

一种基于多令牌桶的数据风暴抑制单元

马徐瀚, 吴龙胜, 赵坤鹏, 陈庆宇

(西安微电子技术研究所, 陕西 西安 710065)

摘 要: 为了避免网络中突发数据导致的交换机拥塞, 提出了一种面向交换机数据风暴的预防机制. 首先该机制以自适应的速率增加令牌桶内的令牌数, 同时根据用户软件配置监控端口数据帧的类型, 并根据其类型计算转发该数据帧所需的令牌数, 最后对桶内的令牌数和数据帧转发所需的令牌数进行比较. 当桶内的令牌数小于转发所需的令牌数时, 启动交换机的流量控制; 否则实现正常的的数据转发. 在实际工程应用中, 通过多令牌桶的相互协同工作实现对以太网中多种数据类型的流量管理. 仿真结果表明, 所述机制基于多令牌桶技术可实现单播、组播及广播等多种数据风暴的抑制, 满足项目应用要求.

关键词: 数据风暴抑制; 流量控制; 交换网络拥塞; 多令牌桶技术; 以太网交换机

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)09-0084-05

Data Storm Suppression Unit Based on Multi Token Bucket

MA Xu-han, WU Long-sheng, ZHAO Kun-peng, CHEN Qing-yu

(Xi'an Microelectronics Technology Institute, Xi'an 710065, China)

Abstract: In order to mitigate the switch congestion caused by bust data in the network, a preventive mechanism targeting for the data storm in switch is proposed. Firstly, it promotes the number of tokens in a token bucket with an adaptive rate, and simultaneously monitors the type of data frames from a port based on user software configuration. Then, the number of token required by forwarding that specific type data frame is calculated. Finally, the number of token in the token bucket and required by forwarding data frame is compared. Further, if the token amount in bucket is more than that required by data frame forwarding, the proposed mechanism permits data frame forwarding. Otherwise, the flow control is activated. In the realistic application, the collaborative work of multi-token is used to implement the flow control for multiple kinds of data frame. Simulation results show that the proposed mechanism, based on multi-token technique, can successfully suppress the data storm inducing by unicast, multicast and broadcast. All that can completely satisfy the project's requirements.

Key words: data storm suppression; flow control; switch congestion; multi-token technique; Ethernet switch

1 引言

在互联网中, 交换机作为网络中数据交换的核心部件, 面临海量数据转发的巨大挑战, 同时由于网络中的数据具有突发性和间歇性, 可能导致某些重要的服务得不到相应的网络资源, 难以提供可靠的QoS(Quality of Service, 服务质量^[1-3]), 为解决上述问题, 需要采用数据风暴抑制技术实现对流量的管理控制, 保证交换机的高效工作^[4]. 常用的流量管理控制方法有漏桶算法^[5]和令牌桶算法^[6].

在文献[5]所述的漏桶算法中, 漏桶相当于一个缓冲区, 把将要通过漏桶的数据缓存在该处并按规定速率向网络转发数据. 为保证漏桶具有一定的数据风暴抑止能力, 漏桶的转发速率一般会远低于网络最大传输速率, 在网络传输畅通时, 这种算法不能有效利用网络传输资源, 造成一定的资源浪费; 而当流量突发过大时, 会造成数据从漏桶中溢出而丢失, 因此漏桶算法不能支持突发流量.

针对以上两个问题, 文献[6]提出了令牌桶算

法,该算法解决了漏桶算法不能对突发流量的支持问题,但在以太网实际应用中,该算法不能针对特定几种类型的数据帧进行分别监控,无法满足交换机的实际需要。

本文在研究了上述两种算法及文献[7]所述的以太网 MAC(以太网媒体访问控制器)的背压技术和停止帧技术^[7]的基础上,对令牌桶算法进行了改进:采用多个令牌桶协同工作的方式,使其对特定类型的数据帧进行监控,满足一定条件后采取特定操作,控制其传输速率,预防网络阻塞,并将其成功应用在以太网交换机芯片中。仿真及综合结果表明,本设计具有较高的时钟频率及较小的功耗,满足设计要求。

