直接互连结构中支持优先级业务的自适应路由算法

朱旭东 李乐民 许 都
(电子科技大学宽带光纤传输与通信网络重点实验室 成都 610054)

摘 要：直接互连结构(Direct Interconnection Network，DIN)具有较好的分布式特性逐渐作为可扩展数据交换结构的核心。在数据交换应用中支持服务质量(Quality of Service, QoS)是一个重要的指标。为此，该文提出了在 DIN 结构中支持公平带宽分配和支持优先级业务的路由算法。考虑到在这类结构中路由机制和交换单元的调度策略之间存在紧密联系，该文结合路由机制和调度策略，提出了一种直接互连结构中支持优先级业务的自适应路由(Priority Supporting Adaptive Routing，PSAR)算法。该路由算法可公平分配输出带宽给各个输入端口，同时支持优先级业务。仿真实验验证了公平分配输出带宽和对优先级业务的支持。

关键词：易扩展交换结构，直接互连结构，路由算法，优先级业务
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2005)03-0337-04

A Priority Supporting Adaptive Routing Algorithm in Direct Interconnection Networks

Zhu Xu-dong Li Le-min Xu Du
(*Key Lab of Broadband Optical Transmission and Communication Networks, UEST of China, Chengdu* 610054, *China*)

Abstract Direct Interconnection Networks (DIN) are considered to build scalable switching fabrics for Internet routers/switches, due to its easy scalability. Furthermore, QoS (Quality of Service) guarantee is very important in switching systems. In this paper, fair bandwidth allocation and priority traffic supporting adaptive algorithm in the DIN is presented. In the switching fabric, there are tight relationships between the schedule scheme and the routing strategy. In order to supporting QoS in the DIN, a new Priority Supporting Adaptive Routing Algorithm (PSAR) is presented, which considers the scheduling strategy and routing scheme at the same time. Simulation results show the algorithm can fairly allocate the output port bandwidth to each input port, and support priority traffic.

Key words Scalable switching fabric, Direct interconnection network, Routing algorithm, Priority traffic

1 引言

随着 Internet 用户量和各种新的网络业务(如电子商务、网络视频点播等)的增长，对网络带宽的需求和对服务质量(Quality of Service, QoS)要求都越来越高。由于光纤技术的采用实现了宽带点到点链接，由此整个网络的瓶颈由传输链路转移到网络交换机和路由器上。多维直接互连结构作为一种扩展性很好的交换结构逐渐被考虑应用到高性能可扩展交换/路由器中[1, 2]。

在传统的交换结构，如单级 Crossbar，三级 Clos 结构中各个端口在物理位置上时均匀的，也就是每个端口在接收，或转发业务时与端口的位置无关。例如，在不考虑其他业务流的影响时，从 1 号输入端口转发分组到 2 号输出端口，与从 3 号输入端口转发一个分组到 2 号输出端口，处理过程是相等效的，所以其性能(例如，时延，等待时间等)也是相同的。然后在直接互连结构(Direct Interconnection Network，DIN)由于结构是非均匀等价的，所以同样是上面的例子，在 DIN 中就会出现不同的情形，因为 DIN 中不同位置的节点，发送业务到同一个输出端口，其所获得的带宽是不相同的，距离远的节点，由于中间经过的转发跳数较多，在竞争同一个输出端口带宽时吃亏，这是由 DIN 本身的结构特性决定的。

在现有文献中采用的支持优先级业务的策略仅考虑交换节点上的调度[3, 4]，通过在交换节点的调度策略中加入对优先级的支持来实现公平带宽分配，即已经完成分组发送的端口，在下一轮发送时其发送的优先级降低，降低的量正比于其上一轮发送的分组长度。在这类策略中每发送完一个分组需记录已发送分组的长度，同时更新各个输入端口的优先

