一种多优先级变长调度星载 IP 交换机交换结构的设计

沈泽民,乔庐峰,陈庆华,邵世雷
(解放军理工大学通信工程学院,江苏南京 210007)

摘 要: 针对星载 IP 交换机中硬件资源使用受限的情况,设计实现了一种具有 8 个优先级、采用指针复制和变长分组调度机制的大容量共享存储交换结构,给出了电路的具体组成、关键调度算法和工作流程.使用 Xilinx V4sx55 FPGA 实现了完整的 8×8 交换结构,电路共占用了 164K 字节片上存储器资源和 5982 个 4 输入查找表,可以满足三模冗余设计要求.在系统工作主频为 100MHz、片外采用 SRAM、数据位宽为 64 的情况下,交换结构的峰值吞吐率可以达到 1.6Gbps;片外采用 133MHz DDR 存储器、位宽为 64 时,交换结构的峰值吞吐率可以达到 4.25Gbps;该交换单元进行多级扩展后,可以满足 10Gbps 以上的系统设计需求.

关键词: 星载 IP 交换机; 变长调度; 队列管理
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2014)10-2045-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.10.027

Design of Switch Fabric in Satellite Onboard IP Switch Based on a Multi-Priority Variable-Length Packets Scheduling

SHEN Ze-min, QIAO Lu-feng, CHEN Qing-hua, SHAO Shi-lei
(*Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, Jiangsu* 210007, *China*)

Abstract: Considering the hardware resources limitation in the satellite onboard IP switch, a high capacity shared memory switch fabric with 8 priorities, variable-length packets scheduling mechanisms and pointer copy is presented. The specific structure, the key scheduling algorithm and the working processes of the circuits are provided. Xilinx V4sx55 FPGA are used to realize a 8×8 switch fabric, the circuits occupy 164K bytes on-chip memory resources and 5982 4 input lookup tables, which can meet the requirements of triple modular redundancy. When the working frequency is 100MHz, 64 bit width off-chip SRAM, the peak throughput of the switch fabric can reach 1.6Gbps. Using 64 bit width, 133MHz DDR off-chip memory, the peak throughputs can reach 4.25Gbps. The fabric can be used in multi-stage switch fabrics which can meet the demand of throughputs over 10Gbps.

Key words: satellite onboard IP switch; variable-length scheduling; queue management

1 引言

早期的卫星通信系统中,卫星主要起中继作用,以“弯管”方式进行数据转发,存在频带利用率和功率利用率低、时延较大的问题.随着地面网中 ATM 技术的发展,卫星 ATM 网络和星载 ATM 交换技术被广泛研究[1,2],星载 ATM 交换技术可在硬件规模较小时实现较高的数据转发能力.目前 IP 交换机在地面网中已经广泛应用,在协议体制上可以和地面网无缝链接的星载 IP 交换技术的研究成为热点[3].

目前地面网中 IP 交换机的研究重点是多级交换网络和服务质量保证技术[4~7].而设计星载 IP 交换机的最大困难是微电子器件的限制.太空中剧烈的环境温度变化和空间粒子流都会对器件的正常工作产生严重影响.空间粒子流对元器件最主要的影响之一是单粒子翻转问题,即空间粒子流造成的电路逻辑错误.采用 FPGA 进行系统实现时,解决单粒子翻转最主要的方法之一是三模冗余.三模冗余方法在提高电路抗单粒子效应能力的同时,要求一片 FPGA 的主要资源使用量原则上不能超过总资源量的 1/3.这对交换机的硬件资源消耗提出了苛刻的要求.另外,星载 IP 交换机中可用存储器也受到了严格限制,宇航级 SRAM、SDRAM 和 DDR 的容量和工作时钟频率通常远低于普通商用级器件.以上因素要求星载 IP 交换机必须在满足系统性能要求的同时严格控制逻辑资源的消耗,这需要对交换结构的类型选择、调度算法、处理效率进行全面的优化设计.

本文设计实现了具有 8 个优先级、可以进行变长分组调度的大容量共享存储输出排队交换结构,该结构采用基于指针复制的输出排队(Output Queuing, OQ)方式,采用经过改进的加权轮询(Weighted Round Robin, WRR)

收稿日期:2013-05-24;修回日期:2013-11-11;责任编辑:李勇锋

调度算法和公平轮询(Round Robin,RR)调度算法,在高吞吐率的前提下有效降低了硬件资源消耗.

