区的数据

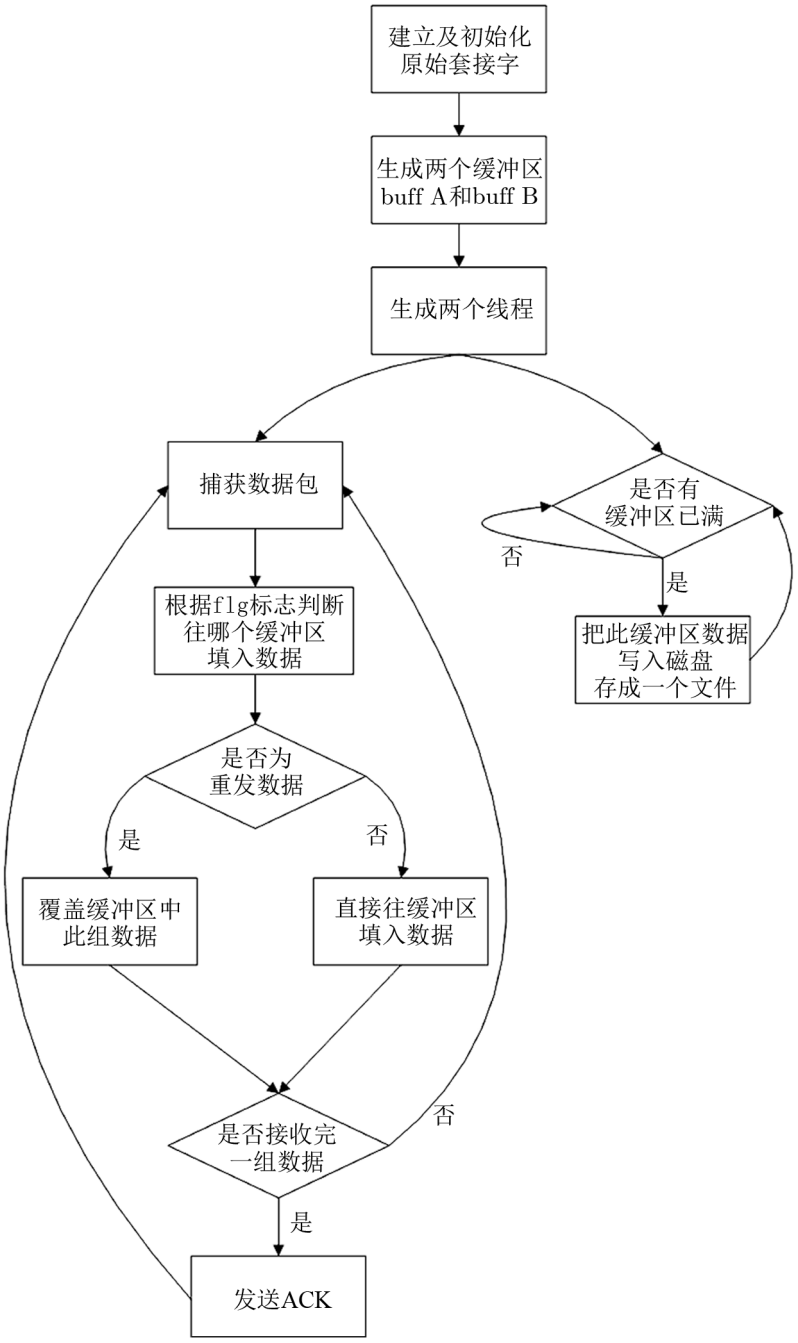
建立及初始化
原始套接字
生成两个缓冲区
buff A和buff B
生成两个线程
捕获数据包
是否有
缓冲区已满
否
是
把此缓冲区数据
写入磁盘
存成一个文件
根据flg标志判断
往哪个缓冲区
填入数据
是否为
重发数据
是
否
覆盖缓冲区中
此组数据
直接往缓冲区
填入数据
是否接收完
一组数据
否
是
发送ACK

