计算机科学 2000Vol. 27№. 1

SNOW 系统通信协议——Homer*)

Homer—A Communication Protocol for SNOW System

龙 翔 吴文峻 高小鹏 李 未

(北京航空航天大学计算机系 软件开发环境国家重点实验室 北京 100083)

Abstract SNOW is a scalable parallel computer cluster which suports both massage passing and shared memory programming model and Homer is a high performance multiport reduced protocol used in the SNOW system. After carefully considering the requirement of the system network, the general hierarchy of Homer protocol is firstly introduced. And then a detailed description of the functionality, structure, interface and services of each layer is given

Keywords SAN, Reduced communication protocol, Multicast, Active message

1. 引言

NOW(Network of Workstations)系统是以商用工作站或高档微机作为处理节点,通过高速商用网或专用网络互联而构成的一种并行计算机系统,一般称之为并行机群系统。它具有可扩展性好、性能价格比高、用户投资风险小及软件资源丰富等优点。由国家863计划支持的、由北航承担研制的并行机群系统SNOW(Shared-memory-based NOW)是一套可同时支持共享存储和消息传递模式的、以高性能PC机为节点的规模可扩展的多计算机系统。该系统采用根据机群特点自行研制的、具有高带宽低延迟的互连网络。本文将在分析机群网络特点的基础上,详细给出SNOW系统网络互连协议Homer(High perf ormance multiport r educed p rotocol)各个层的具体设计。

2. 协议设计的总体考虑

构造高性能可扩展计算机群系统要求处理机间的互连网络具有高带宽、低时延和高可靠性的特点。所谓高带宽是要求互连网络给应用程序提供高传输速率的通信信道,低延时则要求消息从发送方经网络传送到接收方的时间延时短,高可靠性则要求互连网络能够容忍物理信道可能出现的各种错误,保证在异常情况下通信系统不会崩溃。要实现这些目标必须从机群互连网络协议的设计和实现两个方面着手,机群互连网络通信协议应当充分考虑到性能的要求,并有利于互连网络的高效实现。

整个机群互连网络由网络开关节点所组成的通信网络和主机互连接口两部分组成,这两者决定了互连网络的性能。开关节点性能以及相应的互连网络构造方式决定通信网络的带宽和时延性能,而互连接口的性能则决定了应用程序所能利用的网络带宽和用户消息在节点机部分的传输时延。大量的研究表明,往往是互连接口的性能跟不上通信网络性能的提高,使得互连接口成为整个互连网络的"瓶颈"。所以机群互连接口的性能对实现机群互连网络高带宽、低时延和高可靠性的目标是至关重要的。

在传统网络互连接口层次结构中,不同层的协议对应相应的协议处理部分。一般来说,除链路层协议大部分由互连接口硬件实现外,部分链路层功能和上面各层协议功能都是由操作系统各级软件承担的。

大量研究表明,以TCP/IP为代表的传统网络结构不适应机群系统的需要。主要体现在:

(1)对传送的数据单元所做的处理过多,这在机群通信网络中是完全不必要的。例如:传送层中传送层地址到网络层地址的映射问题,不同通信子网之间传输数据格式的转换问题等等。

(2)传统网络中许多功能,例如:差错控制、流控制功能在链路层、网络层和传送层重复出现,这降低了通信系统的效率。

(3)同一数据反复拷贝,极大增加了消息发送的时间。许多研究都表明,数据拷贝占整个发送、接收时间的65%。

TCP/IP等上层复杂协议管理机制,不但增加了

*)本文的工作得到国家863计划(863-306-ZD-03)资助。

消息收发的开销，而且占用了大量的 CPU 资源和存储资源。研究结果表明，在连接以太网的主机上，35%的通信时间都消耗在 TCP/IP 的协议处理开销和操作系统的开销之上[2]。

基于消息传递的 MPP 系统，其互连网络结构简单、功能精简。它表现在：

(1)没有冗余的处理功能，如：消息传送层和网络控制层直接使用底层互连部件的物理地址，避免了地址映射问题。

(2)各层协议功能不重复，如：消息传递层不负责差错控制，而由端口层协议完成。

(3)协议设计充分面向高带宽和低时延，如：消息传送层采用 Active Message 机制，降低了消息缓冲区管理的开销，减少了通信时延[3]。

机群间互连网络与基于 TCP/IP 的传统网络和基于 MPP 系统的互连网络都有所不同。其互连的物理介质一般都安置在室内，甚至就在一个房间里，信号传输时延在几个纳秒～上百纳秒之间，信道出错概率为 10^{-15}。因此，机群互连网络的数据链路层必须采用数据帧作为数据交换的基本单位，也必须具有数据单元映射功能，并解决帧定界和同步问题。尽管其信道出错概率比较低，但是精简的错误检测与恢复机制仍然是必不可少的，否则就无法满足高可靠性通信的要求。流控制机制在机群互连网络中是很重要的。由于机群系统各节点之间通信的频率和通信的数据量将远大于一般 LAN 网，所以网络中通信发生冲突的概率大大增加，通信的"热点"现象会变得频繁。有效的流控制机制可以保证数据帧不会因为收方缓冲区变满而丢失，而且将有助于网络的拥塞控制。