2 交换结构的构成与工作机制

由于共享存储交换结构具有结构简单、存储资源利用率高、时延低的特性被广泛研究[8],所以本设计采用了基于输出排队的共享存储交换结构.本文中的共享存储交换结构包括输入接口、输出接口、片外存储器、调度器、空闲地址队列管理器、组播计数管理器和队列处理器,如图 1 所示.队列管理器中包含 N(N 为输出端口数)个队列控制器($QC_0 \sim QC_{N-1}$),每个 QC(Queue Controller)内部包括具有 8 个优先级的队列,所有用户数据共享同一个片外存储器.

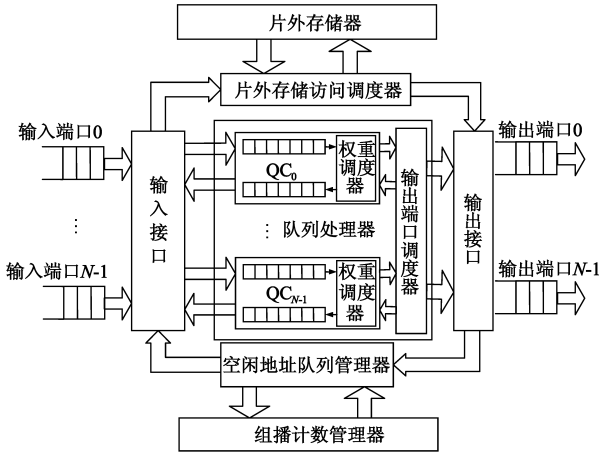

图1 共享存储交换结构

IP 包经过路由查找并获得输出端口信息后进入交换结构,在输入接口处被分割成为一个或多个定长的内部信元(简称信元),输入接口根据其信头中的输出端口信息,从空闲地址队列管理器中取出一个或多个空闲地址指针,这些指针所指向的片外存储区用于存储该分组的所有信元.此后根据信头中的输出端口信息将这些指针链接到不同输出端口对应的逻辑队列中.如果是组播信元,那么其指针会同时进入多个 QC 中并链接到多个逻辑队列中.本设计中采用指针复制的方法实现信元的多播转发,可有效节省片外存储空间,提高组播时交换机抗数据流波动的能力.每个输出队列都有一个变长分组计数器,用于记录在此队列内排队的完整的变长分组个数,只有变长分组计数器不为 0 的输出队列才会获得输出调度机会.

针对每个输出端口,都有 8 个具有不同优先级的队列和一个 WRR 调度器.WRR 调度器根据不同队列的优先级按照预先配置的调度权重选择从哪个队列输出数据.输出端口调度器和片外存储访问调度器用于选择可以输出数据的端口和对片外存储器读写访问的权限.

对于整个交换结构来说,为了减少资源消耗并保证系统性能,需要在交换结构队列模型和调机制上进行系统的优化设计.

3 交换结构队列模型的优化设计

对内部信元采用定长调度方式的共享存储交换结构的队列模型如图 2 所示.由于分组在交换结构内是以定长信元为单位进行排队、调度和管理的,所以在输出端口上需要重组为原变长分组.重组时,在每个输出端口上需要根据内部信元的入端口建立虚拟输入队列(Virtual Input Queuing,VIQ)以保证分组的正确重建.当端口数量较多、存在多个优先级时,这部分电路会消耗一定的资源.

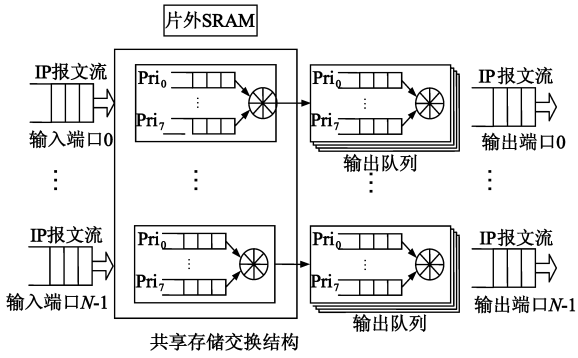

图2 定长调度时的队列模型

本设计中采用的队列模型如图 3 所示,输出端口为 $0 \sim N-1$ 的信元依次存储在 $QC_0 \sim QC_{N-1}$ 中,每个 QC 对应一个输出端口.QC 内部有 8 个逻辑上的先入先出队列(First Input First Output,FIFO),分别对应优先级为 0 ~ 7 的信元.输入的信元被存储在片外存储器中,其地址指针根据优先级以链表的方式链接到对应 QC 中的相应逻辑队列之后.为了提高变长分组交换的效率、降低硬件实现的复杂度,本设计采用了逻辑切分的方法来实现分组的变长调度,即当分组到达输入端口时不对分组采用真正的物理切割,而是将分组逻辑切割成定长信元,分组的所有信元地址指针在 FIFO 内仍是按顺序排列的.当 QC 内的 $FIFO_m$(m 为 0 ~ 7)获得输出权限后,此 QC 的输出通道就一直被 $FIFO_m$ 所拥有,直到这个分组输出完毕才可以为其它优先级队列服务,这样到达输出端口的分组就是完整的,避免了在输出端口处采用 VIQ 进行信元重组.

