一个分布式微机系统及其分布式操作系统的设计与实现

何炎祥 刘 蓬

(武汉大学计算机科学系，430072)

摘要：本文扼要讨论了分布式微机系统MDS的设计与构造，主要探讨MDS的互连结构、通信接口、通信机制，以及分布式操作系统MDS/DOS的设计思想与部分实现技术。MDS系统已初步实现了分布式系统的基本功能。

关键字：互连结构，通信机制，分布式命令解释器，报文交换。

一、引 言

结合工作和科研的需要，我们以广为流行的IBM PC/XT及其兼容机为结点处理机，设计并初步实现了一个分布式微机系统MDS。本文首先简单介绍MDS的硬件环境设计，内容包括网络的拓扑结构、通信接口及互连技术等，然后扼要讨论分布式操作系统MDS/DOS的设计与实现。

二、MDS系统的硬件环境设计[1~4]

1. MDS的拓扑结构设计

(1) 拓扑结构设计原则：

① 当网络中的结点数增加时，网络的直径增加缓慢。

② 与每台机器直接相连的机器数，不会随网中机器台数增加而变大，即网中结点的邻接结点数不大于C(C是一常数)。

③ 存在一个有效的算法，使得从一台机器的地址编号很快就可得知与其通信的路径。

④ 网络中一台或几台机器不能正常工作时，网络仍能保持连通性。

⑤ 网络中各结点的负载比较均匀。

(2) MDS的互连结构：

MDS允许方便地搭接成各种非共享通路拓扑结构，以适应不同处理的需要，如图1所示。其中，结点A用接口a_1与B结点的接口b_1互连，B结点通过接口b_2与D结点的接口d_1互连等等。由于拓扑结构是任意的，使得我们很难用一种算法求出从某结点到另一结点的信息传输路径。对此，我们为每个结点设置一个"结点配置"文件，它与实际的网络拓扑相对应。每个结点的"结点配置"文件是不一样的，其主要内容为：从本(地)结点到系统中其它结点所需经过的路径表。图2示出了结点A到其它结点的路径表。例如，若打算从A结点传送信息到D结点，其过程如下：在A结点，

本文于1992年5月收到。何炎祥：1952年1月生，1975年毕业于武汉大学数学系软件研究生班，1986年8月获美国俄勒冈(Oregon)大学计算机及信息科学系硕士学位，1989年晋升为武汉大学副教授。1993年升为教授，主要研究领域为分布式操作系统、并行程序设计语言及编译系统。刘蓬：1991年毕业于武汉大学计算机系，获硕士学位，主要研究领域是分布式计算系统。

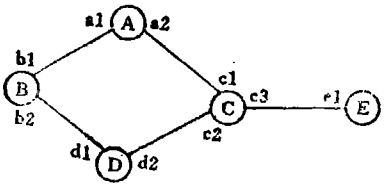

图 1　MDS 的拓扑结构示意(注：图中上圆圈内为A)

到 结 点	通过接口
B	a_1
C	a_2
D	a_1
E	a_2

图2　结点 A 的路径表

查得去D结点的接口为 a_1，则通过 a_1将信息传到结点 B。在结点 B 查得到 D 结点 的接口为 b_2，于是通过 b_2依次将信息传到D结点。采用这种信息转发技术，可实现网内点到点的通信。

2。MDS 的接口设计

(1) 异步通信接口及串行接口的连接方式

IBM PC 及其兼容机已提供了 8250 异步通信接口板，其中 8250 异步通信控制芯片是整个接口的心脏，它负责将主机来的并行数据加上适当的起始位、停止位、校验位后串行地发送出去，同时又可接收串行数据，并自动地去掉校验位等，拼成数据字节交给主机。它提供如下功能：

