直接内存通信技术及其网卡原型研究

陈颖图[1] 王爱林[1] 张 炎[1] 刘君瑞[2]
(中航一集团第六三一研究所 西安 710068)[1] (西北工业大学计算机学院 西安 710072)[2]

摘 要 针对基于PCI等传统I/O总线的网络I/O方式中网络通信性能受到相应总线接口限制的问题,提出了直接内存通信技术DMC(Direct Memory Communication,DMC)。使用此技术的DMC网卡可直接插入内存插槽中,DMC网卡上的存储空间被系统预留作为通信专用区,并使用与普通内存相同的方法进行管理和访问。待发送的数据使用写普通内存的方法直接写入DMC网卡的通信专用区中,对DMC网卡的通信专用区中收到的网络数据,用户可使用读普通内存的方法获得,从而实现了计算机内存之间的直接通信。因此,DMC技术使网络通信速度不受PCI等传统I/O总线的限制,省略了传统通信机制中网卡设备和内存之间的数据拷贝工作,具有通信速率高、通信延迟小及操作简单的特点。在高速光纤通道交换网中设计了DMC网卡原型,证明了DMC技术的正确性和可行性。

关键词 网络I/O,直接内存通信,通信专用区,内存预留,FIFO
中图法分类号 TP393.02 **文献标识码** A

Research of Direct Memory Communication and NIC Prototype

CHEN Ying-tu[1] WANG Ai-lin[1] ZHANG Yan[1] LIU Jun-rui[2]
(AVIC Xian Aeronautical Computing Technique Research Institute,Xi'an 710068,China)[1]
(College of Computer Science,Northwestern Polytechnical University,Xi'an 710072,China)[2]

Abstract At present,the communications speed on PCI is limited by PCI bus,the number of the computer's memory slots is increasing,and the management of memory becomes advanced. So,in this paper,the direct memory communication method,abbreviated as DMC,was put forward. The NIC(the network interface card) based on DMC is inserted into a memory slot,and the memory on DMC-NIC is reserved as communication precinct. When user sends data,he uses the method of writing memory to place the data into communication space,and when user receives data,he uses the method of reading memory to get the data from communication space. So,the user can accomplished the direct point-to-point communications between the computers through accessing memory. The communications speed is not limited by I/O bus,and the copy between memory and NIC is omitted. We also applied DMC into the high-speed fibre channel switch network,and designed the FIFO-DMC NIC to prove that DMC is right.

Keywords Communication I/O,Direct memory communication,Communication precinct,Memory reserved,FIFO

1 引言

现有的计算机通信网卡大多是插在PCI或PCI扩展总线等传统I/O总线插槽中,如以太网网卡、Myrinet网卡等,通信活动要经过PCI等总线接口,通信速度受到PCI等总线接口的限制,用户必须采用设备驱动程序才能实现通信。在目前计算机内存插槽较充裕的情况下,本文提出了一种直接内存通信(Direct Memory Communication,DMC)技术,基于该通信机制的DMC网卡插在内存插槽中,DMC网卡上的存储空间位于内存物理地址最高区,被操作系统预留作为通信专用区,用户使用读写普通内存的方法读写DMC网卡的通信专用区,计算机之间可以通过内存访问进行直接通信,无需在网卡设备和内存之间进行数据拷贝,通信速度不受PCI等传统总线接口的限制,通信延迟减小,通信带宽提高。作者还设计了高速光纤通道技术交换网中的DMC网卡原型,证明DMC技术是正确的、可行的。

下面详细描述DMC通信机制以及高速光纤通道交换网中DMC网卡原型机的设计方案。

2 直接内存通信技术DMC

2.1 DMC技术思想

现有的网卡通信是把网卡作为计算机的一个外围设备来进行操作的,用户先要把数据从内存送到网卡设备里,网卡才能把数据发送出去,接收数据时则需要把数据先接收到网卡设备的存储空间中,然后再把数据拷贝至内存中。这种实现方式避免不了网卡设备和内存之间的数据拷贝,并且网卡作为一种外围设备通信活动也受到相应总线接口如PCI总线接口的限制。