对比上面两个队列管理模型,采用定长分组交换的队列管理器需要在输出线卡上建立 N 个 VIQ 队列来重组信元,而每个 VIQ 队列中又有 8 个优先级,因此每个输出端口处需要 $8N$ 个队列用于分组重建,对于 N 个端口,共需要建立 $8N^2$ 个重建队列.采用变长分组交

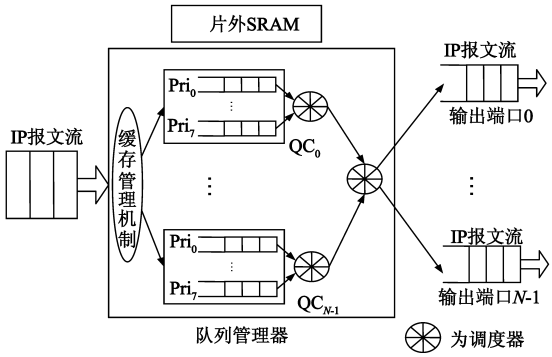

图3　变长调度时的队列管理模型

换的队列管理器可以直接交换变长分组,每个输出线卡上只需要一个输出队列,N 个端口只需要 N 个队列.例如 $N=8$ 时,每个重建队列存储深度平均为 4K 字节,那么两种实现方式需要的存储空间分别为 2.048M 字节和 32K 字节.可见后者可以有效降低输出线卡的设计复杂度和电路资源消耗.

采用变长调度时,由于不同队列中排在队首的帧长度差异可能很大,如何保证不同端口之间调度的公平性、对片外数据缓冲区访问的公平性及保证同一端口内不同队列调度的权重就成为一个需要重点解决的问题,这需要对交换结构中的多个调度器进行整体优化设计.

4　交换结构内部调度器的优化设计

为了在较少资源消耗的情况下解决变长调度带来的调度公平性、带宽分配均匀性问题,本设计在交换结构中采用了三种不同的调度器:输出端口 QC 内部的变长分组权重轮询(Weighted Round Robin,WRR)调度器、基于信元的输出端口公平轮询(Round Robin,RR)调度器和基于信元的片外存储器读写访问的公平轮询调度器.

4.1　QC 内部 WRR 调度器的设计

为了实现对不同优先级队列的有效调度,WRR 算法被广泛研究[9~11].对于共享缓存输出排队的交换结构而言,每个输出端口存在具有不同优先级的多个队列,此时需要设计专用的权重调度器,以保证不同优先级的队列得到不同的输出带宽,从而保证高优先级业务的服务质量.图 4 所示的是某个输出端口内部的队列结构.它包括 8 个具有不同优先级的队列,队列中的每个方格代表一个内部信元,可见,一个变长分组通常包括多个信元.每个优先级队列对应着一个深度计数器(Depth Counter,DC),记录着该队列中完整数据帧的个数.WRR 调度器根据预先的配置,依次轮询每个队列,只有当该队列的 DC 计数器深度不为 0 时,才能够输出一个完整分组所包括的多个信元.图 4 中权重调度器右侧的是依次输出的信元,可以看出调度器每次都完整的调度一个分组所对应的多个信元.