① 提供一个 RS-232-C 接口，需要时还可用电流环方式工作；

② 数据传输率可在 50 baud 到 9600 baud 之间选择；

③ 具有控制 MODEM 功能和完整的状态报告功能；

④ 具有线路隔离、故障模拟等内部诊断功能；

⑤ 具有独立的中断优先功能，

⑥ 传输数据格式可选择 5、6、7 或 8 位字符；奇校验、偶校验或无校验；1.5 或 2 个停止位。

在 MDS 中，各结点间的距离较近，故在设计时未使用 MODEM，而采用"零"调制解调器连接方式。图 3 给出了两结点连接图。

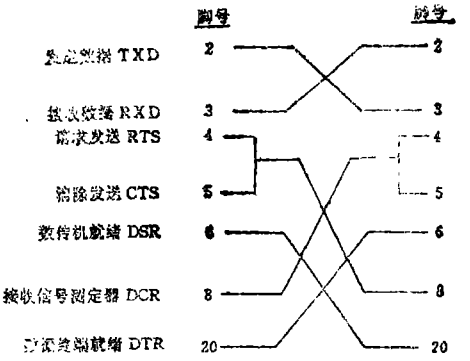

图 3　"零"调制解调器的接法

(2) 并行接口及其连接方式

由于每台 PC 机最多允许两个标准的异步通信接口，且其通信的最高速率为 9600 baud，难以满足 MDS 的要求。为了扩大系统中每个结点的连接度数，提高传输率，我们还为每个结点配置了多个并行通信接口。(注：图 3 中 TXD 及 RXD 中的 X 应为零)

并行接口板的核心芯片为 8255。一个并行通信接口板上可允许有多个 8255 芯片，以便 MDS 系统中的一个结点可以和多个其它结点相连接。由于 IBM PC/XT 及其兼容机只有中断请求 2···IRQ 2 信号保留给用户使用，故需采用一定的中断控制逻辑来连接众多的 8255 中断请求源。解决的办法是将所有 8255 的中断请求信号"逻辑或"之后接到主机的 IRQ2 上，当进入中断服务程序后，用软件进行判断，找出发中断请求的 8255 芯片。图 4 示出了结点 i 的 a 口与接点 j 的 b 口相连的逻辑图。

一个结点要与另一个结点进行半双工的并行通信，在并行通信板的连接器中，须包含下述的信号线：8 条数据输入线，8 条数据输出线，4 条选通信号线，地线等。这样，连接两个结点以实现半双工进行通信的连接逻辑图，如图 5 所示。

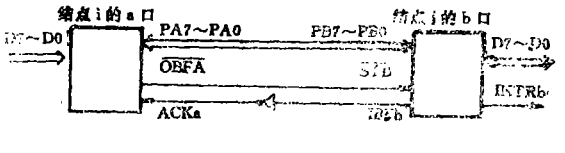

图 4 结点 i 与结点 j 相连的逻辑图

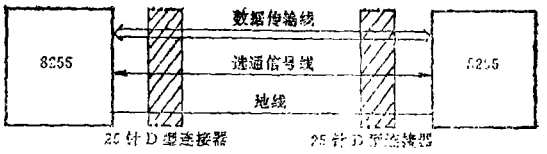

图 5 两结点通过并行接口相连的逻辑图

三、分布式操作系统 MDS/DOS 的设计

下面扼要讨论分布式操作系统 MDS/DOS 的设计与实现，内容包括：MDS/DOS 的开发策略、通信机制、分布式通信及资源管理模块 DCRM 和分布式命令解释器 DCI 的设计等[5,6,7]。

1. MDS/DOS 的开发策略

MDS/DOS 是在 MS-DOS 的基础上，主要通过增加分布式命令解释器，分布式通信与资源管理模块 DCRM，并建立相应的通信机制后修改扩充而成，其基本构成如图 6 所示。其中 DCRM 完成系统资源的申请、分配和回收；完成对串、并行接口板的控制及报文的发送、接收及多缓冲空间的管理；完成对远程文件的操作和执行等。DCI 完成对用户程序的管理及分布式命令的解释执行。

MDS 中的每个结点都配有相同的 MDS/DOS 核部分。它们各自处于平等的地位。整个系统支持多台处理机共同执行系统管理任务和用户程序，但"主处理机"是浮动的，可由一台切换为另一台。

2. MDS/DOS 的通信机制

(1) MDS 的通信系统

分布式软件由一组异步成份所组成。每一成份可在系统中某一结点上运行。在不同结点上执行的成份常常要交互作用，以协同完成指定功能。这种交互作用是通过信息交换来实现的，即通过一个分布式通信系统，把信息从源地传送到目的地。设计通信系统时，应考虑以下问题：通信协议；介质通路、链路和路径选择；程序语言中的通信成份等。