2003-11-18 收到，2004-04-26 改回
国家自然科学基金 (60372011)资助课题

级值。此策略由于需要进行记录和更新操作，增加硬件实现难度。但在实际结构中路由机制同样会影响结构的公平性特性。DIN 中的路由机制是指根据当前节点和目标节点的地址来获取分组传送的下一跳节点，也就是说，路由机制的作用是决定分组的传送路径，而调度策略则负责解决分组传送冲突。本文中我们把路由机制和调度策略的结合统称为路由算法。当拓扑结构固定时，路由机制和调度策略决定整个结构性能的关键因素。同时这两者间又存在紧密联系。首先，不同路由机制的选择会影响调度策略的选择，关于这点具体分析见第 2 节。其次，对于仅在节点调度策略中支持公平带宽分配的算法，如果分组传送请求被调度器拒绝，就只有在分组所在节点缓存器中等待或者被丢弃。但在结合动态路由机制时就可使该分组选择其他空闲通道，从而避开拥塞通道，达到平衡负载的目的。因此本文的目的在于结合这两者来实现一种具有公平带宽分配特性的路由算法，同时考虑了对优先级业务的支持。与动态路由机制相结合，提出了一种支持优先级业务的自适应路由(Priority Supporting Adaptive Routing, PSAR)算法。

本文首先讨论路由策略和调度机制两者之间存在的关联，随后对 PSAR 算法进行详细描述。接着，仿真分析算法的带宽公平分配特性以及对优先级业务的支持特性。最后是对全文的总结。

2 路由策略和调度机制

本节首先介绍我们采用的结构模型，随后分析路由机制与调度策略间的关联。直接互连结构是多个节点通过直接互连的方式构成一个 n 维拓扑结构，每个交换节点既是业务源、目标节点，同时也负责业务转发，即每个交换节点连接输入、输出端口，同时也负责其他输入、输出端口间的互连。一个 DIN 结构可用$(k_1, k_2, \cdots, k_i, \cdots, k_n)$来表示，其中 k_i 表示每一维上的节点数。在我们采用的结构模型中，采用了虚拟通道[5]的策略，具体实现时每个物理链路在与它相连的交换节点的输入缓存器，该缓存器划分为几个虚拟队列，每个队列对应物理链路上的一个虚拟通道。分组传送采用类似与虫孔路由[6]的策略，即每个变长包进入交换结构前切分成等长的分组，只在第一个分组（头分组）中包含路由信息和控制信息，后续分组沿着头分组经过的路径传送。不同于虫孔路由之处在于允许多个分组在交换节点的缓存单元中缓存。

在 DIN 的每个节点上，采用路由策略和调度机制相结合来实现分组传送，我们亦称为两级实现。第一级是路由机制，它负责虚拟通道的分配，即根据路由机制为每个输入队列分配下一跳虚拟通道；第二级是调度策略，负责输出物理端口的调度，为所有已经获得虚拟通道分配的输入队列分配物理端口。图 1 是一个路由机制和调度策略流程示意图。我们用 IV(i,j)来表示输入端口 i 上的虚拟队列 j，输入队列集合为 IV$=\{$IV$(i,j)|$ $i=0,1,2,3,\cdots,P-1$, $j=0,1,2,\cdots,N-1\}$，其中 P 为每个节点的物理端口总数，N 为每条物理链路上的虚拟通道总数。OV(l,m)表示输出端口连接的物理链路 l 上的虚拟通道 m，同理，虚拟通道的集合为 OV$=\{$OV$(l,m)|l=0,1,2,3,\cdots,P-1.$ $m=0,1,2,\cdots, N-1\}$。在图 1 中 IV 通道上的 OV(l,m)表示该输入队列存在分组,其输出虚拟通道为 OV(l,m); 同理，在 OV 上，表示该输出虚拟通道分配给输入队列 IV(i,j)上的分组。