DMC技术是把网卡作为一块特殊的内存,插在物理地址最高内存区的内存插槽中,使得网卡和主机之间的数据交换如同主机访问内存一样,主机的网络通信活动与读写内存一样,这就避免了原有的通信过程中网卡设备和内存之间的数

陈颖图 男,高级工程师,主要研究方向为计算机应用;王爱林 男,高级工程师,主要研究方向为计算机应用;张 炎 女,高级工程师,主要研究方向为计算机管理;刘君瑞 女,博士,副教授,主要研究方向为计算机系统结构和计算机基础教育。

据传输，因此把这种通信机制称为直接内存通信(Direct Memory Connection，DMC)，并且把基于 DMC 技术实现的网络适配器称为 DMC 网卡。DMC 通信机制可以应用于多种网络环境中，下面以高速光纤通道交换网作为应用环境，对 DMC 通信机制技术细节进行介绍。

DMC 技术的详细过程描述如下：首先，把网卡作为一块特殊的内存插入最高内存区的内存插槽中，修改操作系统对内存的最高物理地址区部分空间，即 DMC 网卡上的部分存储空间进行注册预留，将其作为 CPU 和网卡共享的通信专用区，只允许与网络通信活动相关的用户读写，其他系统进程无权访问。我们把通信专用区按照以下 4 种用途进行分配：接收缓冲区、发送缓冲区、网卡命令区以及网卡状态区。然后，根据相应的通信协议如 FC 协议进行网络通信活动，通过访问通信专用区，控制具有通信控制逻辑、并串转换/串并转换器和光收发器等部件的 DMC 网卡，进行计算机之间的点对点直接内存通信。

2.2 DMC 通信活动描述

DMC 通信机制中主要通信活动描述如下：

(1)发送数据：通信源节点发送数据时，只需用户使用写普通内存的方法将数据写入通信专用区的发送缓冲区中，同时把发送命令写入通信专用区的网卡命令区。DMC 网卡上的通信控制逻辑根据网卡命令区的命令解析结果，从通信专用区的发送缓冲区中取出数据发送至网络；

(2)接收数据：在 DMC 通信机制的应用环境下，通信目的节点配置有相同的 DMC 网卡，网络上的数据经网卡的通信逻辑接收后放入通信专用区的接收缓冲区，同时网卡控制逻辑修改通信专用区中的网卡状态信息。当用户需要获得网络数据时，只需使用访问普通内存的方法读通信专用区的接收缓冲区数据即可。

因此，DMC 通信机制实现了两台计算机内存之间的直接通信。用户感觉不到 DMC 网卡的存在，使用访问普通内存的方法就可以实现计算机间的点对点直接通信。

2.3 DMC 通信机制的体系结构

图 1 描述了直接内存通信技术的体系结构。

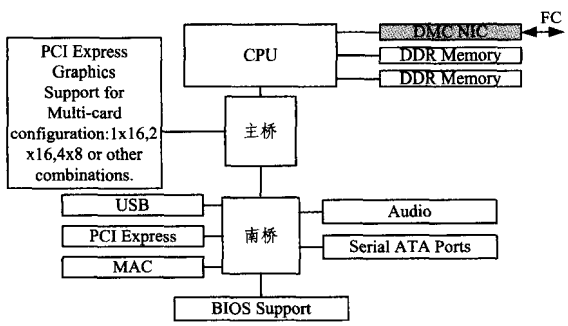

图 1　直接内存通信机制的体系结构

注：图 1 参考了 Intel X58 Express 芯片组的布局图，其中内存控制器在 CPU 芯片内。目前很多高端的通用 CPU 都在内部集成了内存控制器，而相对一些低端的计算机系统，内存控制器在主桥芯片内。

从图 1 可以看出，在直接内存通信体系结构中，DMC 网卡和内存处于对等的位置，对 CPU 是透明的，CPU 使用操作普通内存的方法操作 DMC 网卡的通信专用区，用户通过对 DMC 网卡的通信专用区进行读写来完成网络通信活动。因此，DMC 通信机制避免了数据在网卡设备和内存之间的拷贝，并且通信速率也不再受传统 I/O 总线的限制。