图 4 接收端工作流程图

写入磁盘存成文件。两线程并行工作。

线程 1 工作流程如下：

① 调用原始套接字网络接收函数捕获数据包。

② 捕获到一帧数据后根据 flg 标志判断把数据填入哪个缓冲区。判断方式为：当 flg 为 0 时，往缓冲区 buff A 中填入数据；当 flg 为 1 时，往缓冲区 buff B 中填入数据。flg 初始值0。

③ 根据每一帧数据中序列号的高 24 位 (即组号) 判断是否为重发数据，判断方式为：如果组号与上一组数据的组号相同则为重发数据，否则为新数据。

④ 如果为重发数据，则覆盖缓冲区中此组数据，否则直接把数据存入缓冲区中。当缓冲区数据帧个数达到 20 000 时 flg 的值改变，同时此缓冲区的满标志 (buff A 满标志为 full A，buff B 满标志为 full B) 被置为 1。

⑤ 判断每帧数据低 8 位的值是否为 19，如果不是，则说明发送端没有发完一组数据，还在继续发送数据，返回步骤 1；如果是，则说明发送端已经发完一组数据，进入步骤 6。

⑥ 调用原始套接字网络发送函数发送 ACK 信号，ACK 信号为全零的以太网数据帧。发送完 ACK 信号后返回步骤 1。

线程 2 工作流程如下：

① 根据 buff A 和 buff B 的值，判断是否有缓冲区已满，如果有则进入步骤 2，否则继续判断。

② 把该缓冲区的数据写入磁盘，存成一个单独的文件，并把满标志置为 0，返回步骤 1。

3　结果测试及分析

用两台安装有 Linux 操作系统的高性能 PC 机接收数据进行测试。两台 PC 机配置如表 1 所示。表 2 为测试结果，其中数据丢失率是指所丢失的数据占发送端发送数据总量的比例。在测试时，发送端发送 32 位不断累加的数据 (数据从 0 累加到 0xFFFFFFFF，再返回到 0)，每次测试接收端接收 50 个数据文件，每个文件大小为 300 MB。对接收到的 50 个文件的数据，判断每两个相邻的 32 位数据之间的差是否为 1，如果不为 1，说明这两个数据间有数据丢失，记录下丢失的数据的总数目。每个速度下测试 5 次，然后对 5 次的结果求平均。

表 1　高性能 PC 机配置

型号	戴尔 precision T1500	戴尔 OPTIPLEX 755
CPU	Intel 酷睿 2 四核 i7-860	Intel 酷睿 2 四核 Q6600
CPU 主频	2.8 G	2.4 G
内存大小	4 GB	4 GB
硬盘	7200 转，SATA 接口，500 GB	7200 转，SATA 接口，250 GB
网卡	集成 Broadcom(R)5 777 780 千兆以太网卡	Intel(R) 82 566DM-2 千兆以太网卡
操作系统	Ubuntu	Redhat

数据丢失有两方面原因，一是抓包程序没有抓到某个数据包，那么此帧数据便丢失。此

表 2 定时器超时时间为 2 ms 时丢包率测试结果

发送端数据发送速率/Mb·s^{-1}	数据丢失率 (PC 机 1)	数据丢失率 (PC 机 2)
64	0	0
128	0	0
256	0.000 001%	0.000 001%
512	0.000 01%	0.000 01%
640	0.000 01%	0.000 01%
768	0.01%	0.01%
1000	非常高	非常高