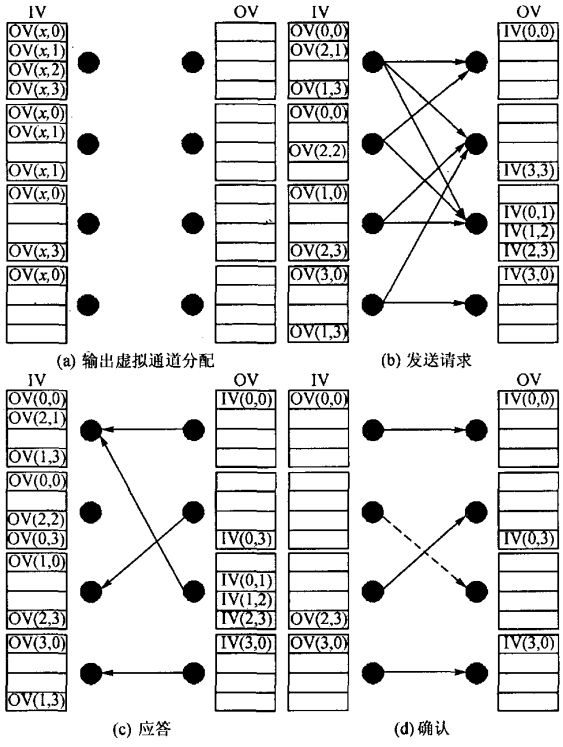

图 1 路由策略和调度机制流程

图 1 给出了整个算法的实现过程，首先在输入队列 IV 中的 OV(l,m)分组向路由单元发送输出虚拟通道分配请求，接着，在第二级中分配完输出虚拟通道的输入队列向输出端口发送传送请求。输出端口在收到请求后随机选择其中一个输入队列返回应答信息，发送端在收到应答后随机选择一个输出端口返回确认信息。下面我们分析选择不同的路由机制对调度策略实现的影响。

在采用确定路由机制时由于分组对下一个通道的选择是确定的，所以路由机制不会影响调度策略的实现。然而，在动态路由机制中，情形将不同。由于采用动态路由策略需要避开冲突的通道，所以在第一级的输出虚拟通道分配应考虑下面两个因素：(1) 所请求的虚拟通道是否空闲；(2) 如果空闲那么该虚拟通道所在的物理链路是否与别的虚拟通道的请求存在冲突。其中第(2)个因数的目标是判断输出物理链路是否冲突。这一点是通过调度策略来实现的，因此为实现

负载平衡，路由机制和调度策略须结合在一起考虑。在实现时可采用返回多个位于不同输出物理链路上的虚拟通道，从而调度策略可选择的输出端口范围增加，避开冲突端口的概率增加。针对是否考虑物理链路冲突，可有下面两种不同的策略：一种是不考虑输出物理链路冲突，直接选择虚拟通道，即路由机制任意选择一个空闲的虚拟通道 OV(l, m)返回给 IV(i, j)。这样调度策略的判决最大空间就等于物理链路数 n_{phy}。另一种策略是在虚拟通道分配时选择多个空闲的虚拟通路{空闲的 OV(l, m)}，由第二级调度策略来决定对虚拟通道的选择。此时调度策略的判决最大空间就是 $n_{phy}\times N$。综上可知选择不同的路由机制可影响调度策略。因此，本文结合路由机制和调度策略，考虑实现一种公平分配带宽的路由算法——PSAR，同时利用该算法实现了对优先级业务的支持。

3　PSAR 算法

PSAR 算法允许分组选择更接近目标节点的任意一条物理链路，即采用最短路径。下面是该算法的描述：首先是虚拟通道的设置，为了避免不同输出端口间分组的冲突，我们采用输出端口虚拟缓存，即在每个节点为每个可到达的目标端口配置一个排队队列。根据该设计思想，可计算出每个节点的输入队列总数。图 2 是在(X, Y)二维结构上任取一个节点(x, y)，根据虚拟通道的设置规则，又由于采用最短路径算法，故通过 $X+$ 方向的链路可到达的目标节点的个数为 $(X-x)\times Y$，同理可得其他方向可达目标节点的个数。所以，总计所需的虚拟队列数 C_{IV} 为：

$$C_{IV}=(X-x)\times Y+(Y-y)\times X+xY+yX=2XY \qquad (1)$$

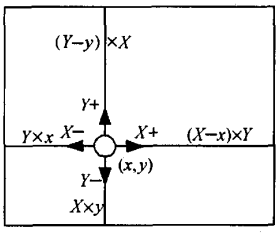

图 2　二维结构中交换节点的虚拟通道配置计算

PSAR 中的路由策略：分组选择的输出虚拟通道包括两个值，输出物理链路 O_{str} 和该物理链路上的通道号 O_{ch}。根据上面提到的虚拟通道设置规则，可用下面的公式来计算目标节点为(x_1, x_2, …, x_n)分组的虚拟通道 O_{ch}：$O_{ch}=\sum_{i=0}^{n}x_i k_i$。在获得 O_{ch} 后，用下面所示的输出物理链路选择策略为每个输入队列计算输出物理链路 O_{str}：

```
while(Ostr)∈输出物理链路集合)
{
    if(OV(Ostr, Och)没有被其它输入队列保留)
    {
        输入队列分配该输出物理链路;
        break;
    }
    Ostr++;
}
```