2 基本原理

2.1 令牌桶的基本工作原理

令牌桶的基本工作原理如下:令牌按一定速率添加到具有一定容量的令牌桶中,每个令牌允许通过一定数量的数据。当一帧报文到达时,把报文数据存入缓存,比较该报文长度和令牌桶内的令牌数。若令牌数足够,则向缓存管理模块发出 PASS 信号,根据报文的长度从令牌桶中删除相应数量的令牌;若令牌桶中的令牌数消耗到不足以转发一帧报文数据,则不再允许报文通过,将该报文数据丢弃,发出指示信号。如果一段时间经过令牌桶的数据流量不够多,桶内令牌数可能会达到令牌桶的最大容量,新添加的令牌将被溢出丢弃。令牌桶的示意图如图 1 所示。

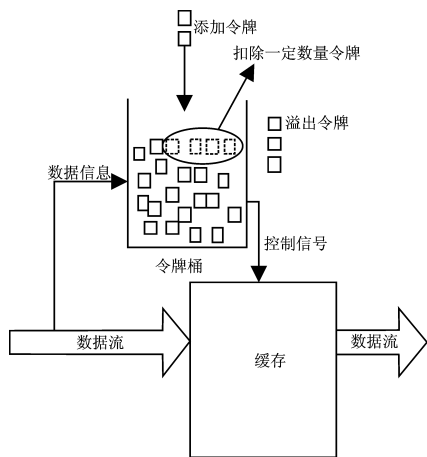


图1 令牌桶的示意图

2.2 设计原理改进

以太网交换机需要同时对多种类型的数据帧进行实时监控,并测算当前网络能否满足其转发。

而多种数据帧的 QoS 要求并不相同,单一的令牌桶只能公平对待所有类型的数据帧,无法有效区分不同的 QoS 要求。为了满足需要,本文提出了基于多令牌桶的数据风暴抑制机制,每个交换机的端口使用三个令牌桶对数据流量进行监控,其流程如图 2 所示。

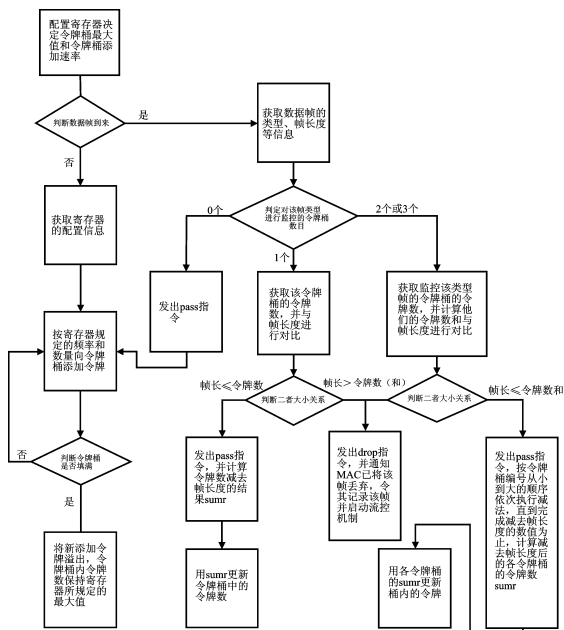


图2 数据风暴抑制单元原理流程图

首先,根据网络的具体情况,通过配置寄存器,指定令牌桶需要监控的数据帧的类型、令牌桶中令牌数的最大值以及向令牌桶添加令牌的速率。

然后,判断该端口是否有数据帧到来,若没有,则以寄存器指定的速率向令牌桶中添加令牌,直到令牌数达到令牌桶的最大值为止。若一帧数据到来,则获取该帧的类型及长度,判断该端口有哪几个令牌桶的指定监控类型与该数据帧类型匹配,按照匹配的令牌桶个数分为以下三种情况,分别进行处理。

(1)如果没有令牌桶的指定监控类型与该帧的类型匹配,则不对该帧进行监控,直接给出 pass 指令。三个令牌桶按寄存器规定的速率添加令牌,直到令牌数达到令牌桶的最大值。