总之，在物理层和链路层上，机群系统的互连网络将不同于 MPP 互连网，应当基于现有的 LAN 网络技术，对其进行简化和改造，以适应高速中距离数据传输的特点，满足机群高速互连网络高带宽、低时延、高可靠性的要求。同时，在构造整个机群系统的互连网络系统高层协议时应借鉴 MPP 系统的经验，尽可能减少协议层次，减少协议中机群通信不需要的复杂功能，建立精简高效的机群互连网络。为此 Homer 的设计着重考虑了以下两点：

1)精简消息层协议，合理划分节点接口和网络的功能。整个 Homer 通信系统必须向上层用户提供可靠的消息传递平台。用户发出的消息应由 Homer 通信系统按序、可靠地发送到接收方。同时，由于发送和接收用户的异步行为，整个网络协议机制必须具备流量控制和拥塞控制机制，以避免出现网络死锁，避免因缓冲区的溢出导致用户数据的丢失。

在机群系统通信环境下，应当由机群互连网络确保消息数据帧按序、可靠地传送。这样互连接口的消息层协议就不必检查接收到消息数据的正确性、处理接收方的确认应答及对消息帧进行排序等工作。所以，Homer 互连网络必须向消息层提供永久虚电路的服务。一方面满足减少消息层协议的功能，另一方面避免因建立、撤消连接带来的不必要开销。

2)直接支持选播通信的交换式互连网络。Homer 高速互连网络采用交换式的方式，避免共享介质对互连网络可扩展性的限制。同时，以多端口的交叉开关构成拓扑灵活的互连网络。Homer 互连网络将针对其互连拓扑的特点，直接支持广播和选播通信方式。

综合上述考虑，我们最终给出了如图 1 所示的 Homer 的协议层次结构，它由消息传递层(Homer Message，HM)、数据帧传递层(Frame Delivery，FD)和物理层(Physical Layer)组成。

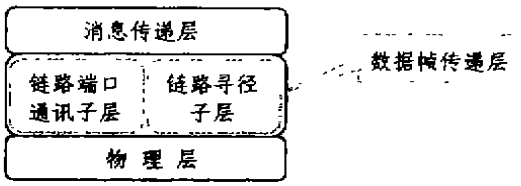

图 1 Homer 的结构

Homer 消息传递层负责上层用户消息数据的封装、端对端的流量控制等任务。数据帧传递层完成 Homer 网络的数据帧传送功能，它可以细分成链路端口通信(Link Port Communication，LPC)子层和链路寻径(Link Routing Communication，LRC)子层。LPC 子层完成相邻物理端口之间的帧传递工作，包括差错控制、流量控制等；而 LRC 子层实现数据帧在 Homer 网络上的寻径、转发等功能。物理层协议则给 FD 层提供字节流在传输媒介上的传送服务，以及相应的物理链路维护功能。在以下的小节中我们将详细介绍各层的特点、结构和功能。

3. 消息传递层

HM 支持多用户同时访问 Homer 网络。它为每个进程分配和管理该进程在通信接口部分的现场(Endpoint)。每个进程通过通信原语库提供的通信原语申请、访问 Endpoint，实现进程间的消息通信。

HM 采用能动消息机制(Active Message)的思想，在发送消息头上携带 Handler(消息处理函数)的标识，当消息到达接收方，由消息头上 Handler-id 域指明的 Handler 函数来对该消息加以分析处理。

由于 Homer 网络本身将承担消息数据的可靠、按序传送，所以 HM 不必负责对消息传送进行差错控制，以及消息数据重组排序工作。

3.1 用户 API

HM 提供给用户进程的通信原语包括 Endpoint 连接原语和消息发送、接收原语。

Endpoint 连接原语包括 Attach(req_endp_num) 和 DeAttach(endp_key)，Attach 实现用户进程与指定 Endpoint 端口的连接，DeAttach 用于删除用户进程与端口的连接。其中 req_endp_num 为用户将访问的 Endpoint 端口号；endp_key 为用户执行 Attach 成功后的返回值。用户在访问 Endpoint 之前，必须调用 Attach；当不再使用该 Endpoint 后应当调用 DeAttach 释放它。