在 PSAR 算法中第一级分配完后的 OV(l, m)不被输入队列保留(Reserve)，而是允许其他输入队列 IV(i, j)继续分配该通道。因此，在第一级分配完成后可能出现两个或者两个以上的不同输入队列 IV(i, j)分配到相同的输出虚拟通道 OV(l, m)。之所以采用这种策略是为了防止分配是按照一定顺序进行，而造成不公平性，在 PSAR 中该冲突由调度策略来解决。在完成虚拟通道分配后接着进行第二级，下面是调度策略的 3 基本个步骤：

(1) 请求（Request）：完成输出虚拟通道分配的 IV(i, j)向对应的输出端口发送请求,同时包含 IV(i, j)号。

(2) 应答(Grant)：没有匹配的输出端口收到请求，随机选择一个 IV(i, j)，返回应答信息给该输入队列。

(3) 确认(Accept)：没有匹配的输入端口收到多个输出端口返回的应答，随机选择一个返回给输出端口一个确认信息。

在支持优先级业务时，采用独立的虚拟缓存来存放高优先级分组。高优先级输入队列优先分配输出通道，调度器在完成了对高优先级业务的物理端口分配后，才对低一级的业务进行分配。图 1 是整个算法的实现过程，可以看出输出通道 OV(0,0)与 OV(1,3)被分配了两次，该类冲突由调度来解决。另外，IV(1,2)虽然请求的输出端口空闲，仍然没有获得匹配，可采用增加迭代次数来解决，即进行多次请求-应答-确认的过程。

4　仿真结果

为了分析该算法性能，我们使用 OPNET MODELER[7] 建立仿真模型，OPNET 是一个模型化仿真工具，通过节点和链路来构建通信网络和分布式系统仿真模型。我们实现了一个 8×8 Mesh 拓扑，从而构成一个 64 个输入、输出 DIN 交换结构。其中，每个交换节点是一个内部无阻塞的 Crossbar 交换。仿真模型设置每个分组在节点上的处理时间和通道上的传送时间分别为一个单位时隙(cycle)。每个交换节点的输入业务源相同，平均包长度为 20 个分组,发送包的间隔时间为固定常数 20 个时隙。每隔 20 个时隙以概率 λ 发送分组，λ 为业务源的输入负载强度。设定交换节点的每个队列的缓存长度为平均包长度，20 个分组。忽略本地业务，即交换结构的输入端口 In_i 不发送业务到输出端口 Out_i。仿真模型设定分组在到达目标节点后不需要等待立即处理。端口吞吐率是指平稳状态时输入端口平均每个时隙发送的分组数。

4.1 公平分配带宽特性

由于在路由策略中采用确定路由计算不影响调度策略的实现。所以，我们采用一种常见的确定路由机制维序路由(Dimension Order Routing, DOR)来与 PSAR 算法进行比较。在DOR算法中采用与PSAR中类似的调度策略，但由于DOR为确定路由机制，所以它不需要判断通道冲突。仿真业务源为随机均匀业务，即输入端口发送到各个输出端口的概率相同。通过采集每个输入端口吞吐率来比较各个节点的负载平衡特性。图 3 是对几个不同位置的节点负载强度和端口吞吐率特性比较。从图中可以看出在 DOR 路由算法中各个不同位置的节点吞吐率之间变化较大。而在采用了 PSAR 算法后(图 4)，几个端口的吞吐率接近相同。图 5，图 6 是输入负载为 1 时各个端口的吞吐率。图中 z 坐标表示各个端口的吞吐率，x，y 坐标表示各个节点在二维交换结构上的坐标值。图 5 仿真结果显示最大值与最小值间的差值大约为 50%，而在 PSAR 算法(图 6)中最大最小值吞吐率间的差值为 9%，可见 PSAR 算法可达到平衡负载，公平分配带宽的作用，为支持分优先级业务提供了基础。