(2)如果只有一个令牌桶匹配,则其他两个令牌桶继续按寄存器规定的速率添加令牌,直到令牌数达到令牌桶的最大值。帧计算模块读取匹配的令牌桶中的令牌数与该帧的帧长度比较大小,若令牌数足够,则向缓存管理模块发出 pass 指令,令该令牌桶的令牌数减去帧长度的值;若令牌数不够,则发出 drop 指令,用 flowctrl 信号通知 MAC(媒体访问控制器)模块已将该帧丢弃,并令其暂停接收新的

帧, MAC 将丢掉的帧记录, 并启动流控机制, 在半双工和全双工模式下分别采用背压算法和发送停止帧的方式停止接收信源发送的数据。

(3)如果有不少于两个令牌桶监控的帧类型匹配, 则计算匹配的令牌桶中令牌数之和, 并与帧长度进行比较, 若令牌数之和不小于帧长度, 则优先从编号较小的令牌桶中减去帧长度, 倘若该桶中令牌数不够, 则依次从编号大一号的令牌桶中继续扣减, 直至完成扣减该帧长度对应的值, 发出 pass 指令; 若令牌数之和不够, 则按照情况 2 中的令牌数不够时的处理方式进行处理。

表 1 令牌桶位速率与寄存器配置关系及可实现数据传输速率

Rate Count(RC)	Bit Rate Mode	Link Speed	Bucket Bit Rate Equation	Bucket Bit Rate Values
1—28	0	Any	$(RC \times 8 \times 1 \text{ MB}) / 125$	64 kB, 128 kB, 192 kB, ..., 1.792 MB
29—127	0	Any	$(RC - 27) \times 1 \text{ MB}$	2 MB, 3 MB, 4 MB, ..., 100 MB
128—240	0	Any	$(RC - 115) \times 1 \text{ MB} \times 8$	104 MB, 112 MB, 120 MB, ..., 1 000 MB
1—125	1	10 Mb/s	$(RC \times 8 \times 1 \text{ MB}) / 100$	0.08 MB, 0.16 MB, 0.24 MB, ..., 10 MB
1—125	1	100 Mb/s	$(RC \times 8 \times 1 \text{ MB}) / 10$	0.8 MB, 1.6 MB, 2.4 MB, ..., 100 MB
1—125	1	1000Mb/s	$RC \times 8 \times 1 \text{ MB}$	8 MB, 16 MB, 24 MB, ..., 1 000 MB

从公式表述可知, 为方便设计, 可采用对时钟分频的方法来简化桶位速率(bucket bit rate)的实现。按表 1 的顺序, 从上至下分别将时钟频率分为: 8 kHz、1 MHz、8 MHz、80 kHz、800 kHz, 这样在不同时钟下, 就能以整数个令牌数向令牌桶添加令牌, 简化了设计实现, 降低了系统功耗。根据所给参数不同, 一个时钟上升沿所添加令牌数及其范围如表 2 所示。

表 2 时钟分频后令牌添加速率关系

RC	Bit rate link mode speed		分频频率/ MHz	令牌数	范围/ MB
1—28	0	any	0.008	8RC	8~224
29—127	0	any	1	RC-27	2~100
128—240	0	any	8	RC-115	13~125
1—125	1	10 Mb/s	0.08	RC	1~125
1—125	1	100 Mb/s	0.8	RC	1~125
1—125	1	1 000 Mb/s	8	RC	1~125

3 数据风暴抑制单元的设计实现

3.1 数据风暴抑制单元在交换芯片中的实现

本设计应用于某型号以太网交换机芯片, 该芯片支持八个端口的数据转发。在本设计中, 采用了多令牌桶协同工作的方式对八个端口的数据风暴抑制

2.3 参数设定

表 1 是向令牌桶添加令牌的速率(bucket bit rate)计算公式, 公式中 Rate Count(RC)是控制令牌桶位速率的重要参量, 其值越大, 令牌桶位速率越大。Bit Rate Mode 是位速率模式, 分为绝对速率(值为 0 时)和相对速率(值为 1 时)两种模式。绝对速率时, 桶位速率与链接速率无关, 相对速率时有关。Link Speed 是交换机端口的链接速率, 分为 10 Mb/s, 100 Mb/s 和 1 000 Mb/s 三种。这三个参数均由用户根据实际应用配置, 经过计算得出 bucket bit rate, 即为一定时间向计数器添加的令牌数。