在 HM 中，消息发送、接收原语对长、短消息发送、接收分别处理。对于长消息的发送和接收请求，采用"零拷贝"的 DMA 方式，实现用户缓冲区与网络的数据交换。即不在系统核心设置额外的消息缓冲区。由互连接口板以 DMA 方式实现互连接口板上的数据缓冲区和用户数据缓冲区的数据交换。对于短消息的发送和接收请求，则要求用户以程序控制的 I/O 方式直接访问连互接口电路的数据缓冲区来完成。这样做可以避免设置复杂的 DMA 操作所带来的系统开销，减少短消息的传递时延。

HM 的发送原语包括 Send_short_msg 和 Send_long_msg，即发送一条短消息和发送一条长消息，其具体格式如下：

```
Send_short_msg(Endp_key, SendMode, MsgTag, Destination, MsgHandle, MsgAttr, MsgLen, SrcDataAddr);
Send_long_msg(Endp_key, SendMode, MsgTag, Destination, MsgHandle, MsgAttr, MsgLen, SrcDataChan);
```

HM 接收原语为 Extract(Endp_key, Signalmode)。用户接收线程调用 Extract 原语，等待消息的到达；当消息到达后，转而执行相应的 Handler 函数。接收函数共有 3 个：Rev_Direct_msg(Endp_key, DesBuff)、Rev_Int_msg(Endp_key, DesChain) 和 Rev_Map_msg(Endp_key, DesChain)，它们分别对应着 3 种不同的接受方式，即直接接受方式(Direct_Rev)、中断接受方式(Int_Rev)和映射接收方式(Map_Rev)。

当接收到的消息为短消息时，可采用 Direc_Rev 方式，使用户进程直接从接口硬件缓冲区读取数据；当接收到的消息为长消息并且在接收前发送方已知道接收方用户数据缓冲区的地址，可采用 Map_Rev；当接收到长消息，并且无法在接收前确定用户数据缓冲区地址，采用 Int_Rev。灵活使用这三种接收方式，可在不同通信模式下取得较高的通信效率。

3.2 消息格式

Homer 消息由消息头和消息数据体组成。消息头格式及定义如下：

域名称	长度(Bytes)
源地址(Msg_Src)	2
目的地址(Msg_Des)	2
消息端口号(Msg_Endp_id)	1
消息长度(Msg_Len)	4
消息接收标记(Msg_Tag)	1
消息属性(Msg_Attr)	1
消息处理函数(Msg_Handler)	1

源地址：消息发送节点机物理编号；

目的地址：目的节点机位图编码；

消息端口号：接收方 Endpoint 号；

消息接收标记：作为使用 Map-Rev 方式下，指示接收方用户缓冲区地址描述；

消息属性：接收模式定义；

消息处理函数：目的端消息处理函数。

3.3 连接的管理和控制

HM 为每个 Endpoint 维护一个与 Homer 网络的连接(Elink)，并为每个 Elink 进行流量控制。流控制分为基于 Credit 的和基于 Rate 的流控制方式。Rate 方式适合于传送数据流量比较平稳的情况。在这种情况下，可以准确计算出当前网络数据传送速率，正确调整发送方的窗口尺寸。但是在阵发性通信发生时，很难迅速计算出网络当前的数据流速。这种做法可以很好地解决暴发性的数据通信冲突，并保证数据帧在接收方不会被丢弃。但 Credit 方式比起 Rate 的方式，要多占用一些缓冲区资源。考虑到 SNOW 系统的通信特点，我们采用基于 Credit 的流控制方式。

4. 数据帧传递层

数据帧传递层完成 Homer 网络的数据帧传送功能，它可以细分成链路端口通信子层(LPC)和链路寻径子层(LRC)。LPC 子层完成相邻物理端口之间的帧传递工作，包括差错控制、流量控制等；而 LRC 子层实现数据帧在 Homer 网络上的寻径、转发等功能。

4.1 链路端口通信子层(LPC)

LPC 层协议可分为非流水式的停-等协议(stop and wait)、流水线式的退后 N 帧协议(go back N)和选择性重传协议(selective repeat)。停-等协议如果等待的时间过长，就会严重影响物理信道带宽的利用率；退后 N 帧协议使发送方不必等待接收方的应答，但在物理信道可靠性较差的情况下，大量的数据帧重传会降低信道利用率；在选择性重传协议中，由于接收方返回应答帧，指定发送方重传的数据帧，减少了重发的数据