图 7 给出了 MDS 的通信系统模型。其中，通信设施级约定结点之间的物理连接，并对通信的

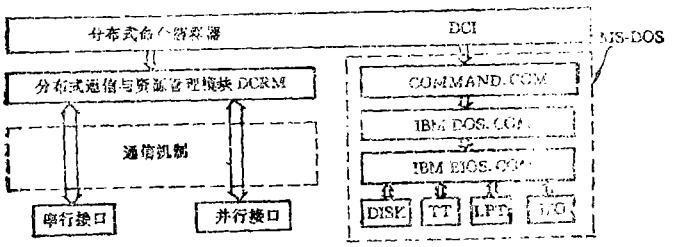

图 6 MDS/DOS 的基本结构图

物理参数作出规定等。操作系统级约定信息在通信线路中的表示以及传输路径的确定、目的地址的识别、信箱及消息缓冲空间的管理等。用户进程级约定各类信件的功能及使用方法。在 MDS/DOS 中，信件分为系统信件和用户信件两类。用户信件的含义由用户自行定义并由用户自行处理。系统信件由系统处理，且对用户是透明的。

在我们设计的通信机制中，除了前面提到的信息转发技术外，采用报文分组交换来实现结点

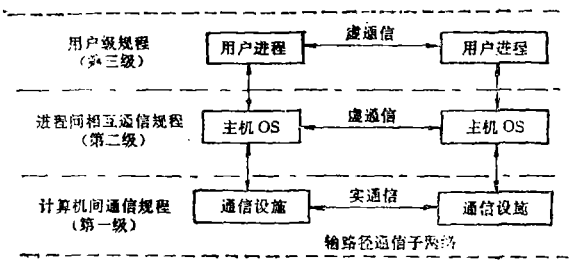

图 7　MDS 的通信系统模型

（注：图 7 中的输路径…，应为传输路径…）

到结点的通信也是它的一个特点。报文分组解决了变长报文的交换问题，方法是将长“报文”分割成许多“报文分组”，并在前面加上标示符，提供重新组装的信息。此外，还采用了报文与缓冲区的统一和多缓冲空间技术。报文与缓冲区的统一是指系统或用户程序在发送和接收消息的过程中，通信机制都在缓冲空间中工作，缓冲空间的大小与分组报文的大小一致。多缓冲空间技术提供了大量缓冲区，为实现异步型信息传递奠定了物质基础。

（2）信箱和消息缓冲栈通信方式

信箱是一个先进先出的队列链，它将所接收的报文（分组）按其到来的先后次序链接起来，并用头指针 MBhead 和尾指针 MBtail 分别指向报文的首尾。当 MBhead 和 MBtail 均为 NIL 时，表示信箱为空。基于信箱的操作原语有三个：1．加入一个报文：Addinmailbox(p)；2．按报文类型 t 取一个报文：Typefetch(t,p)；3．按发送地址 a 与报文类型 t 取一个报文：Typeaddrfetch(t,a,p)

在报文的发送和接收过程中，为了更有效地使用内存缓冲区，在 MDS 系统中定义了一种消息缓冲栈 MBS，它是由指针链接的多个空闲缓冲区所构成的链栈，如图 8 所示。当 MBStop 为

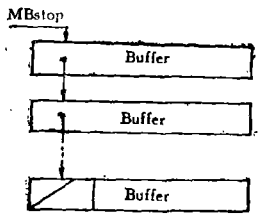

图 8　消息缓冲栈 MBS 的结构

长度（字节数）	内　容
1	报文长度
1	数据类型（报文类型）
1	发送地址
1	接收地址
60	正文数据（报文正文）

图 9　Buffer 的定义

NIL 时，表示消息缓冲空间栈为空。其中缓冲空间 Buffer 的结构见图 9，且 Buffer 的定义与报文格式是一致的。

基于消息缓冲空间栈的操作原语有两个：1．申请一个缓冲空间 Newb(p)；2．归还一个缓冲空间 Disposeb(p)。其中 p 为指针型变量。由于报文的发送和接收都要在缓冲空间进行，都要执行 Newb 和 Disposeb 操作原语，因此，若缓冲空间栈 MBS 为空，则不能再接收报文。

