计算机科学2002Vol. 29№. 7

LRSP 路由器链路调度策略及实现方法*)

LRSP Link Arbitration Policy and Its Implementation

安学军　高文学　郑为民　祝明发
(中国科学院计算技术研究所　北京100080)

Abstract The performance of a router chip is crucially influenced by its link arbitration policy. In this paper, the LRSP(Least Recently Served Preference) link arbitration policy and its implementation applied to the Dawning-UX8 routing chips are presented. Practical results show that this policy provides equality of bandwidth for each output channel and the implementation method is robust and reliable with low latencies.

Keywords LRSP, Router, Link arbitration policy

1 引言

从硬件上看,机群系统是由若干独立的高性能计算机(结点机)和起互连作用的高速网络组成的,各个结点间通过传递消息实现结点间的通信。因此,内部互联网络性能的优劣极大地影响着机群系统整体性能的好坏。一种互联网络可通过以下几个要素来表达:拓扑结构、路由选择方式、网络切换机制和流量控制技术。其中拓扑结构决定了网络的连接方式,一般分成直接网和多级网两大类。路由选择方式决定了如何为源结点到目的结点的数据包选取传送路径。路由选择在算法上分成确定性和自适应性两种。在确定性路由选择中,源结点到目的结点间的路径唯一;在自适应性路由选择中,源结点到目的结点间的路径不唯一,网络中的各级路由芯片根据当前的网络状况来选择路径。网络切换机制指消息通过互联网络时,网络内的路由选择芯片如何把消息从输入通道切换到输出通道。网络切换机制一般分为以下几种。包交换、电路交换、信元交换、虚拟切入交换(Virtual cut-through switching)和蛀洞路由(Wormhole-routing)。蛀洞路由是虚拟切入交换路由机制的一种特例,它们之间的差别在于发生阻塞时对消息包的处理不同。流量控制策略是指当数据传送期间发生资源冲突时,对消息如何处理,是丢弃、阻塞、缓冲、还是重新寻径等。

曙光3000超级服务器机群系统的内部互联网络采用多级网络拓扑结构和确定性路由选择算法,网络切换机制采用蛀洞路由,在发生资源冲突时选用阻塞方法实现流量控制。该网络中的核心部件是具有8个双向端口的路由器芯片(Dawning-UX8)。本文着重研究了路由器芯片的链路调度问题,提出了 LRSP(Least Recently Served Preference)优先级轮换调度策略并详细介绍了该策略的硬件实现方法。

2 链路调度概述

路由器链路调度的目的是实现输出通道的动态分配。对于任意一个输出通道,可能同时收到一个或几个输入通道的请求,但在某一时间内该输出通道只能响应其中一个请求。如何合理有效地响应输入通道的请求是链路调度策略需要解决的核心问题。

链路调度的单位有两种,一种是以数据包为单位进行调度,另一种是以数据片为单位进行调度。数据片是把一个数据包分成若干个小的数据片段,同一个数据包的各个数据片沿相同虚通道传送。以数据包为单位调度时,每次传送完一个数据包就检查输入通道的请求并给出仲裁结果。如果输出通道被分配给了某一个输入通道的数据包,在整个数据包传送完成之前不切换给其他输入通道使用。以数据片为单位调度时,每传送完一个数据片就检查一次输入通道的请求并给出新的仲裁结果。

常见的链路调度策略有循环调度(Round Robin)、先到先服务(First Come First Served)和短包优先(Shortest Message First)。

循环(RR)调度策略是目前应用最广泛的链路调度策略。其工作过程是:把输入通道的请求设置成一个循环队列,仲裁器扫描该队列,若第 i 个请求获得响应,为其传送一个数据片或传送完整个数据包。仲裁器再从请求队列的第 i+1个请求开始扫描。很明显,循环调度策略可保持输入通道请求的公平性,不会发生饿死现象,但缺点是请求仲裁时间长。Cray T3E 网络就是采用这种调度策略。