3 DMC 网卡原型机——FIFO-DMC 网卡研究

3.1 FIFO-DMC 网卡的体系结构

基于直接内存通信技术 DMC 以及 FPGA 片上的存储区域 FIFO(First Input First Output，FIFO)和寄存器，作者设计了 DMC 技术，并将其应用于高速光纤通道交换网的原理样机 FIFO-DMC 网卡。该网卡是在 DIMM DDR 内存总线规范上扩展实现的，其逻辑结构可用图 2 描述。

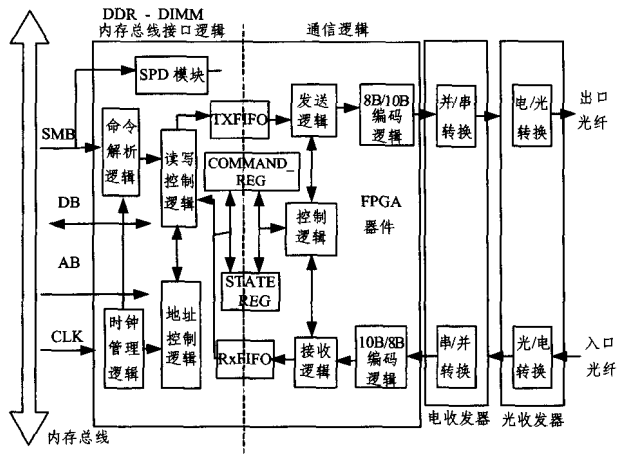

图 2　FIFO-DMC 网卡逻辑结构图

由图 2 可以看出，FIFO-DMC 网卡由一对出口/入口光纤、光收发器、电收发器、FPGA 可编程器件组成。FPGA 为网卡的控制芯片，是整个系统的核心。其内部逻辑可分为 DDR-DIMM 内存总线接口逻辑和通信逻辑两个功能模块，各部分器件的功能分别如下：

(1)光纤用于连接 FIFO-DMC 网卡和高速光纤交换网络的对应端口。

(2)光收发器负责进行光信号与差分电信号之间的转换。

(1)电收发器用于数据的串/并转换。

(4)FPGA 可编程器件从功能上可以分为两大部分：DDR-DIMM 内存总线接口逻辑和通信逻辑，其中 DDR-DIMM 内存总线接口逻辑包括 5 个模块，分别为 SPD 模块、命令解析逻辑、时钟管理逻辑、地址控制逻辑、读写控制逻辑，用以完成网卡的通信逻辑以及网络用户和通信专用区之间的信息交互。通信逻辑包括发送逻辑、CRC 校验逻辑、接收逻辑、控制逻辑、8B/10B 编码逻辑与 8B/10B 解码逻辑 5 个模块，用以实现真正的网络传输活动。

3.2 FIFO-DMC 网卡的主要功能模块

下面对 FIFO-DMC 网卡中的主要功能模块进行简要介绍。

(1)FIFO-DMC 网卡的通信专用区

按照 DMC 通信机制的要求，FIFO-DMC 网卡的通信专用区按用途分为 4 块：接收缓冲区、发送缓冲区、网卡命令区以及网卡状态区。

接收缓冲区和发送缓冲区采用 FPGA 片上 FIFO 实现，数据接收 FIFO 命名为 RxFIFO，用来存放网卡接受逻辑从网上接收的数据，用户使用读普通内存的方法就可获取。数据发送 FIFO 命名为 TxFIFO，用来存放用户待发送的数据，用户使用写普通内存的方法把数据放入发送 FIFO 中，而后网卡的发送逻辑读取 FIFO 的内容进行传输。数据接收 FIFO

和数据发送 FIFO 的容量都为一帧数据的大小。

网卡命令区和网卡状态区采用 FPGA 片上 64 位寄存器实现，网卡命令区即网卡命令寄存器 COMMAND_REG 存放用户发出的网卡命令。网卡状态区即网卡状态寄存器 STATE_REG 存放网卡的各种状态信息。DMC 软件或网卡通信逻辑在对网卡进行操作前，读取 COMMAND_REG 和 STATE_REG 的内容，判断相应位，再根据结果执行相应动作来防止冲突。寄存器中各位置"1"表示有效，在系统初始化时全部清零。

