第33卷 第14期 计 算 机 工 程 2007年7月
Vol.33 No.14 Computer Engineering July 2007
·工程应用技术与实现· 文章编号：1000—3428(2007)14—0207—03 文献标识码：A 中图分类号：TP393.05

机群系统DDR DIMM总线光互联网络适配器设计

刘兰军，张晓彤，王景存
(北京科技大学信息工程学院，北京 100083)

摘 要：介绍了一种面向机群系统双环形网络拓扑结构的高速光互联网络适配器的设计和实现方法。该网络适配器基于FPGA技术实现，总线接口采用高速、高带宽的DDR DIMM总线，网络传输介质采用光纤，底层路由协议采用FPGA内部硬件逻辑实现，全方位保证了高带宽、低延迟、高可靠的网络特性。

关键词：机群；DDR DIMM；FPGA；光互联网络适配器；硬件路由

Design of Optical Interconnection Network Adapter Based on DDR DIMM Bus for Cluster System

LIU Lanjun, ZHANG Xiaotong, WANG Jingcun
(Information Engineering School, University of Science and Technology Beijing, Beijing 100083)

【Abstract】This paper introduces the implement methods of a high-speed optical interconnection network adapter for double token-ring network topology cluster. This network adapter is implemented based on FPGA technology and adopts high-speed high-bandwidth DDR DIMM bus as interface, optical fiber as network transmission media. Its routing protocols are implemented based on hardware logic embedded in FPGA chip. The network characteristics such as high-bandwidth, low-latency, high-reliability are ensured.

【Key words】cluster; DDR DIMM; FPGA; optical interconnection network adapter; hardware routing

机群系统是基于网络互联由多个计算机节点构成的并行计算机系统，其性能主要受两方面制约：计算机节点性能和互联网络性能[1]。目前，随着桌面PC机性能的提高以及大规模采用商业工作站作为计算机节点，互联网络的性能逐渐成为提高机群系统性能的“瓶颈”。因此，开发高带宽、低延迟、高可靠的互联网络是目前机群系统研究的一个热门课题。

对于如何提高机群系统的互联网络性能，有文献从网络传输介质的角度，设计开发了采用光纤作为网络传输介质的PCI总线光互联网络适配器[1]，这在一定程度上提高了网络性能，但其采用低带宽、低速率的PCI总线作为网络适配器与计算机节点的接口，总线接口成为通信的“瓶颈”[2]。

为此，文献[2~5]进一步提出采用内存总线接口作为网络适配器的接口。内存总线接口是一个高速高带宽接口，因此基于内存总线的光互联网络适配器是提高机群系统互联网络性能的有效方案。本文介绍了面向机群系统的一种基于FPGA技术实现的DDR DIMM总线光互联网络适配器的设计方法。

网络适配器总线接口采用目前流行的DDR200总线接口，网络传输介质采用光纤，控制逻辑和底层路由协议基于FPGA技术由硬件实现，从各个方面保证了网络适配器的高带宽、低延迟、高可靠等特性。

1 DDR DIMM总线光互联网络适配器的设计方案

本文设计的DDR DIMM总线光互联网络适配器基于Xilinx公司Virtex-Ⅱ Pro系列高档FPGA实现，整体结构框图如图1所示。虚线框内为FPGA内部的硬件逻辑。

由图1可以看出，DDR DIMM总线光互联网络适配器的功能逻辑全部集成到一片FPGA中，主要包括3部分：(1)DDR DIMM总线接口逻辑及数据缓冲区部分，包括DDR DIMM总线接口逻辑、模式控制、读写控制和数据接收/发送缓冲区(FIFO)；(2)数据接收/发送控制及路由选择部分，包括数据接收/发送控制逻辑和路由选择控制逻辑；(3)高速数据并/串(串/并)转换及串行发送/接收部分，通过配置FPGA内嵌的ROCKETIO模块来实现。FPGA的高速串行信号通过光/电、电/光转换电路与网络传输介质光纤相连。

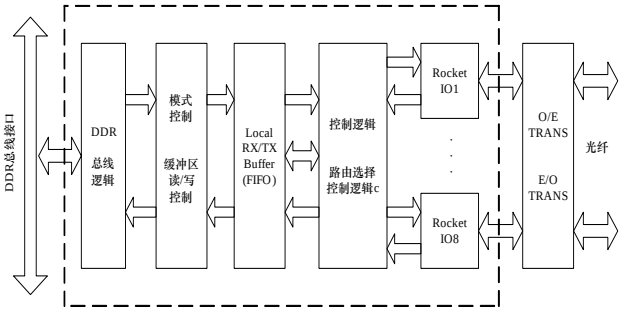

图1 DDR DIMM总线光互联网络适配器整体框图

2 网络结构及工作原理

本文设计的DDR DIMM总线光互联网络适配器具有硬件路由功能，其路由协议是针对双环形网络设计的，基于该网络适配器搭建的机群系统的网络结构如图2所示。

该机群系统是一个双环系统，内外环可分别独立的工作，各个节点计算机之间可以进行全双工的数据通信。

基金项目：中科院计算所知识创新工程资助项目“HPC-OG模拟系统及相关技术研究”(20036040)