先到先服务(FCFS)调度策略,先到达的服务请求优先获得服务。实现上可在路由器内为每个到达的数据包或数据片设置一个"年龄"计数器,当一个新的数据包或数据节片到达时,置计数器的初值为零。然后"年龄"计数器按照一个可编程的速率增加。仲裁器依照各个数据包或数据节片的"年龄"进行调度,对于"年龄"大的数据包或节片优先分配输出通道。显然,先到先服务的调度策略也不会发生饿死现象。SGI Spider 路由器芯片使用八位"年龄"计数器来实现 FCFS 策略。

短包优先(SMF)调度策略,数据包本身携带包长信息,仲裁器优先为短包分配输出通道。这一策略对不同输入通道不具有公平性,如果某一输入通道所传送的数据包是一组连续的多个短包,输出通道就可能一直被该输入通道占用,而其他输入通道就可能发生饿死现象。因此,在目前的路由芯片中都没有使用 SMF 调度策略实现路由器的链路调度。

*)本课题得到国家"八六三"高技术研究发展计划(863-306-ZD01-01)资助。**安学军**　博士生,主要研究方向为计算机体系结构及计算机网络;**高文学**　硕士,主要研究方向为计算机网络;**郑为民**　博士后,主要研究方向为 SOC 系统芯片设计及计算机网络;**祝明发**　研究员,主要研究方向为计算机体系结构、并行计算及人工智能。

3 LRSP 优先级轮换调度策略

Dawning-UX8路由器芯片采用 LRSP 优先级轮换调度策略实现对输出通道请求的调度，其基本思想是：为输入通道的请求设置一个优先级队列并设定初始优先级顺序，当有数据传送请求时，优先级最高的请求获得响应。数据传送完成后，对优先级队列按照近期最少服务优先(LRSP)原则轮换，即刚完成服务的请求置成最低优先级，原来优先级高于被服务请求的通道其优先级保持不变，原来优先级低于被服务请求的通道其优先级依次提高一个级别。链路调度以数据包为单位，即每次仲裁发生在一个数据包传送结束时刻。

图1 是 LRSP 优先级轮换调度策略的算法描述，P_req[7:0]表示来自八个输入通道的请求，ACP[7:0]是对输入请求的响应，RST 是初始状态，prior[7:0]表示优先级队列，maxprior 表示当前优先级队列中处于优先级最高位置的输入通道请求号，取值范围是0～7。

```
UX_Schedule (ACP[7:0],P_req[7:0],RST)
{
  if (RST)
   {
    ACP[7:0] = 0;
    prior[7:0] = {7,6,5,4,3,2,1,0};
       //赋优先级初值
    maxprior = 7;
    index = 0;
   }
  else
   {
    while ( Req[7:0] == 8'H00)
      {
       ;// 等待有效请求
      }
     maxprior = MAX ( Req[7:0] * Prior[7:0]);
  // 从8个有效请求中挑选优先级最高的一个
    for ( i=0; i<8; i++)
   {
    if (prior[i]<maxprior)
     {
      prior[i]++;
     }
    else if (prior[i]>maxprior)
    {
     ;// 空操作
    }
    else // if (prior[i]=maxprior)
    {
     ACP[i]=1;
     prior[i]= 0;
     index=i;
     }
   }
  while (P_req[index]=1)
  {
   transfer packete();
    // 传送数据包并等待被响应请求撤销
   }
   ACP[index]=0;
   }
```

图1

显然，LRSP 优先级轮换调度策略可保持对输入通道请求的公平性，不会发生饿死现象。

4 LRSP 优先级轮换调度策略的实现

Dawning-UX8路由器芯片在硬件设计上采用并行调度方法实现对8个输出通道的调度，即每个输出通道都具有独立的请求仲裁模块，8个仲裁模块并行工作。每个仲裁模块都由优先级轮换交叉开关、优先级队列、优先级编码和控制信号生成逻辑等模块组成。一个独立的仲裁模块如图2所示。

仲裁模块的输入信号是输入通道发出的请求 P-req[7:0]，输出信号包括两组，一组是对输入通道请求的响应信号 Acp[7:0]，分别是对输入通道7到输入通道0的响应；另一组是对路由器链路的控制信号 Selx[3:0]，在这4位控制信号中低三位控制链路的连接关系，最高位控制链路的连接状态，连通或断开。以下分别介绍仲裁模块中各功能单元的作用及工作过程。

