高速分组交换网络中调度器的设计

魏艳艳[1],孟李林[2]
(1.西安邮电学院 计算机学院,陕西 西安 710061;
2.西安邮电学院 电子工程学院,陕西 西安 710061)

摘 要:为了满足迅猛发展的网络业务对网络服务质量提出的更高要求,使用高速分组网络交换机中的队列调度器可以有效地提供高质量的网络服务。通过采用分级式队列调度和四种队列调度算法有效地实现了队列调度器的设计。并且深入地比较和分析了队列调度器中多种队列调度算法的优缺点,尤其是对DRR调度算法进行了优化和改进。最后,对所设计的电路进行了仿真验证和电路综合,结果表明该调度器可以满足网络对服务质量的更高要求,并且能够应用到高速分组交换网络的调度器设计中。

关键词:分组交换;队列调度算法;调度器
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2012)01-0025-04

Design of Scheduler in High Speed Packet Switching Networks

WEI Yan-yan[1],MENG Li-lin[2]
(1. School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710061, China;
2. School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710061, China)

Abstract: The rapid development of internet business puts forward higher request for the quality of service in the network, the queue scheduler of the high speed packet switching network can effectively provide the higher quality of service in the network. It uses hierarchical scheduling and four queue scheduling algorithms to realize the design of queue scheduler. What's more, it compares and analyses deeply advantages and disadvantages among kinds of scheduling algorithms in the queue scheduler, especially, improves and optimizes DRR scheduling algorithm. At last, It completes the simulation verification and circuit synthesis for the circuit design, and the results show that the scheduler can satisfy the higher quality of service of the network and can be applied to scheduler design of high speed packet switching network.

Key words: packet switching; queue scheduling algorithm; scheduler

0 引 言

随着互联网业务和下一代网络(NGN)的飞速发展,迫切需要研制高速大容量网络数据包交换芯片。分组交换转发器是包交换芯片中很关键的一部分。使用网络的各类用户可能要求提供不同的网络服务质量(Quality of Service,QoS),诸如低延时、高带宽等。不同的QoS[1~3]要求是由网络交换节点上的分组交换转发器通过流量控制来实现的,而队列调度又是流量控制的重要环节,因此有必要对队列调度算法进行分析与研究。

队列调度算法[4]运行在网络节点中发生冲突需排队等待调度之处,它按照一定的服务规则对交换节点的不同输入数据包进行调度和服务,使所有的输入数据包能够按预定的方式共享交换节点的输出链路带宽。

如图1所示,输入数据包到达交换节点后,分别暂存到相应的队列中,队列调度器的任务是如何从这N个队列中选择下一个要传输的数据包。不同的网络环境中有不同的调度算法,这里只讨论对不同业务流[5]所属的队列的调度。

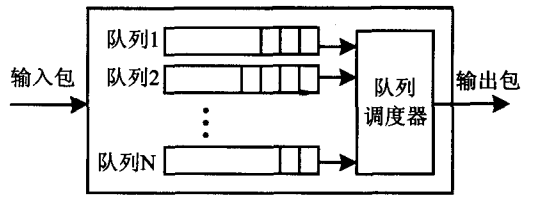

图1 队列调度器功能示意图

收稿日期:2011-06-21;**修回日期**:2011-09-27
基金项目:陕西省"13115"科技创新工程重大科技专项(2009ZDKG-43);陕西省教育科研计划项目(2010JK840)
作者简介:魏艳艳(1986-),女,延安人,硕士研究生,研究方向为高速数据网络包交换技术、数字集成电路系统设计;孟李林,教授,研究方向为专用集成电路设计等。

1 队列调度算法研究

根据不同的服务规则,队列调度算法可以分为以下几种[6]:先到先服务、循环调度、优先级服务、处理机共享、随机服务等。实际上,对于某一特定的调度算法,根据不同的分类标准,又可属于多个不同的类,所以并没有一种统一的分类标准。这里主要对以下几种调度算法进行分析比较,并对DRR算法进行了改进。