(2)SPD 模块

SPD 模块使设计出的 FIFO-DMC 网卡设备保持与普通内存相同的稳定性，能够正确地被北桥芯片或者 CPU 芯片中的存储管理器识别。在 FIFO-DMC 网卡中使用 VHDL 语言编程模拟 SPD 芯片的工作。

通过分析，FIFO-DMC 网卡的 SPD 模块只需使用 SPD 芯片的 5 个引脚：SA0、SA1、SA2、SDA 和 SCL，并且 BIOS 对 SPD 模块只执行读操作(Random Address Read)，所以 SPD 模块的结构比较简单，主要包括 START 状态控制以及 Random Address Read 命令响应两个功能模块。其中，START 状态的控制逻辑比较简单，主要依靠作为从设备的 SPD 模块监听串行数据线(SDA)和串行时钟线(SCL)来产生，此处不再赘述。Random Address Read 命令的响应由一个状态机来实现，在不同的状态完成相应的工作。其状态转换如图 3 所示(该状态图由软件 Synplify Pro 编译程序代码后生成)。

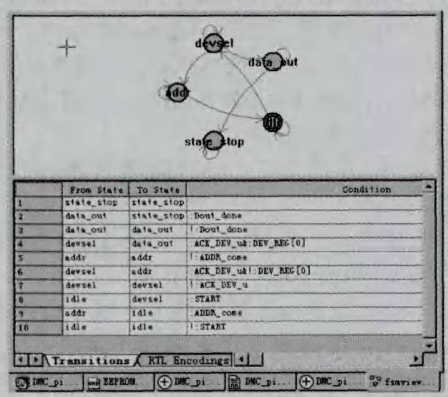

图 3　SPD 状态转换图

(3)命令解析逻辑

命令解析逻辑主要接收来自 DDR-DIMM 内存总线接口的各种内存访问命令，并对命令进行解析。FIFO-DMC 网卡的命令解析逻辑由一个状态机控制，状态转移时设置特定的信号，由该信号触发相应的读、写逻辑。状态转换图如图 4 所示。

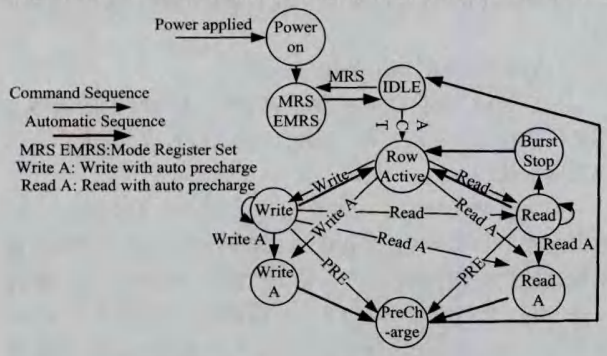

图 4　DDR DIMM 状态转换图

(4)地址解析逻辑

此模块在内存访问命令到来时，控制地址总线上的行地址和列地址等信息进行地址译码工作，寻址被访问的存储单元，使得各种数据信息能够在网络用户、网卡的通信逻辑和内存之间正确地完成读写工作，协助 FPGA 中控制逻辑实现网卡的通信活动。

由于 FIFO-DMC 网卡中使用了 FPGA 芯片上的 FIFO 和寄存器来模拟通信专用区，因此用户操作通信专用区时只有 4 个地址信息：TxFIFO 写端口对应的虚拟地址，RxFIFO 读端口对应的虚拟地址，命令寄存器 COMMAND_REG 对应的虚拟地址以及网卡状态寄存器 STATE_REG 对应的虚拟地址。地址解析逻辑根据用户访问的虚拟地址信息定位到通信专用区的某个部分即可。