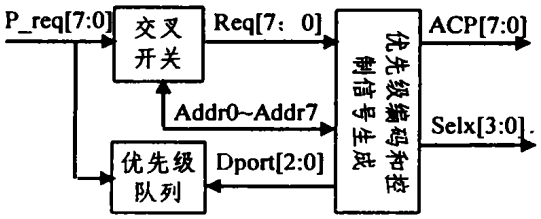

图2 仲裁模块逻辑框图

4.1 优先级轮换交叉开关

优先级轮换交叉开关的功能是实现输入请求信号 P_req[7:0]到输出请求信号 req[7:0]之间的映射关系的变化。因为后面接的是固定优先级编码器，所以当改变映射关系时就相当于改变了输入请求信号的优先级顺序。此交叉开关由8个八选一数据选择器构成。每个数据选择器的输入都相同，8个八选一的输出分别对应 Req[7]、Req[6]、Req[5]、Req[4]、Req[3]、Req[2]、Req[1]、Req[0]，各个数据选择器的选择控制信号分别是优先级控制队列中保存的优先级序号 Addr7[2:0]、Addr6[2:0]、Addr5[2:0]、Addr4[2:0]、Addr3[2:0]、Addr2[2:0]、Addr1[2:0]、Addr0[2:0]。其中 Addr7到 Addr0的取值是7～0且不重复。系统复位后，Addr7＝7，Addr6＝6，Addr5＝5，Addr4＝4，Addr3＝3，Addr2＝2，Addr1＝1，Addr0＝0，此时输出与输入的对应关系是：

Req[7]= P_req[7] 最高优先级，
Req[6]= P_req[6]
Req[5]= P_req[5]
Req[4]= P_req[4]
Req[3]= P_req[3]
Req[2]= P_req[2]
Req[1]= P_req[1]
Req[0]= P_req[0] 最低优先级。

当优先级顺序发生变化时，表现为优先级控制队列中的优先级序号的排列顺序发生变化，从而改变交叉开关输入与输出间的映射关系。

4.2 优先级控制队列

优先级控制队列的作用是保存优先级顺序号，并在一个数据包传送结束时对优先级顺序号按3节中算法进行更新。优先级控制队列的优先级更新在请求的信号 P_req[7:0]中被响应请求的下降沿发生，更新的顺序由控制信号生成逻辑的输出响应序号 Dport[2:0]指明。控制队列的输出(Addr7～Addr0)一方面作为优先级轮换交叉开关的地址信号，以实现优先级轮换；另一方面作为输出控制信号生成逻辑产生响应信号的依据，以确定响应哪个输入通道的请求。

4.3 优先级编码器和控制信号生成逻辑

优先级编码器从收到的所有请求中，选出一个当前优先级最高的请求，并锁定该请求直到请求结束，锁定的要求是正在被响应的通道请求不允许被比它的优先级高的请求所中断。优先级编码器的输入是经过优先级排列后的输入请求信号(Req[0]～ Req[7])，输出是 Ack[0]～ Ack[7](在模块内部)，这八个信号中，同一时刻只能有一个为1，并且为1的持续

时间与对应的 Req 持续时间一致。当编码输出信号有效时 Ack[i]=1，Ack[i]=Req[i]，(i 是0到7之间的数)，而 Req[i] 是对应序号为 Addri[2:0]的输入请求。如 Ack[3]=1，则 Req[3]=1，此时响应第 Addr3号输入请求，若 Addr3=5，说明第5号输入请求(P_req[5])获得响应。仲裁模块应送出 Selx=B1101的链路选择信号，同时使 Acp[5]=1，表示响应第5号输入通道的请求。

结束语 本文所提出的路由器链路调度策略和实现方法已经在 Dawning-UX8路由器芯片上实现，运行结果表明：此调度策略可保证对各输入通道的公平性，无饿死现象发生。这一实现方法工作稳定可靠，延迟时间短，实测的延迟时间为2个时钟周期。