（3）通信协议

MDS 系统中报文的传输类似于 DMA 传输，因为除了“报文传输请求”是利用中断被接收方响应外，控制信息、数据正文等都以查询方式连续地读入与输出，最后才中断返回，为便于叙述，用 IRQ 代表通信板的中断请求信号。通信协议如下：

接收方：在 MBS 未溢出且接收方没有传输报文时，可以响应发送方送来的“报文传输中断请求”。响应后，关掉 IRQ，发送一个“准备好接收”信号，开始以查询计数方式接收报文字符，最后开放 IRQ 中断返回。

发送方：需传输报文时，先关掉 IRQ，向接收方发出“报文传输请求”信号。若接收方未回答，

则应给一定的时间来解决可能存在的通信线路竞争问题(接收方同时向发送方发送报文)，直到对方响应为止。然后以计数方式发报文，最后开放 IRQ 返回。

(4) 报文的发送和接收

发送一个报文时，先应申请一个空闲缓冲空间 P，并按规定的格式写好报文长度、接收地址及报文正文，然后调用 Submitsend(p)检查该报文的目的站点是否为本结点，若是，则调用 Addinmailbox(p)将报文放入信箱、否则调用 SendPKG(P)将报文 P 发送到指定的远程结点。

为了防止结点 i 向结点 j 发送报文时，恰好 j 也同时试图向 i 发送报文，在串行接口和并行接口发送报文的流程中，都须测试对方是否响应了"报文传输请求"没有。若对方没有响应，则有可能是对方也正试图向本结点发送报文。因此，必须留出一定的时间等待对方，以防止因彼此同时都等待对方的响应而造成死锁。

报文的接收程序是由中断驱动的。接收方一旦收到发送方发出的"报文传输请求"信号，在 MBS 非空的情况下，可以响应这个中断信号。响应后进入中断服务程序，按协议接收报文。由于 MBS 系统中同时使用了异步串行接口板和并行接口板，所以报文接收有两个中断服务程序，分别处理用异步通信口和用并行通信口接收报文的事务。

3．分布式通信与资源管理模块 DCRM

DCRM 是 MDS/DOS 中关于分布式通信与资源管理的低层软件，它应向高层的分布式命令解释器及用户程序提供服务，包括消息的发送与接收、系统变量的读取与修改，处理机的申请和释放，远程文件的存取与操作等。对此，高层软件有两种调用方式：(1) 利用 INT OCOH 功能调用指令；(2) 提交给"后台"命令处理器处理。这里所称的"后台"命令处理器是这样一个处理系统，它处理的命令需要多机之间的相互配合才能完成。这种命令是用与普通报文格式兼容的称之为"后台命令报文"来表示的。有关对 DCRM 的进一步讨论请见另文。

4．MDS/DOS 的分布式命令解释器 DCI

DCI 是系统与用户的界面，是用户使用 MDS 的直接工具。用户可通过它发各种命令，如文件的编辑、编译、运行分布式程序等等。用户须正确地回答用户名和口令后方可进入系统。DCI 在初始化时，要检查用户档案文件 user·sys 是否存在，若不存在，则必须建立一个名为 system、口令为 password 的系统用户。这是一个特殊用户，它可以建立其它上机用户的档案，使它们可以进入系统并使用 MDS/DOS 的命令。系统用户 system 的服务菜单如下：(1) 增加一个用户；(2) 删除一个用户；(3) 修改用户档案；(4) 显示所有用户信息。

系统在增加一个用户时，首先搜索用户表，若没有重复的用户名，则将增加的用户档案加到用户表的最末。删除一个用户时，只需将用户档案找到，然后从 userhead 链中删除它(指针 userhead 是指向系统中由所有用户档案记录链接而成的链表表头。初始化时，这个链表从 user·sys 文件中读入，退出时写入磁盘)。修改用户档案包括修改用户名、用户口令，以及系统用户的名字和口令。系统总是以 user·sys 中的第一个记录作为系统用户的档案。