1.1 算法的分析比较

SP(Strict Priority)调度算法:每个队列都有优先级,在发送数据分组的时候,高优先级队列中有数据包则首先发送,只有在高优先级队列中的数据包为空时,次优先级的队列才能发送其中的数据包。也就是说一个队列若要发送数据包,那么所有优先级高于它的队列中一定没有数据包,即为空。

RR(Round Robin)调度算法:所有队列采用轮询方式,对有数据包的队列即得到服务,空的队列则跳过。各个数据包的队列没有优先级,遵循哪个队列有数据包,就先发送哪个队列的数据包,多个队列同时有数据包时,则按照默认顺序进行发送,而且每个队列只许发送一个包,就轮到下个队列发送。

WRR(Weighted Round Robin)调度算法[7,8]:其是在RR算法的基础上给每个队列赋予一个权重,同时为每个队列维护一个计数器,初始值为该权重,每次服务一个队列,发送一个数据包,计数器减1,然后服务下一个队列,直到所有队列的计数器为0,则重置所有计数器为各自的权值。

DRR(Deficit Round Robin)调度算法[9,10]:该算法可适合不同封包长度的应用,每个队列设置有一个量子值(Quantum)和一个量子计数器,量子计数器初始化为量子值,量子值对应到包的字节数。每次轮询时,将队列中数据包的长度和计数器值进行比较,如果包长度小于计数器值,则从该队列中发送数据包,同时在计数器值中消耗掉发送数据包的长度,然后重复上述操作,直到数据包的长度大于计数器的值或该队列为空才服务下一个队列。下一次轮询时,量子计数器还需加上该队列的量子值。每个调度轮回中,只有激活队列才能获得服务,若某个激活队列读空但量子值未用完,则将其清0。

上述四种调度算法各有特点,可满足不同QoS的要求。表1对四种调度算法的优缺点进行了比较。

1.2 算法的改进

采用DRR的调度算法,当队列中数据包的长度大于计数器值时,此时调度器会产生一次空调度,从而造成调度流水线的资源浪费。为了提高调度器的效率和简化队列调度器的电路结构,对DRR算法进行如下改进:

该算法机制与WRR算法相类似,每个队列设置有一个权重值(weight),一个量子字节数和一个量子计数器,量子计数器初始化为量子值(Quantum),量子值对应到包的字节数。量子值为权重值与量子字节数的乘积。量子字节数有四种(1k,2k,4k,8k),可以由量子选择信号经选择器选择确定。每次轮询时,都从该队列中发送数据包,同时在计数器值中消耗掉发送数据包的长度,然后重复上述操作,直到计数器值为负数或该队列读空则该队列调度结束,转而服务下一个队列。下一次轮询时,量子计数器还需加上该队列的量子值。每个调度轮回中,只有激活队列才能获得服务,若某个激活队列读空但量子值未用完,则将其清0。采用这种改进的DWRR(Deficit Weighted Round Robin)算法,在电路结构实现上可明显简化电路设计,同时可防止队列的空调度,从而提高调度器的工作效率。

表1 四种调度算法比较

算法	优点	缺点
SP	高优先级队列有较低时延和较高带宽	低优先级的队列不能得到实时的调度
RR	不同队列具有相同级别的调度服务	无法对业务提供时延保证,无法改变变长分组引起的不公平性
WRR	可以改善实时业务的时延	无法提供时延上限保证,无法改变变长分组引起的不公平性
DRR	可改变变长分组带来的不公平性,实现简单	不能很好的满足业务的时延特性

2 队列调度器的设计

针对高速分组交换网络中不同业务流所需求的服务质量,所设计的队列调度器采用分级式的四类调度算法,优先级别[11,12]从高到低分别为:级别3,级别2,级别1,级别0。