作者简介：刘兰军(1979－)，男，博士研究生，主研方向：无线传感器网络，高性能网络通信及数字逻辑设计；张晓彤，博士、副教授；王景存，博士研究生

收稿日期：2006-07-21 **E-mail：**liulanjun123@sohu.com

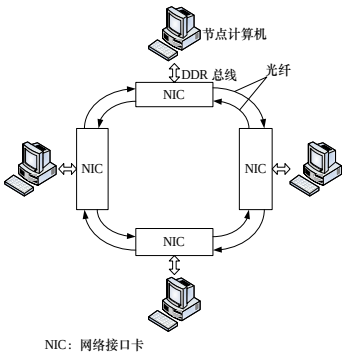

图 2　双环形网络结构机群系统

本设计定义的数据包结构如图 3 所示。

8bit	8bit	16bit	variable	32bit
数据源地址	数据目的地址	数据长度标识	可变长度数据	数据校验（CRC）

图 3　环形网络结构机群系统数据包结构

数据包由数据源地址、数据目的地址、数据长度标识、可变长度数据及循环冗余校验(CRC)构成。地址编码设计为 8 位，可以支持 256 结点规模的机群系统；数据长度标识定义为 16 位，可以支持 64KB的数据包；CRC校验定义为 32 位，可以保证数据的可靠传输。环形网络的路由算法[1]比较简单，节点计算机接收到数据包后，首先判断目的地址是否跟本节点地址相同，若相同且源地址与本节点地址不同，则接收数据包到本节点数据缓冲区并处理；若相同且源地址与本节点地址相同，则表明数据包是检测链路连通的回归数据包；若不同且源地址与本节点地址不同，则表明是另一节点的数据包，从另一输出端口转发；若不同但源地址与本节点地址相同，则表明数据包是错误数据包或目的节点不存在的数据包，予以舍弃处理。数据包中没有定义包的起始标志和结束标志，节点计算机可以根据数据长度标识来自动识别数据的开始和结束[1]。

3　DDR DIMM 总线光互联网络适配器的设计

下面将按模块分别阐述 DDR DIMM 总线光互联网络适配器的 FPGA 内部具体逻辑实现。

3.1　DDR DIMM 总线接口逻辑及数据缓冲区设计

DDR DIMM 总线接口逻辑及数据缓冲区部分的功能逻辑框图如图 4 所示。

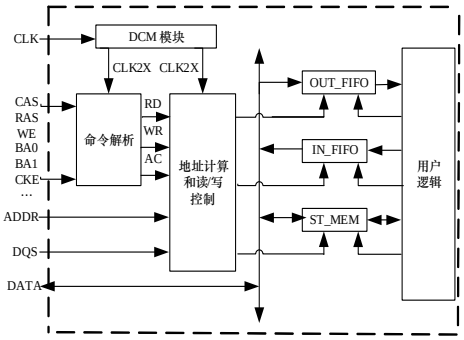

图 4　DDR DIMM 总线接口逻辑及数据缓冲区功能逻辑框图

JESD79D 标准规定的 DDR DIMM 总线的操作命令集合是针对 DDR SDRAM 的特性而设置的。在本设计中，考虑到接口逻辑只需完成数据的读写，因此，对总线命令进行了解析过滤，只考虑了激活命令 Active、读命令 Read 和写命令 Write。

DCM 模块由 FPGA 内部的 Digital Clock Management 实现，为整个接口逻辑提供工作时钟。DDR DIMM 内存槽上的时钟源 CLK 经 DCM 模块倍频后产生两倍频于 CLK 的内部工作时钟 CLK2X，驱动接口逻辑的其他模块。

命令解析模块，功能是对 DDR DIMM 内存接口的各种命令进行解析，对读、写和激活命令以外的命令予以过滤。实现的方法是在每个 CLK2X 的上升沿对 Cas、Ras、Cs、We 等信号进行采样，根据其组合识别出主机发出的读、写和激活命令。