(5)读写控制逻辑

根据地址解析逻辑寻址出的通信专用区空间以及命令解析逻辑解析的结果，对 FPGA 的寄存器或者 FIFO 进行读写操作。外部与通信专用区之间传输的数据信息主要有 3 类，分别是通信活动中的数据、用户写入网卡命令区的网卡命令以及网卡的状态信息。

3.3　FIFO-DMC 网卡的软件实现

直接内存通信技术 DMC 得以实现的重要根基是预留部分内存空间供 DMC 通信机制进程专用，这依赖于 Linux 操作系统提供的灵活机制。因此，DMC 网卡的软件功能包括：

(1)实现通信专用区的物理内存预留：依据 Linux 操作系统对内存的管理办法，将 FIFO-DMC 网卡插入内存插槽的高端，使其存储空间即通信专用区处于内存区的物理地址最高端。然后，我们借助于 Linux 内核启动时能接收某些命令行选项或启动时参数的特性，修改系统引导程序中的启动配置参数 mem，限定内核使用的内存数量。实际物理内存中大于 mem 值的部分就是预留的内存空间，系统不会使用这片物理内存。

(2)实现通信专用区内存的映射：由于 Linux 操作系统是一个虚拟内存系统，访问内存是基于虚拟地址空间的，因此为了能够使用被预留的通信专用区空间，需要把这部分物理内存正确映射到虚拟内存空间中。Linux 操作系统提供了至少 3 种实现内存映射的方法，可以在系统不同时刻将通信专用区映射为 I/O 内存、内核空间内存或普通用户空间，考虑到 DMC 技术中通信专用区需要在用户态下进行访问，作者最终选择使用 mmap 设备操作方法来实现通信专用区的内存映射。并且，由于在 FIFO-DMC 网卡的设计中使用 FPGA 片上 FIFO 和寄存器模拟实现通信专用区，因此 DMC 软件实现对通信专用区映射之后，只需要网卡命令寄存器、网卡状态寄存器、数据发送 FIFO 的写端口和数据接收 FIFO 的读端口 4 个虚拟地址。

(3)实现用户对通信专用区的访问接口：由于 FIFO-DMC 网卡硬件逻辑中提供了将通信专用区作为普通内存管理和访问的功能，因此用户可以使用访问普通内存的方法访问通信专用区。

3.4　FIFO-DMC 网卡的功能验证测试

3.4.1　FIFO-DMC 网卡的运行测试平台

FIFO-DMC 网卡的测试平台采用 PC 机，CPU 为 Intel(R) Pentium(R) 4，北桥芯片为 Intel 的 RG82865PE SL722。FPGA 采用 ALTERA 公司 Cyclone 的 EP1C4 芯片，串/并转

换使用德州仪器的 tlk2501，光电转换则选用美国 Finisar 公司的产品 FTRJ 8519。示波器为 Tektronix 的 TDS3052。为了降低调试的难度，通过 BIOS 设置，将内存时钟频率 200MHz 改为 100MHz。

3.4.2 FIFO-DMC 网卡的运行测试结果

图 5 为开机时 SPD 模块从示波器上看到的波形图。参照 I²C 协议的相关内容可加深理解。

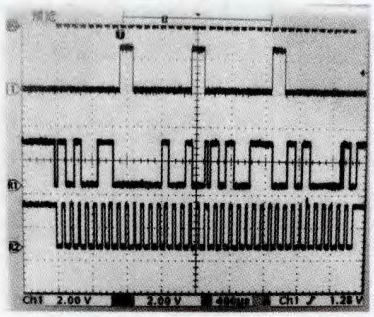

图 5 SPD 模块波形图

图 6－图 9 为 DDR DIMM 接口模块接收到各种网卡命令后从示波器上看到的波形图。图中，FIFO-DMC 网卡在通道 1(图中标号为 1 的那条线)的上升沿锁存信号。