帧数目。三种协议的性能依次提高，而协议复杂程度也相应增加。

通过理论计算和对试验结果的分析，最终确定采用 GO_Back_N 的滑动窗口协议[4]。下面是帧格式的说明及接收、发送方的链路控制行为描述：

· 数据帧格式定义

内　容	长度(bytes)
Start of Frame	2
Frame Type+SeqNo	1
FrameHead	4
Data Field	128
Crc	2
End of Frame	1

Start of Frame：帧起始同步字符；

Frame Type(3 bit)：帧类型指示(数据帧、应答帧、链路流控制帧、消息流控制帧)；

SeqNo(3 bit)：帧滑动窗口序列号；

FrameHead：数据帧头；

End of Frame：帧尾字符；

Crc 校验和校验多项式为 $G(x)=x^{16}+x^{15}+x^2+1$

FrameHead 定义为：

7　　0

D_ID(信宿地址)	
D_Endpoint(信宿端口号)	
S_ID(信源地址)	
Reserved	LastFrame

LastFrame：为 00 时，表明此数据帧为消息的第一帧；为 11 时，表明此数据帧为消息的最后一帧。

· 消息流控制帧格式定义

内　容	长度(bytes)
Start of Frame	2
FrameType+SeqNo	1
FrameHead	4
Credit Field	1
Crc	2
End of Frame	1

其中 Credit Field 域指明流控制帧携带的 Credit 的数量。

· 应答帧和链路流控制帧格式定义

内　容	长度(bytes)
Start of Frame	2
Control Field+Frame Type	1
End of Frame	1

其中 Frame Type 为应答帧和链路流控制帧；Control Field 的定义为：应答帧号 AckFrmNo(3 bits)+ConnFlag(1 bits)。

· LPC 协议发送、接收方控制流程

```
Sender:
    Conn-state:(断连、连接);
    FrameUpp:integer0..7;
    Begin
        Event:发送请求;[Effect:
                        if Conn_state=连接 & i<
                                    FrameUpp
                            Send (Frame(i)) * 发送
                                    第 i 帧
                            i=(i+1)mod 8
                        end if]
        Event:Receive Ack Frame;[Effect:调整
                                    FrameUpp]
        Event:Receive Flow Control Frame;[Effect:修
                                    改 Conn_state]
        Event:超时;[Effect:重发当前未获应答的数据帧]
    End
Receiver:
    Begin
        Event:帧到达;[Effect:Rev(Frame(i))]
        Event:正确接收;[Effect:发送 Ack Frame]
        Event:缓冲区状况发生变化;[Effect:发送 Flow
                                    Control Frame]
    End
```

4.2　链路控制通信子层(LRC)

LRC 层协议由路由、广播(或选播)、拥塞控制等部分组成。

4.2.1　路由。Homer 互连网是由多个 HUB 级联组成的，每个 HUB 是由16个端口的全互连的交叉开关构成的，其中每个端口连接一个接口板(SMIF)。帧路由就是源端 SMIF 根据帧头路由信息选择正确的端口以联接目的端 SMIF。

每个 SMIF 都维护一张由硬件支持的路由表，表的入口参数为目的节点号，出口参数为本 HUB 的目的端口号。

4.2.2　广播(或选播)。Homer 协议采用位图方式来描述广播帧地址，即位图中的每一位对应于系统中的一个节点。因此全广播与选播没有本质区别。帧广播包括广播帧路由、广播帧发送两个部分。

·广播帧路由。是指当源端 SMIF 接收到一个广播帧后，要根据路由信息计算出必须向本 HUB 上哪些端口(即 SMIF)发送该帧的过程。该计算过程和所需数据结构描述为，在 HUB 上的每一个 SMIF 都应建立如下的数据结构：

NN：系统规模，即可连接的最大的节点数目；

PN：一个 HUB 的端口数目；

NIND[1..NN](节点指示字)：位图，每一位对应于一个节点。如果某位为1，表明消息应发送至对应节点；

NMSK[1..PN](节点屏蔽字)：位图，每一位对应于 HUB 上的一个端口。如果某位为1，表明对应的端口连接了一个节点；

PIND[1..PN](端口指示字):位图,每一位对应于HUB上的一个端口。当PIND被计算完毕后,若某位为1,表明需向对应的端口发送消息;