地址计算和读写控制模块，为了方便用户逻辑的设计，DDR DIMM 总线接口部分采用输入 FIFO、输出 FIFO 及双端口 RAM(ST_MEM)3 种接口与用户逻辑相连。地址计算和读写控制模块主要完成以下任务：在激活命令到来时，锁存 ADDR 上的行地址；在读或写命令到来时，锁存 ADDR 上的列地址；然后进行地址译码，发出读写控制信号，完成数据的读写操作。

3.2　数据接收/发送控制及路由选择逻辑设计

数据接收/发送控制及路由选择逻辑设计是网络适配器设计的核心部分，其功能逻辑框图如图 5 所示。

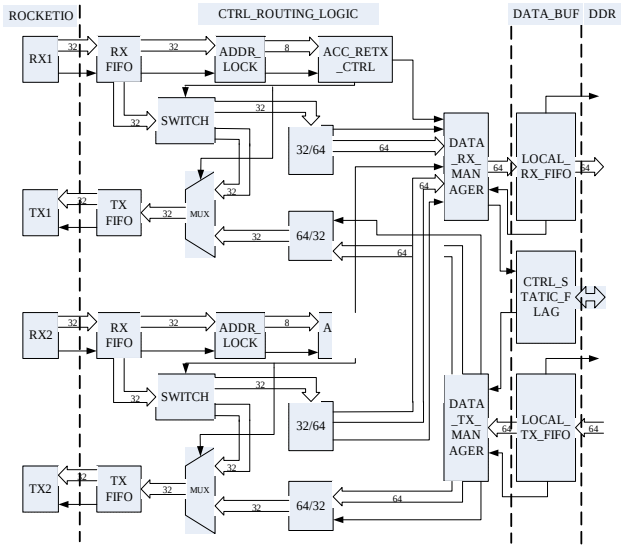

图 5　数据接收/发送控制及路由选择逻辑设计功能框图

由图 5 可以看出，数据接收/发送控制及路由选择逻辑部分的功能主要体现在以下几个方面：

(1)单卡支持全双工的数据通信，网络适配器中设计了 2 套收发电路 RX1、TX1 和 RX2、TX2，分别对应双环网的内环和外环。

(2)设计了接收、发送仲裁控制器 DATA_RX_MANAGER 和 DATA_TX_MANAGER，分别用于仲裁接收和发送数据的通道。在令牌环网中，某一时刻只有取得令牌的节点计算机可以发送数据，从而数据发送不存在冲突问题，因此，数据发送控制逻辑比较简单，发送仲裁控制器 DATA_TX_MANAGE 只是根据控制信号选通相应的数据通路；但在有双令牌的双环网结构中，数据包的接收存在两个节点计算机向同一节点计算机发送数据而产生冲突的情况，为此，接收仲裁控制器 DATA_RX_MANAGER 中集成了解决数据包冲突的算法：先到达的数据包得到优先接收，若同时到达，则外环的数据包优先。

(3)硬件实现数据包路由协议[1]，节点计算机接收到数据包后，数据首先存入接收缓冲区(RX FIFO)，地址锁存逻辑(ADDR_LOCK)自动提取数据包的目的地址和源地址给数据接受/转发控制器 ACC_RETX_CTRL，数据接受/转发控制器ACC_RETX_CTRL内部集成了第 2 节所述的数据包路由判断算法，根据数据包的目的地址、源地址及本节点计算机的地址对数据包的处理做出判断，并输出相应的控制信号，实现数据包路由的硬件自动识别。

此外，为了减小数据包传输的延迟，在本设计中，各个

数据缓冲区FIFO的存取机制采用了高效率的ON-THE-FLY机制[6]，在数据包未完全到达之前，根据数据包头的信息，对数据包的处理做出判断，提高了数据包的传输效率。

3.3 高速数据并/串(串/并)转换及串行发送/接收逻辑设计

高速数据并/串、串/并转换逻辑设计通过配置 FPGA 内嵌的高速串行收发器 ROCKETIO 模块来实现，最高速率可达到 3.125Gb/s，收发器完成 CRC 校验、编解码、串/并(并/串)转换和串行数据的发送/接收。在本设计中，数据位宽设为 32 位，编码方式采用 8B/10B 编码，串行传输速率设置为 2 Gb/s。由于篇幅所限，对 ROCKETIO 模块的具体设置不再详述，读者可参阅 Xilinx 公司的设计规范。

4 DDR DIMM 总线光互联网络适配器的性能分析

本文设计的 DDR DIMM 总线光互联网络适配器基于 Xilinx 公司 Virtex-Ⅱ Pro 系列的高档 FPGA 实现，DDR DIMM 总线接口逻辑及高速串行收发器 ROCKETIO 部分的时序逻辑仿真波形分别如图 6、图 7 所示。