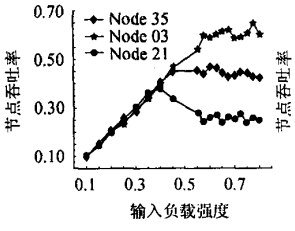

图 3 DOR 路由算法的节点负载-吞吐率

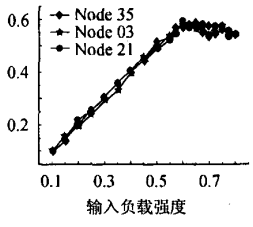

图 4 PSAR 算法的节点负载-吞吐率

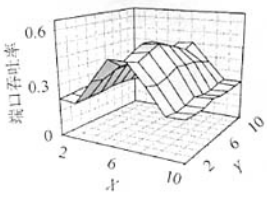

图 5 DOR 路由算法每个输入端口吞吐率

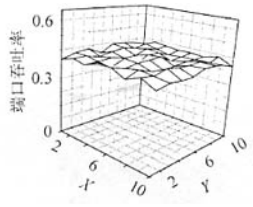

图 6 PSAR 算法每个输入端口吞吐率

4.2 对优先级业务的支持

为了分析 PSAR 对优先级业务的支持，我们在输入端口混合输入两种不同的业务(CBR,VBR)以相同的负载容量混合成一个输入业务，业务流输出端口均匀分布。采用统计每个节点平均每个时隙接收到的分组数作为交换结构的吞吐率。高优先级业务(CBR)和低优先级业务(VBR)分别排队。在输出端统计各个优先级业务的吞吐率。图 7 是仿真结果，分别是高优先级和低优先级业务的负载-吞吐率曲线。图 7 中 HighPri 表示高优先级业务，LowPri 为低优先级业务。结果显示 PSAR 算法可支持优先级业务。当高优先级业务和低优先级业务共存时，能保证高优先级业务的带宽，而低优先级业务的吞吐率下降。

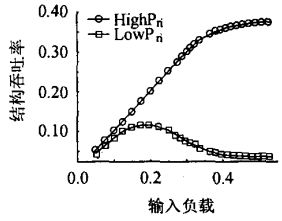

图 7 PSAR 算法时优先级业务的负载-吞吐率曲线

5 结束语

公平带宽分配，支持优先级业务是数据交换结构中的重要指标。在采用直接互连结构构建可扩展数据交换时，需要把路由策略与调度机制结合在一起考虑。本文就此进行了分析，提出了一种路由机制与调度策略结合的支持优先级业务的路由算法，通过结合自适应路由策略实现了各个输入端口公平使用输出端口带宽。这种策略无需记忆或记录最近一次的连接就能达到公平分配带宽的目标。采用在路由策略中对分优先级业务的调度，实现了在直接互连结构对优先级业务支持。仿真结果验证了 PSAR 算法可实现公平带宽分配及支持优先级业务。

参 考 文 献

[1] Park J S, Davis N J. The folded hypercube ATM switch. IEEE International Conference on Networking, Colmar, France, 2001: 370 – 379.

[2] Dally W J. Scalable switching fabrics for Internet routers. 1999, Avici Systems Inc., http://www.avici.com.

[3] Chien A A, Kim J H. Approaches to quality of service in high performance networks. Proc. of the Workshop on Parallel Computer Routing and Communication, Atlanta, Georgia, 1997: 1 – 20.

[4] Kanhere S S, Sethu H, Parekh A B. Fair and efficient packet scheduling using elastic round robin. *IEEE Trans. on Parallel Distributed Syst.*, 2002, 13(3): 324 – 336.

[5] Dally W J. Virtual channel flow control. *IEEE Trans. on Parallel and Distributed Syst.*, 1992, 3(3): 194 – 205.

[6] Ni L M, McKinley P K. A survey of wormhole routing techniques in direct network. *IEEE Computer*, 1993, 26(2): 62 – 76.

[7] OPNET Modeler documentation, OPNET Technologies, Inc., http://www.opnet.com/.

朱旭东： 男，1974 年生，博士，主要从事宽带网络中分组数据交换结构、结构中的调度策略、路由算法等的研究.

李乐民： 男，1932 年生，教授，博士生导师，中国工程院院士，目前主要研究方向为宽带通信网.

许 都 ： 男，1968 年生，副教授，主要从事网络性能分析、核心路由器体系结构中的关键技术研究.