要求加以实现。根据芯片需求, 对每个端口配置 3 个令牌桶, 共使用了 24 个令牌桶对数据流量进行监控。其在交换机芯片中的结构如图 3 所示, 当从物理层传输来一帧数据, MAC(介质访问控制器)给出该帧的起始定界符, 标志一个新的数据帧到来, 并将该数据帧通过 APB 接口存储于缓存。之后, 帧解析模块获取该帧的端口号、长度、类型等信息。数据风暴抑制单元根据端口号将该帧的长度、类型等信息送至对应端口的单元进行处理, 如果该帧的类型符合所要求监控的帧类型, 这时令牌桶中若有足够的令牌发送数据帧, 则减去相应数量的令牌, 并发出允许通过的信号; 若没有足够的令牌发送数据帧, 则向缓存管理模块发出丢弃信号, 并向 MAC 发出控制信号, 令 MAC 记录丢弃的帧并采取流控机制, 停止接收信源发出的报文数据。

3.2 一个端口的数据风暴抑制单元结构

图 4 是一个端口的数据风暴抑制单元结构图, 主要由寄存器模块、令牌桶模块、计算控制模块组成。用户根据网络实际需要, 通过寄存器配置令牌桶的最大尺寸、单位时间向令牌桶添加的令牌数以及需要令牌桶监控的数据帧类型, 计算控制模块读取令牌桶的令牌计数值, 结合到来的数据帧信息, 给出对该帧的控制信号并更新令牌桶的计数值, 令牌桶

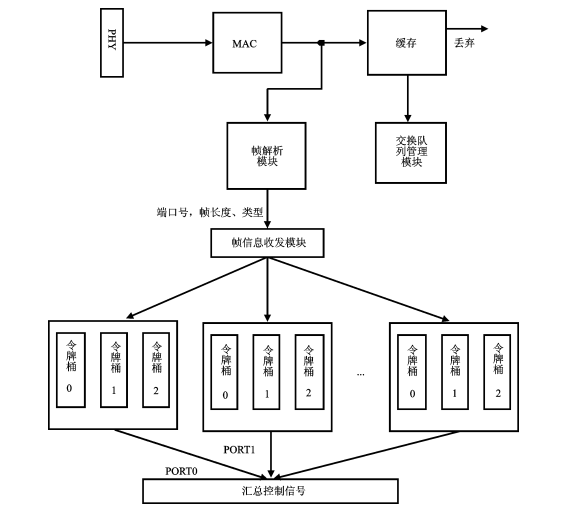


图 3 数据风暴抑制单元在交换机芯片中的结构示意图

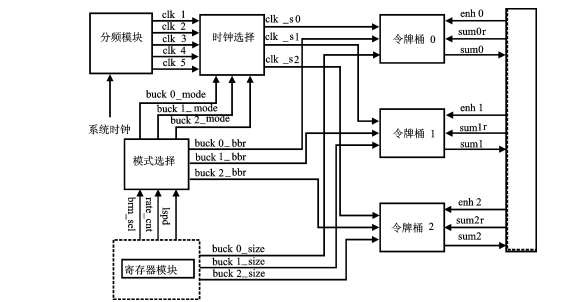


图 5 令牌桶模块结构

更新计数值的操作,将增加的计数值寄存一个系统时钟周期,在下一个系统时钟上升沿进行增加。

3.4 计算控制模块的设计

图 6 是计算控制模块结构图,首先对到来的一帧数据进行解析,获取其端口号、帧长度、帧类型,并产生 sof 信号标志该帧的到来. 通过端口号将该帧的信息送至对应的端口,并产生端口使能信号 port_en,使能该端口的模块工作. 用户通过寄存器的 en_buck_num 信号设定要使用哪几个令牌桶, buck_type 信号设定该令牌桶监控的数据帧类型. 当系统检测到 sof 信号有效,在帧比较模块根据该帧的类型与用户设定的该端口三个令牌桶需要监控的帧类型进行匹配比较,哪个令牌桶的监控帧类型匹配则对应的 buck_en 信号为高电平. 此时,即出现按照数据帧类型与令牌桶设定监控类型匹配情况所分的的三种情况(如 2.2 节所述),最后按照各情况对应的方案进行处理.