分级式调度算法将队列分为多个级别,每个级别包含1个或多个队列,相同级别的队列调度采用同样的调度算法,不同级别的队列调度可以采用不同的调度算法,也可采用相同的调度算法。每个调度级别优先级不同,高优先级的队列优先调度,只有高优先级调度级别中所有的队列为空,才能对低优先级调度级别的队列进行调度。不同优先级的调度级别之间采用SP调度方法。

2.1 结构设计

图2所示为队列调度器的内部结构图,该队列调度器可实现16个队列的调度控制,从内部结构上划分为数据处理、配置ram、队列调度、权值更新等4个子模块。

图 2 中所示,所有队列的输入数据包同时进入到数据处理模块,然后由配置 ram 模块决定各个队列的调度级别和调度算法,从而确定哪些队列可以进行调度(包括队列的分级处理和级别选择)。由权值更新模块产生各个队列屏蔽信号提供给队列调度模块,队列调度模块依据所调度队列的屏蔽信号和相应的配置调度算法(SP、RR、WRR、DWRR 算法)实现对该队列的调度操作,若本次调度有效,则将该队列号和其调度算法反馈到权值更新模块用于更新队列屏蔽信号和 WRR,DRR 的计数器值。下面主要介绍该队列调度器中队列调度模块和权值更新模块的详细电路设计。

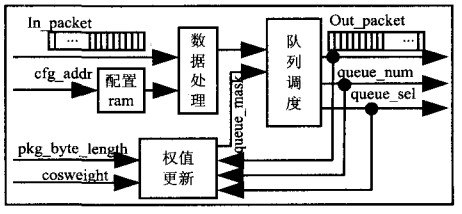

图 2　队列调度器的内部结构图

2.2　队列调度模块设计

队列调度模块完成由数据处理和配置 ram 两个模块确定的可调度队列的调度功能,综合实现了 4 种调度算法。根据队列的调度算法确定下次可调度的队列需要屏蔽与否,再在 SP 算法的基础上来实现该队列的调度。如果是 SP 算法,则不需要屏蔽信号;如果是 RR 算法,或 WRR 算法,或 DWRR 算法,则由权值更新模块产生相应的屏蔽信号,在 SP 算法基础上通过屏蔽某些队列来实现该算法。

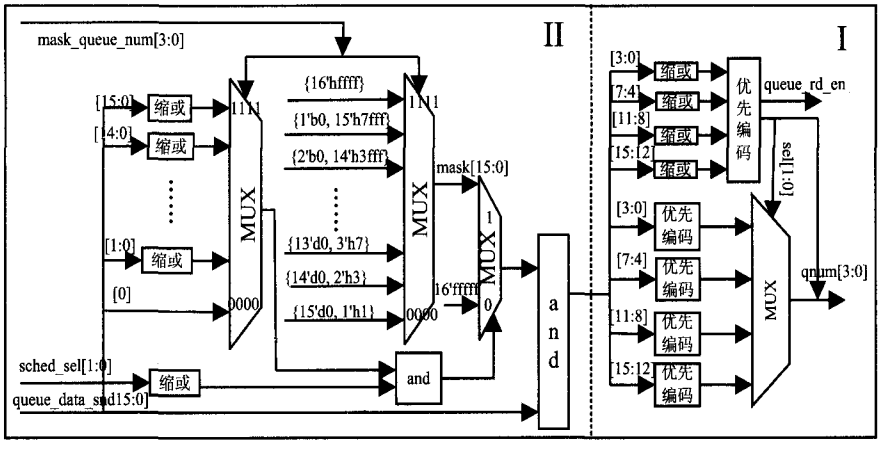

图 3　队列调度模块的电路实现结构

如图 3,第 I 部分是 SP 算法的实现,将 16 个队列分成 4 组,把每组内的四个信号,按位缩或后,得到四个组间信号,再进行组间优先编码,编码结果作为最终调度的高两位,同时也作为四个组的组内优先编码结果的选择信号。同时,每组都进行组内优先编码,经过组间选择后,根据组间编码结果,从四组组内编码的结果选择一组作为最终调度结果的低两位输出。最后,根据当前调度结果,选择屏蔽信号,在进行下次调度时,判断所有队列号小于本次调度的队列号的队列是否有包发送,如果有,则需要将所有队列号大于或等于这次调度的队列号的队列都屏蔽;否则,不需要屏蔽任何队列,从队列 15 开始。