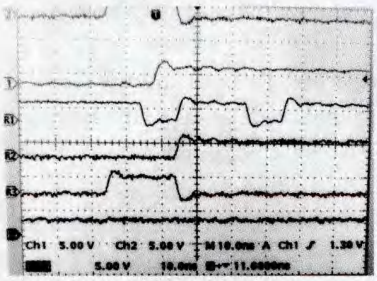

图 6 Active 命令

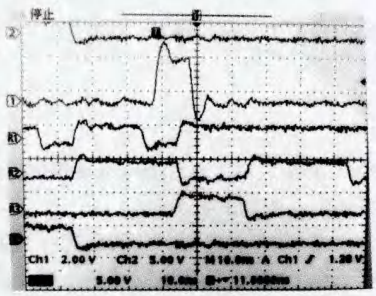

图 7 Write 命令

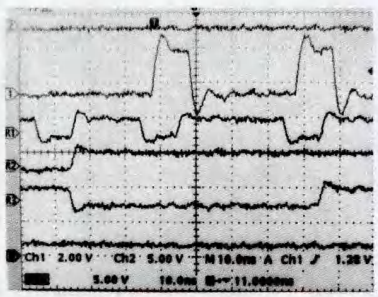

图 8 Read 命令

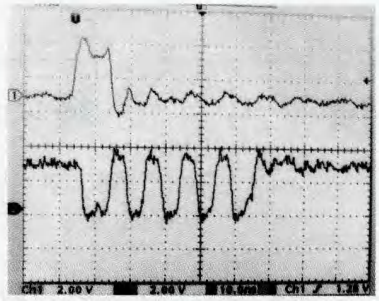

图 9 DQS 波形

经测试，FIFO-DMC 网卡能在开机的 BIOS 自检中被识别为内存设备，正确响应 CPU 的读写命令，并能在操作系统引导时预留共享存储区，证明了直接内存通信 DMC 通信机制是正确的和可行的。

结束语 本文提出的直接内存通信技术 DMC 把通信网卡作为一块特殊的内存插在内存插槽中，用户通过内存访问操作实现计算机之间的直接通信，使得计算机的网络通信活动不受 PCI 等传统 I/O 总线接口的限制。相比基于传统 I/O 总线的网络通信方式，基于 DMC 技术的通信机制中数据无需在网卡设备和内存之间进行拷贝，并且网络通信性能不再受传统总线接口的限制，通信速度更高，传输延迟更小，通信手段更简单。在后续工作中，笔者将继续研究内存管理方法和 Linux 操作系统细节，在保证 DMC 网卡的工作稳定性和通信性能不受影响的前提下，增加 DMC 技术的实用性、操作简便性以及可推广性。

参 考 文 献

[1] Budruk R, Anderson D, Shanley T. PCI Express System Architecture[M]. 北京：电子工业出版，2005

[2] Tanabe N, Hamada Y, Nakajo H. A low latency high bandwidth network interface prototype for PC cluster[C]// Proceedings of the International Workshop on Innovative Architecture for Future Generation High2Performance Processors and Systems. Big Island, 2002：87

[3] Gorman M. 深入理解 Linux 虚拟内存管理[M]. 北京：北京航空航天大学出版社，2006. 5

[4] Matzigkeit G, Okuji Y K. The GUN Grub manual[OL]. http://www. gnu. org/software/grub/manual/grub. html

[5] Corbet J. Alessandro rubini & Greg Kroah-Hartman, Linux Device Drivers[M]. O'REILLY, 2006

[6] Jedec, Sdram P C. Serial Presence Detect (SPD) Specification [S]. Revision 1. 2A

[7] Jedec, PC2100 and PC1600 DDR SDRAM Registered DIMM Design Specification[S]. January 2002, Revision 1. 3

[8] 张晓彤，王景存，王沁，等. 基于 DDR 内存总线的高速网络接入技术[J]. 北京科技大学学报，2007，29(11)：1158-1162

[9] 王乐，张晓彤，李磊，等，Linux 下的 DDR DIMM 总线接口设备检测方法[J]. 计算机工程 2007，33：256-258

[10] 雷艳静，魏建军，王玥，等. 面向集群系统的高速光纤传输网络接口卡设计[J]. 计算机工程与应用，2006，13：19-23

[11] 雷艳静. 面向 MNWF 的信令寻径式光纤通道先锋交换网研究[D]. 西安：西北工业大学计算机学院，2009