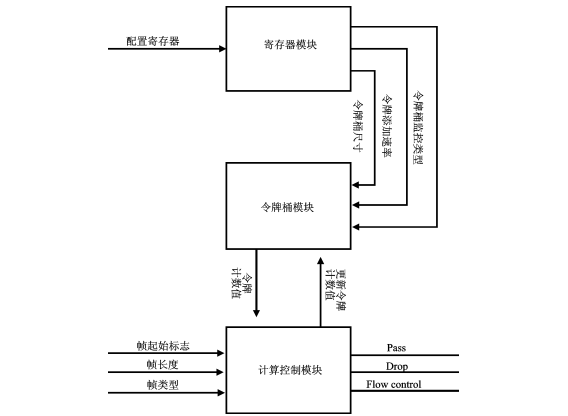


图 4 单端口的数据风暴抑制单元结构

模块对令牌进行计数,将结果送至计算控制模块,接受其对令牌数的更新。

3.3 令牌桶模块的设计

图 5 是令牌桶模块的结构. 为了精确控制添加令牌的时间,对系统时钟进行分频,根据项目实际需求,采用五种时钟频率和与其相对应的令牌数来向令牌桶添加令牌. 三个令牌桶可根据实际需求灵活选择合适的时钟频率、每次添加令牌的数量,以及令牌桶最大值等参数. 令牌桶按照规定的分频频率增加令牌,直到到达令牌桶最大值为止. 期间若有允许数据帧通过的情况,则计算控制模块给出 enh 信号和数据帧通过消耗一定令牌后的令牌数更新值,令牌桶在系统时钟上升沿对计数器当前计数值进行更新。

这时,若分频时钟与系统时钟上升沿重合,则会出现同时要求计数值增加和更新计数值的情况,从而产生矛盾. 因为计算控制模块是根据前一时刻计数器的值来判断数据帧是否允许通过,所以先执行

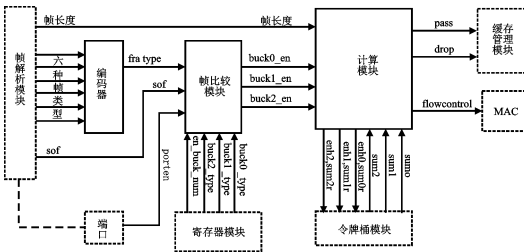


图 6 计算控制模块结构

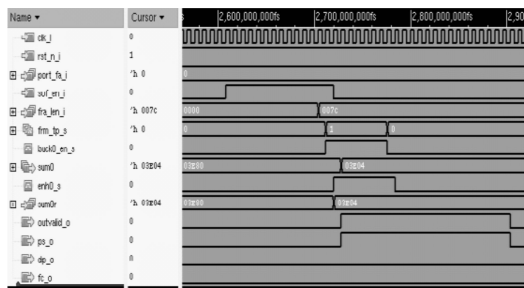


图 7 令牌数满足转发要求时的情况

4 设计验证

本设计采用 VHDL 硬件描述语言设计实现,并搭建了 SOC 验证环境. 通过 NCsim 进行功能仿真,可观察其信号波形,对其功能实现进行验证. 如图 7 所示,从端口 0 接收到一个数据帧,该帧类型与令牌桶 0 监控类型匹配,帧长度为 124 B,长度在允许转发范围内,在输出有效时(outvalid_o 信号为高时)输出 pass 指令,并将令牌数减去帧长度,使令牌数从 16 000 减为15 876.