链路调度策略对路由器的性能具有决定性影响，合理地设计调度策略是路由器研究中的一个重要问题。利用虚通道技术提高路由器的吞吐率是目前路由器设计中广泛采用的技术，引入虚通道后，链路调度策略也要做相应的变化，我们下一步将在这方面作进一步改进。

参 考 文 献

1 Pirvu M, Bhuyan L, Ni N. The Impact of link Arbitration on Switch Performance. The 5th International Symposium on HPCA, 1999

2 Herbordt M C, et al. Design Trade-Offs of Low-Cost Network Switches. In: IEEE Proc. of the Symposium of Massively Parallel Processing, 1998

3 曾嵘，董向军，祝明发. 蛛洞路由机置及其芯片设计. 计算机学报，1997，20(5)，404～411

4 Duato J, et al. MMR: A High-performance Multimedia Router-Architecture and Design Trade-Offs. In: IEEE Proc. of the The Fifth Intl. Symposium on HPCA, 1998

5 Dally W J. Virtual Channel Flow Control. IEEE TRANS. On Parallel and Distributed Systems, 1992, 3(2): 194～205

(上接第120页)

表2 问题2的结果

第 I 次运行	在 f_2 的各区间最优解的数目(为简单，[a,b] 代表 10^6[a,b])										nT		ns	
	[2.7,3.5]		[3.5,4.3]		[4.3,5.1]		[5.1,5.9]		[5.9,6.7]					
	MGA	HGA	MGA	HGA	MGA	HGA	MGA	HGA	MGA	HGA	MGA	HGA	MGA	HGA
1	44	34	30	1	26	0	17	0	11	0	128	35	4619	8956
2	45	54	51	0	30	0	40	0	29	0	195	54	4186	9236
3	38	48	30	1	36	0	34	0	27	0	165	49	4023	8956
4	32	35	38	5	23	0	25	0	30	0	148	40	3982	8876
5	26	21	19	3	24	3	16	0	25	0	110	27	3654	8536

结论 本文利用新的约束处理方法和新的适应值函数构造了一个新的约束多目标优化的进化算法，该算法搜索能力更强，能更容易求出一组均匀分布的 Pareto 最优解。计算机模拟的结果也表明了这一点。

参 考 文 献

1 Ishibuchi H, Murata T. A multi-objective genetic local search algorithm and its application to flowshop scheduling. IEEE Transactions on Syst. Man, Cybern. C, 1998, 28 (3): 392～403

2 Leung Y W, Wang Y P. Multiobjective programming using uniform design and genetic algorithm. IEEE Transactions on Syst. Man, Cybern. C, 2000, 30 (3): 293～304

3 Deb K, Pratab A, Meyarivan T. Constrained test problems for multi-objective evolutionary optimization. In: Zitzer E, et al. eds. First Intl. Conf. on Evolutionary Multi-Criterion Optimization, Springer-Verlag. Lecture Notes in Computer Science No. 1993, 2001. 284～298

4 Deb K, Pratab A, Moitra S. Mechanical component design for multiple objectives using elitist non-dominated sorting GA. In: Schoenauer M, et al. eds. Proc. of the Parallel Problem Solving from Nature Ⅵ Conf. 2000. 859～868

5 Srinivas N, Deb K. Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation, 2 (3): 221～248

6 Jimenez F, Verdegay J L. Constrained multiobjective optimization by evolutionary algorithms. In: Proc. of the Intl. ICSC Symposium on Engineering of Intelligent Systems (EIS'98), 1998. 266～271

7 Deb K, Agrawal S, Pratab A, Meyarivan T. Afirst elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-Ⅱ, In: Schoenauer M, et al. eds. Proc. of the Parallel Problem Solving from Nature Ⅵ Conf. 2000. 849～858

8 胡毓达. 实用多目标最优化. 上海：上海科学技术出版社，1990. 35～55

9 Fang K T. Number-theoretic Methods in Statistics. London : Chapman & Hall , 1994. 123～140