图 6 DDR DIMM 总线接口仿真时序图

图 7 ROCKETIO 发送数据仿真时序图

由时序仿真图可以看出，DDR DIMM 总线接口逻辑部分的工作时钟频率是系统主时钟频率的 2 倍，带宽为 64bit×200MHz=12.8Gb/s；高速串行收发器 ROCKETIO 部分的工作参考时钟频率为 100MHz，设置参考时钟的频率为串行收发速率的 1/20，因此，单路 ROCKETIO 的传输速率为 100MHz × 20=2Gb/s，网络适配器的每路收发通道采用 4 路 ROCKETIO 实现，因此网络传输带宽为 8Gb/s。

本文设计的网络适配器的最大传输带宽为 8Gb/s，是 PCI 总线光互联网络适配器(1.067Gb/s)的 7.5 倍，可以满足机群系统高带宽、低延迟的性能要求。

5 结论

本文设计实现了一种基于 DDR DIMM 总线接口的高带宽、低延迟、高可靠光互联网络适配器。该网络适配器与基于 PCI 总线的光互联网络适配器相比，传输带宽有了比较大的提高；同时还具有支持全双工通信、硬件路由等功能。将其应用到机群系统中，可以解决互联网络的“瓶颈”问题，进而提高机群系统的整体性能。

参考文献

1 井文才，田劲东，张　珣. 用于机群系统的高速光互联网络接口卡设计[J]. 光电子•激光, 2000, 11(1): 7-10.

2 Noboru T, Junji Y, Hiroaki N. MEMOnet: Network Interface Plugged into a Memory Slot[C]//Proc. of IEEE International Conference on Cluster Computing. 2000.

3 Noboru T, Yoshihiro H, Hironori N. A Low Latency High Bandwidth Network Interface Prototype for PC Cluster[C]//Proc. of International Workshop on Innovative Architecture for Future Generation High-performance Processors and Systems. 2002.

4 Noboru T, Akira K, Tomotaka M. Preliminary Evaluations of a FPGA-based-Prototype of DIMMnet-2 Network Interface[C]//Proc. of International Workshop on Innovative Architecture for Future Generation High-performance Processors and Systems. 2005.

5 Akira K, Yoshihiro H, Yasuo M. Evaluation of Network Interface Controller on DIMMnet-2 Prototype Board[C]//Proc. of the 6th International Conference on Parallel and Distributed Computing, Applications and Technologies. 2005.

(上接第 189 页)

6 小结

本文对经典的阈值选取方法 Otsu 准则进行了推广，结合图像熵与 Otsu 法的优点，提出了一种改进的算法，并且采用局部递归的分割思想，在保证分割质量的前提下大大减少了阈值搜索时间，提高了分割效率。该算法计算量相对较小，受灰度值线性变化和平移变化的影响小，有一定的抗噪声的特点，而且算法程序的可扩展性好。实验结果表明，该方法计算速度比传统的方法有实质性的提高，这得益于递归思想的运用，避免了传统方法中准则函数的重复计算。这种改进的方法可运用到一些实时性要求比较高的图像处理系统中。

参考文献

1 罗西平，田　捷. 图像分割方法综述[J]. 模式识别与人工智能, 1999, 9(3): 300-312.

2 Otsu N. A Threshold Selection Method from Gray Level Histogram[J]. IEEE Trans. on Syst. Man, Cybern., 1979, 9(1): 62-66.

3 Kittler J, Illingworth J. Minimum Error Thresholding[J]. Pattern Recognition, 1986, 19(1): 41-47.

4 Dunn S M, Harwood D, Davis L S. Local Estimation of the Uniform Error Threshold[J]. IEEE Trans. on PAMI, 1984, 6(6): 742-745.

5 吴　谨，李　娟. 基于最大熵的灰度阈值选取方法[J]. 武汉科技大学学报(自然科学版), 2004, 27(1): 58-60.

6 Sahoo P K, Tanis S, Wong A K C. A Survey of Thresholding Technique[J]. Computer Vision Graphics Image Processing, 1988, 41(2): 233-260.

7 章毓晋. 图像分割[M]. 北京: 科学出版社, 2001.

8 付忠良. 图像阈值选取方法——Otsu 方法的推广[J]. 计算机应用, 2000, 20(5): 37-39.

9 孔　明，孙希平. 一种改进的基于类间方差的阈值分割法[J]. 华中科技大学学报(自然科学版), 2004, 32(7): 46-47.