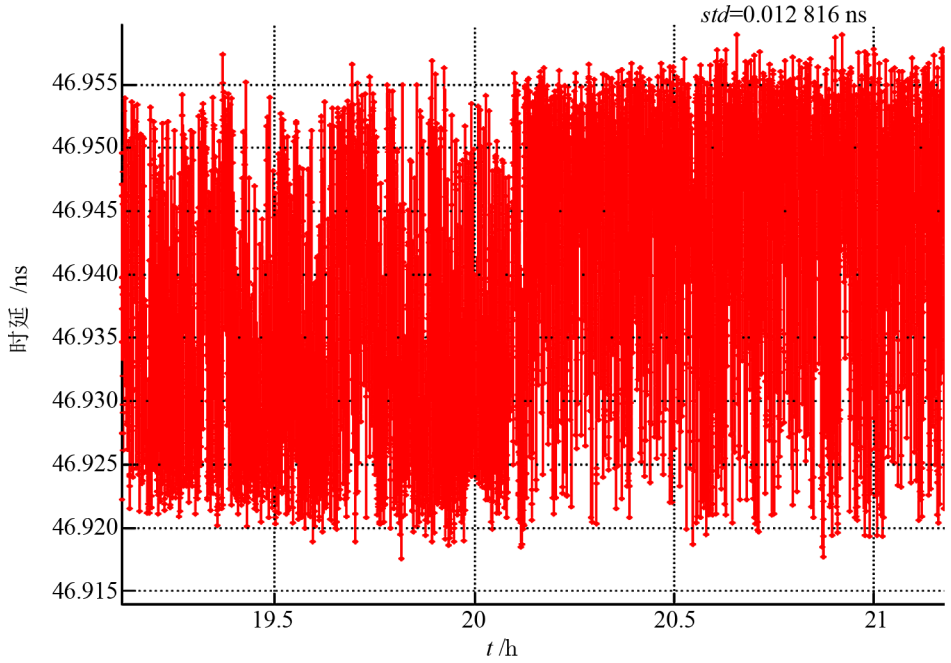

图 5 两面天线的时延

后由于重传和超时可能导致发送端 FIFO 溢出，新生成的数据便丢失。二是由于写盘不及时，虽然写盘的平均速率很高，但也会出现偶尔峰值速率很低的情况，即会出现某个乒乓 RAM 已满，而另外一个乒乓 RAM 中的数据还没有完全写到磁盘中的情况。这时接收端收到的数据就会由于无法放入 RAM 中而丢失。当发送速率为 1 Gb/s 时，数据大量丢失很可能就是由于写盘不及时造成的。由表 2 可以看出，此网络系统在发送端数据发送速率不超过 640 Mb/s的情况下，数据丢失率非常低。

目前该网络系统已经应用于上海天文台 VLBI 技术实验室开发的 DBBC、连续谱终端等设备中，并进行了一系列的实验。图 5 为使用 DBBC 于 2012 年 3 月 19 日在西安临潼用两面相距 10 m、口径均为 3.7 m 的小天线接收同一颗 GEO 卫星的信标，测得的信号到达两根天

线的时延差 (未扣除系统误差)，结果表明，其随机误差小于 13 ps(积分时间为 131 ms，总观测时间为 15 h，数据速率为 128 Mb/s)。图 6 和图 7 为使用连续谱终端于 2012 年 6 月 27 日用上海佘山 25 m 天线观测 3c123, 3c147, 3c286, 3c295 四颗射电源所得到的右极化和左极化的功率图，图中观测频段为 C 波段 (6 cm)，观测过程为 ON, ON-Fire, OFF, OFF-Fire。OFF 在赤经方向 ±1°，总观测时间为 10 h。不过此次实验数据速率较低，只有 32 Mb/s。一系列的实验结果表明，该网络系统在 DBBC 和连续谱终端中能长时间正常工作，但是这些实验数据速率都较低，对于该网络系统的检验还不够充分。后续的工作主要包括：① 使用能产生高速数据的射电天文终端继续检验此网络系统；② 使用带有 raid 卡的服务器作为数据记录设备来提高数据记录速率；③ 基于此千兆网系统开发万兆网网络系统。