在进入系统之前，DCI 需要同时扫描键盘输入和处理机状态，检查有无用户名的输入及本结点是否是其它结点的子结点，然后根据不同情况分别进入前台命令处理和键盘命令处理。

MDS/DOS 除了能执行原有 MS-DOS 的所有命令外，还主要增加了如下分布式命令：(1) 远程文件拷贝；(2) 列另一处理机的磁盘目录；(3) 请求远地结点执行 MS-DOS 的 RUN 命令；(4) 申请一个处理机；(5) 释放一个处理机；(6) 共享打印机；(7) 文件加锁；(8) 文件解锁。上述命令按执行方式分为前台和后台命令，其中(1)、(4)、(5)、(6)为后台命令。"后台命令"是由中断驱动的，是命令发出者发出一个"后台命令"到配合者结点，从而中断唤醒配合者的相应程

序共同完成的一种命令，命令发出者与配合者不需要满足"父子"关系。前台命令的发出者与配合者一定要满足配合者是命令发出者的子结点这种关系，这样，配合者子结点可以时刻监视信箱中有无父结点发来的前台命令型报文，若有，立即取出分析并执行它；对此，若不满足"父子"关系，则命令无法执行。

四、结 束 语

限于篇幅，本文仅扼要讨论了 MDS 及其分布式操作系统 MDS/DOS 的设计思想与部分实现方法。MDS 之所以选用 IBM PC/XT 及其兼容机作为结点处理机，主要是因为它具有很高的性能价格比，拥有广大的用户，硬件上提供了异步通信接口板，软件设计上层次分明、模块性好，易于改造。这样，MDS 系统中的结点既可通过标准的 RS-232-C 串行接口相连，也可通过自制的并行通信接口进行连接，实现起来比较容易。MDS/DOS 是在 MS-DOS 的基础上修改扩充而成，它具有如下特点：

(1) 处理机管理：MDS/DOS 是多用户操作系统。用户可为自己的任务申请系统中的处理机，以建立一个适用的子系统。

(2) 提供了运行分布式程序的环境。DCRM 提供的功能调用可方便地实现信息的发送、接收，处理机的申请与释放，文件的远程存取与执行等等。

(3) 实现了文件和打印机的共享。

(4) MDS/DOS 的命令与 MS-DOS 的命令兼容。

(5) 对整个分布式系统进行管理，不存在集中环节，因而具有较好的协同性、资源共享性和自治性。

我们只是在分布处理技术的实用化方面作了一点尝试。这一工作还是初步的，不少问题还有待继续探讨，如系统中的负载均衡、MDS/DOS 的容错和安全性，以及进一步扩充和完善 MDS/DOS 的功能等。

参 考 文 献

[1] B.W.Lampson et al., Distributed Systems, Architecture and Implementation, Springer Verlag(Berlin), 1983。
[2] 孙钟秀等，一个分布式微型计算机系统，计算机学报，1981，2。
[3] 武汉大学分布式系统研究组，《分布式并行处理系统探索》，武汉大学出版社，1984。
[4] H.Loriu, Aspects of Distributed Computing Systems, John Wiley & Sons, UK, 1988。
[5] 张福保，何炎祥，"分布式多处理机操作系统—DPOS"，小型微型计算机系统，1984，1。
[6] A.S. Tanenbaum et al., Distributed Operating Systems, ACM, Computing Surveys, Vol.17, NO.4, 12, 1985。
[7] 何炎祥，苏开根，刘锦臣，《分布式操作系统导论》，学术期刊出版社，1989。

Design and Implementation of a Distributed Microcomputer System and Its Distributed Operating System

He Yanxiang and Liu Peng
Wuhan University, 430072

Abstract: In this paper, a distributed microcomputer system (MDS) is presented with emphasis on its interconnection construction, communication interfaces, communication mechanisms and design ideas and implementation methods of the distributed operating system MDS DOS. Based on MS-DOS, MDS/DOS is constructed by adding a distributed communication and resources management modale(DCRM)and a distributed command interpreter(DCI), which is compatible with MS-DOS. The MDS system has realized fundamental functions of a distributed system.

Keywords: interconnection construction, communication interfaces, communication mechanisms, message exchange and resource management