如图 8 所示,该数据帧同样是从端口 0 接收的,其类型与令牌桶 1 监控类型匹配,帧长度为 124 B,长度超出了令牌桶 1 所允许转发的范围,在输出有效时(outvalid_o 信号为高时)向缓存管理模块输出丢弃指令(dp_o),向 MAC 控制器输出流控信号(fc_o),令牌数不作改变.

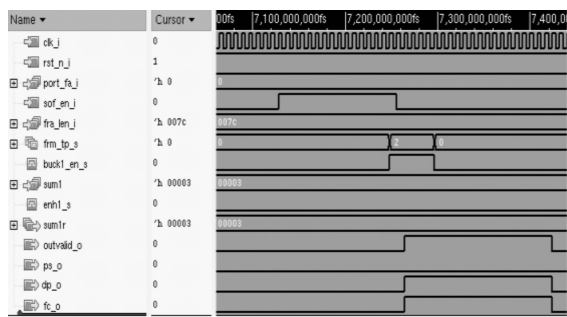


图 8 令牌数不够转发时的情况

在 Altera 公司 STRATIX III 系列的 EP3SL150F1152 芯片进行逻辑综合,得到综合报告如图 9 所示. 综合报告给出了本设计的芯片资源使用情况,工作频率可达 400 MHz 以上,占用面积小. 通过仿真及验证结果表明设计符合预期.

Top view:	switcher_ratectrl_top
Requested Frequency:	401.0 MHz
Total combinational functions	2892
ALUT usage by number of inputs	
7 input functions	22
6 input functions	345
5 input functions	645
4 input functions	487
[=3 input functions	1393
ALUTs by mode	
normal mode	1982
extended LUT mode	22
arithmetic mode	888
shared arithmetic mode	0
Total registers	1647 of 116352 (1%)

图 9 综合报告

5 结束语

本文针对交换网络普遍存在突发数据导致网络拥塞的情况,提出了一种基于令牌桶的数据风暴抑制方法,通过设置,可实现每个端口的三个令牌桶分别对不同的数据帧类型进行监控,也可令三个令牌桶同时对一种帧类型进行监控,每个端口还可以灵活配置各端口的令牌加入速率,进而限制特定类型数据帧的转发速率,以及决定支持突发流量大小的令牌桶尺寸等,从而有效避免由于突发数据流量引发的网络拥塞.

设计中采用 VHDL 硬件描述语言完成了该算法的 VLSI 设计,使用了 NCsim 进行功能仿真验证,使用 synplify 进行综合,结果表明,电路符合设计预期,其工作频率达到 400 MHz 以上,占用面积较小,可以作为数据风暴抑制的独立 IP 使用.

参考文献:

[1] 李鹏,兰巨龙. 一种新型的可重构路由交换通用平台[J]. 计算机应用研究,2009,26(6):2016-2019.

[2] 杨勇. 可重构路由交换平台若干关键技术研究[J]. 信息工程大学学报,2009,10(1):34-40.

[3] Chang IC, Chen SF. An end-to-end QoS adaptation architecture for the integrated intServ and diffServ networks[C]// International Conference Embedded and Ubiquitous Computing. Taipei:IEEE,2007:365-376.

[4] Mingzhi Feng, Hua Zhang, Dong Wang, et al. Design and Implementation of high-speed token bucket based on FPGA[C]// The 2nd International Conference on Information Science and Engineering. Hangzhou, China:IEEE,2010:4474-4476.

[5] Anurag Aeron. Fine tuning of fuzzy token bucket scheme for congestion control in high speed networks[C]// I2010 2ndInternational Conference on Computer Engineering and Applications. Indonesia;Bali,IEEE,2010:170-174.

[6] 李苗,王玉艳,段国东,等. 以太网控制电路流量控制研究与实现[J]. 计算机工程,2007,33(13):268-271.

[7] 冯明智. 可重构路由器流量控制构件的设计与实现[D]. 长沙:国防科学技术,2011.

作者简介:

马徐瀚 男,(1987-),硕士研究生. 研究方向为以太网交换机 SoC 设计、以太网 MAC 研究设计. E-mail: maxuhan0901@163. com.

吴龙胜 男,(1968-),研究员,博士生导师. 研究方向为高速总线、SoC 可靠性研究.