PRMSK[1..PN][1..NN](端口路由屏蔽字数组):二维位图数组,每一个位图对应于一个端口,其中每一位对应于一个节点。若位图中某位为1,则表明其对应的节点在该位图对应的端口的路由控制下。用位图与NIND作AND运算,如果其结果不为0,表明该位图对应的端口在消息广播时应被选中,并建立PIND对应位。

当一个SMIF的输入端口接收到一个广播消息后,其路由选径工作主要由下述的循环完成,选径结果体现在PIND中。

```
For(i=1;i≤PN;i++){
        If(NIND&PRMSK[i]≠0)
                PIND[i]=1;}
```

当计算结束后,SMIF将PIND提交给HUB的集中控制器作为链路建立的参数。

·广播帧发送。是指源端SMIF通过交叉开关向计算出的所有需连接的目的端SMIF发送该帧的过程。从提高整个互连网效率角度出发,广播帧的发送必须解决下述问题,即:在建立广播帧链路时,由于部分目的端口正在与其他端口(或共享存储器)进行通信,因此只有部分目的端口处于可以接收数据的状态,而并非所有目的端口都可以接收数据。为了解决这个问题,互连网允许广播帧可以先发送给已就绪的目的端口,然后再发送给其他目的端口。

广播帧的发送是由集中控制器和源端SMIF配合完成的,集中控制器连续查询目的端SMIF的状态,如果允许建立链路则建立链路并通知源端启动帧传输。集中控制器重复该过程直至所有的目的端SMIF都接收到了该帧,源端SMIF接受集中控制器的控制,重复发送该帧直至集中控制器通知其发送完毕。

·拥塞控制。Homer协议中在解决拥塞问题时是采用帧缓冲和阻塞相结合的方式来抑制拥塞的。

4.3 数据帧传递层与消息传递层的连接

·FD为HM提供的服务。主要包括消息数据发送和接收、消息Credit的发送和接收。消息数据发送和接收函数为:Send_Msg(DAddr,Endpoint,SAddr,MsgData)和Rev_Msg(Endpoint,SAddr,MsgData);消息Credit的发送和接收函数为:Send_Credit(DAddr,Endpoint,SAddr,MsgData)和Rev_Credit(Endpoint,SAddr,MsgData),其中DAddr为消息目的地址,Endpoint为目的端口号,SAddr为消息发送方地址,MsgData为消息数据;HM发送选播和广播消息通过对DAddr的设置来实现。

·消息数据的分片和重组。由于Homer网络提供永久虚连接服务,保证数据帧的可靠、顺序传送,所以整个消息数据只是简单地按长度在发送方分割成定长的数据帧,而不必在数据帧上标记它在消息当中的位置,消息数据的头帧和尾帧由帧的LastFrame域规定。接收方根据该域区分消息的开始和结束,并根据消息长度域抛弃最后数据帧的多余数据。

5. 物理层

物理层可再分成物理链路控制子层(PHYL)和物理链路信号层(PHYS),PHYS完全符合Fibre Channel协议要求。使用IBM Type-1类屏蔽双绞线,这种类型的双绞线为八芯,其中两对分别作为双向数据通道。电缆接头为DB-9插头。信号编码方案为8B/10B。数据传输速率为250Mbps,传输距离为20米范围。信道比特出错概率低于10^{-15}。物理层提供给端口控制层的界面为八位数据线、两位控制数据线和控制命令线。

PHYL层用于监视相邻端口是否存在,接收到的比特流是否失去同步,并把相应的情况通知给链路层;接收链路层的命令,发出命令字符序列,或者关闭物理信道,自动从网络上脱离。

结束语 精简高效的互连协议是保证机群系统网络满足高带宽、低时延要求的关键之一。本文在分析机群系统网络特性的基础上,详尽给出了SNOW系统互连协议各个层次的结构、功能、特点和层间的相互关系。目前Homer协议已用于SNOW机群系统的互连网络系统中。实践表明该协议结构完备合理,完全满足SNOW系统的需要。

参考文献

1 北京航空航天大学计算机系.具有分布式共享存储机制的可扩展机群系统.[项目研究进展报告(NO 863-306-ZD-03)].1997

2 Clark D D, et al. An analysis of TCP processing overhead. IEEE Communication Magazine, June 1989

3 Eicken T, Culler D E. Active Messages: a Mechanism for Integrated Communication and Computation. In: Proc. of the 19th International Symposium on Computer Architecture. May 1992

4 吴文峻,向晓华,龙翔.高速机群互连网络链路层协议设计.北京航空航天大学学报,1998(4):458~461