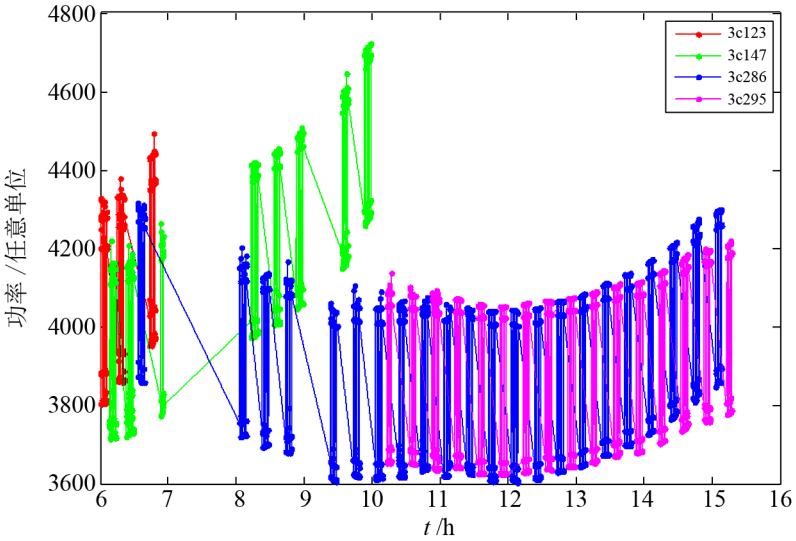

图 6　四颗射电源的右极化功率图

参考文献：

[1] 田耘，徐文波. Xilinx FPGA 开发实用教程. 清华大学出版社, 2008: 1

[2] Logicore IP Tri-mode Ethernet MAC V4.1 Getting Started Guide, datasheet, 2009. http://www.xilinx.com

[3] 孙兵，吴礼章，张丽，等. 通信与信息技术, 2008, 171: 76

[4] 88E1111 Product Brief, datasheet, 2009. http://www.marvell.com/transceivers/assets/Marvell-Alaska-Ultra-88E1111-GbE.pdf

The Design of Network Transmission System Applied in

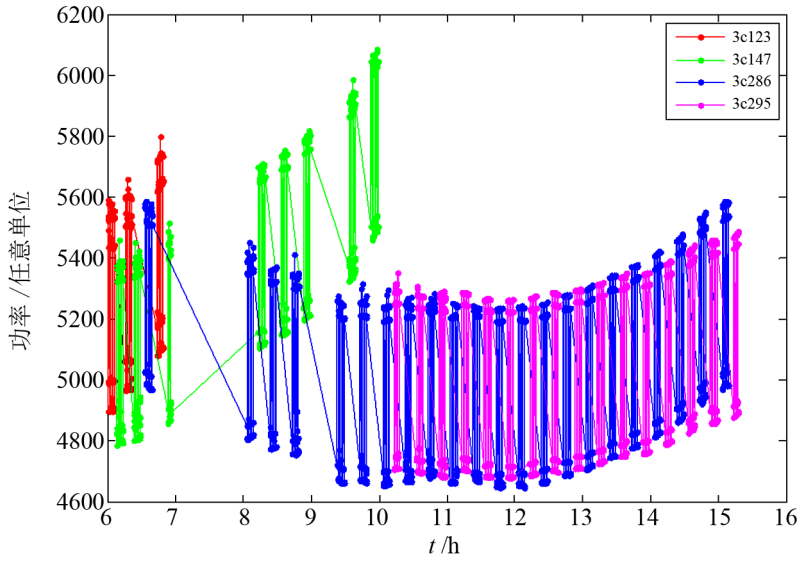

图 7 四颗射电源的左极化功率图

Radio Astronomy Digital Equipment

YU Wei, ZHANG Xiu-zhong, WU Ya-jun, GUO Shao-guang, GAN Jiang-ying

(Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030)

Abstract: When radio signal was processed by radio astronomy digital equipment, we need to transmit it to data record equipment, so that we can undertake further analysis and processing. In order to satisfy the requirement of data transmission, in this paper we designed a network transmission system based on Gigabit Ethernet. The test result showed that this system worked well when the data rate was lower than 640 Mb/s. Now this system has been applied in some radio astronomy equipments such as DBBC, continuous spectrum instrument and so on.

Key words: radio astronomy; digital equipment; Gigabit Ethernet