第 II 部分是屏蔽反馈部分,通过屏蔽信号的控制,实现 RR,WRR,DWRR 算法。两个数选器分别用于选择屏蔽信号和判断小于本次调度的队列号的所有队列中是否有包发送。

2.3　权值更新模块设计

图 4 所示为权值更新模块的电路实现结构,该模块完成了 16 个队列的权重值或量子计数器的更新功能。权重值的更新主要完成了 WRR 算法权重值的计算,量子计数器的更新完成了 DWRR 算法量子值的计算,以及各个队列屏蔽点的缓存和选择。由于每个队列只能选择一种调度算法,所以本模块将 WRR 算法的权值计数器和 DWRR 算法的量子计数器合并在一起,采用存储器实现。当采用 DWRR 算法时,使用对应存储单元的全部 22 位;当采用 WRR 算法时,只用其中的低 7 位。

权重更新模块的电路工作原理:每发送一次数据包 ,若该数据包不丢弃,则需要从存储器中读出该队列的权重值(或量子计数器的值),经过权重值或量子计数器值的计算处理后,将结果再存入存储器中。这里的计算处理包含了 WRR 算法和 DWRR 算法的各种计算情况。若该数据包要丢弃,则不需要减去权值(或量子计数器的值)。最后根据计算处理的结果,选择下次调度的队列的屏蔽点。

3　队列调度器的实现

(1)队列调度器采用 Verilog 语言完成电路设计,使用 Synopsys 公司的 VCS 完成电路功能仿真,仿真结果表明所设计的队列调度器针对输入的 16 队列能够完成 4 种调度算法和 4 种调度级别的队列调度功能,完全满足设计要求。

(2)所设计的队列调度器使用 SMIC0.13μm CMOS 标准单元库,采用 Synopsys 公司的 DC 综合器进行逻辑电路优化综合,结果表明该队列调度器工作速度可达 215MHz,满足设计电路 200MHz 工作速度的要求。

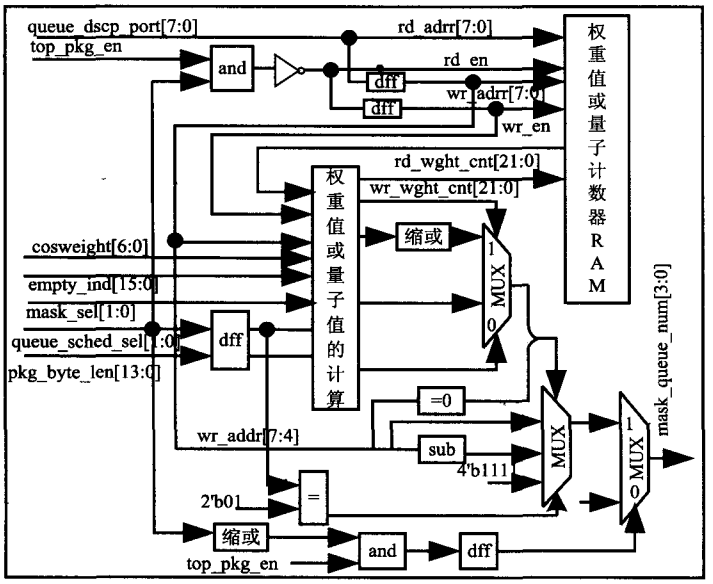

图4 权值更新模块的电路实现结构

4 结束语

文中针对高速分组交换网络中不同业务流所需求的服务质量,优化设计和实现了一个基于高速分组交换网络的队列调度器。该队列调度器采用四种调度算法和四种调度级别可实现对16个队列的调度,且电路实现简单。在实际应用中,各种调度算法和调度优先级别可通过微机接口进行灵活配置。仿真验证结果表明该调度器可以满足不同用户对服务质量的需求,综合结果该调度器工作速度可达到215MHz。

参考文献:

[1] Gurin R, Peris V. Quality-of-service in packet networks basic mechanisms and directions [J]. Computer Networks, 1999, 31(3): 169-179.

[2] 林 闯,李 寅,万剑雄. 计算机网络服务质量优化方法研究综述[J]. 计算机学报, 2011, 34(1): 1-14.

[3] 林 闯,单志广. 计算机网络的服务质量[M]. 北京:清华大学出版社, 2004.

[4] 杨永斌,唐亮贵. 队列调度算法在网络中的应用研究[J]. 计算机科学, 2005, 32(7): 56-58.

[5] 钱光明. 基于业务的多优先级队列区别服务方案[J]. 计算机工程与应用, 2006, 42(10): 118-120.

[6] 王重钢,隆克平,龚向阳,等. 分组交换网络中队列调度算法的研究及其展望[J]. 电子学报, 2001, 29(4): 553-559.

[7] Shimonishi H, Yoshida M. An improvement of weighted round robin cell scheduling in ATM networks [C]//IEEE GLOBECOM'97. [s. l.]: [s. n.], 1997: 1119-1123.

[8] 尹德斌,谢剑英. 一种新的加权公平队列调度算法[J]. 计算机工程, 2008, 34(4): 28-30.

[9] Shreedhar M, Varghese G. Efficient fair queueing using deficit round robin [J]. IEEE/ACM Transaction on Networking, 1996, 4(3): 375-385.

[10] 谢希仁. 计算机网络[M]. 北京:电子工业出版社, 2008: 18-22.

[11] 张登银,许扬扬,蒋 娟. 基于时延的动态优先级调度算法[J]. 计算机技术与发展, 2011, 21(2): 162-165.

[12] 刘化君,刘 斌. 支持多优先级分组交换调度算法研究及其调度器设计[J]. 计算机工程与应用, 2002(14): 92-94.

(上接第24页)

具有层次性。这种仿真过程管理的研究有利于视景仿真系统整体性能的提高,在虚拟视景仿真系统的研究和开发中具有很好的应用价值。

参考文献:

[1] 黄 权,徐学军. 基于OpenGL的卫星跟踪仿真[J]. 计算机技术与发展, 2007, 17(2): 131-134.

[2] 刘 良,黄路炜. 基于OpenGL Performer的视景优化研究[J]. 计算机技术与发展, 2007, 17(8): 77-79.

[3] 张大强,翟素兰,程家兴. OpenGL在视频游戏中的应用[J]. 计算机技术与发展, 2006, 16(2): 73-75.

[4] John S, Carson I I. Introduction to modeling and simulation [C]//Proc of the Winter Simulation Conference. Orlando, Florida: [s. n.], 2005: 16-23.

[5] Burdea G, Coiffet P. Virtual reality technology [J]. Presence: Teleoperators & Virtual Environments, 2003, 12(6): 663-664.

[6] 董 敏. 基于OpenGL的飞行环境虚拟仿真技术研究[D]. 西安:西北工业大学, 2003.

[7] 常 鑫. 三维图形引擎中场景管理的研究与实现[D]. 成都:电子科技大学, 2006.

[8] 陈金水,颜伟琼. 基于OpenGL的三维建模在水利行业中的应用[J]. 计算机技术与发展, 2006, 16(3): 197-199.

[9] Li Ruixian. Flight Environment Virtual Simulation Based on OpenGL[C]//2009 Second International Conference on Information and Computing Science. [s. l.]: [s. n.], 2009.

[10] 潘李亮. 基于LOD的大规模真实感室外场景实时渲染技术的初步研究[D]. 西安:西北工业大学, 2003.

[11] 普建涛,查红彬. 大规模复杂场景的可见性问题研究[J]. 计算机研究与发展, 2005(2): 236-246.