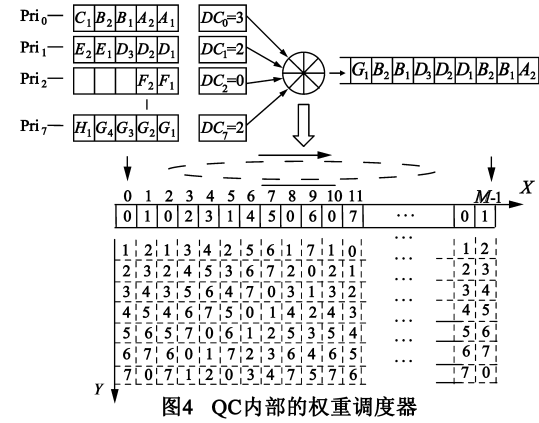

图4　QC内部的权重调度器

本设计对传统的 WRR 算法做出了改进,使其可以支持变长分组调度,同时可以方便的配置不同优先级所占用的输出带宽.在每个 WRR 调度器中都设置了编号分别为 0 到 M-1 的一组优先级配置寄存器,每个寄存器中写入的是一个对应的优先级编号,另外设计了一个从 0 至 M-1 循环移动的指针,其值用于从 M 个寄存器中选择出对应的寄存器并读出其中存储的优先级编号.如图 4 所示,指针值为 0 时,对应优先级配置寄存器中预先配置的是 0,表示需首先查看 Pri_0 队列中是否有完整的分组,如果有则调度出其队首的分组包含的所有信元,如果该优先级队列为空则根据配置,按照从 Pri_1-Pri_7的顺序依次查看.当指针值为 M-1 时,优先级配置寄存器中存储的为 1,表示首先查看 Pri_1 队列中是否有完整的分组,完成调度后,指针值又恢复为 0,开始新的调度操作.

此方案中,可以通过修改优先级配置寄存器中某个优先级出现的次数配置其占用的带宽.例如在 M 个寄存器中,Pri_0 出现了 K 次,那么调度器为其分配的带宽为总带宽的 K/M.由于每个队列首部的分组长度不同,同样一次调度,可能分组的长度差别很大,就单次调度来说存在着带宽分配的不公平性.考虑到整个交换结构采用共享存储方式,分组存储在片外容量较大的主存储区中,从统计来看,可以认为不同优先级中业务流的分布特征相同,此时这种权重带宽分配方式是统计准确的.

这种 WRR 调度机制具有实现简单,配置灵活的特点,由于每次完整的调度一个分组的所有信元,降低了输出端口重组操作的复杂度,有效降低了资源消耗.

4.2　RR 调度模块的设计

除了一个输出端口内部的 WRR 调度器外,为了实现对 N 个输出端口的公平调度和对片外存储区的公平

访问,如图 1 所示,本方案中设计了两个调度器:输出端口调度器和片外存储器读写访问调度器,二者均为基于信元的 RR 调度器.

对于多个输出端口来说,每次只能有一个输出端口访问片外存储区读取数据.如果各个端口具有相同的输出带宽,那么调度器应为各个端口之间提供公平的调度服务,即平均分配其输出调度带宽,因此应该考虑使用 RR 调度器对各个端口进行轮询调度.

在获得调度器的输出许可后,存在两个选择:一次调度过程读出一个完整的分组还是读出一个信元.由于一个端口内部的 WRR 调度器是基于分组进行调度的,其已经保证了输出端口可以方便的进行分组重组,因此输出端口调度器采用基于信元的工作方式更有利于使每个端口获得公平服务的同时减少输出端口内部重组缓冲区的深度.

另外,共享存储结构中所有信元共用片外 SRAM,片外存储器的读写方式直接关系到交换结构的交换性能.为了改进 SRAM 读写访问的带宽均匀性,本方案中采用了基于信元的读写访问 RR 调度器.当同时存在对片外存储器的写入和读出请求时,RR 调度器首先查看写入请求并给予应答.当前等待写入的分组中的一个信元被写入到片外存储器后,无论是否已经写入完整的分组,调度器都将轮询输出端口是否有数据需要读出.这种基于信元的公平轮询方式可以最大限度的保证对片外存储器读写访问的公平性,减少等待读写访问时所需内部数据缓冲区的容量.

将一个端口内部基于分组的变长 WRR 调度器、端口之间的 RR 调度器和片外存储访问的 RR 调度器有效结合起来,可以在尽量保证优先级调度需求和公平性的前提下最大限度的减少硬件逻辑资源消耗.

5 仿真结果与分析

本交换结构设计实现时选用的是 Xilinx 的 V4sx55 FPGA,开发环境采用的是 Xilinx 集成开发环境 ISE13.1,电路核心模块采用 Verilog HDL 编程实现,仿真工具采用的是 ModelSim SE10.0a,下面给出的是关键电路仿真结果.

5.1 片外存储的 RR 调度算法仿真

片外 SRAM 中写入与读出信元时,采用的是基于信元的 RR 调度方式.图 5 是片外 SRAM 读写仿真示意图.图中①为向交换结构写入长度为 256 的数据分组的仿真波形,②是将输入分组划分为 4 个内部信元并写入到片外存储区的过程,由于此时队列中没有完整的分组,所以没有从片外读入并输出信元的操作.当完整的分组写入后,由于同时存在着分组写入和读出的请求,在仲裁器的调度下,对 SRAM 的读写交替进行,如③所示.

5.2 WRR 调度算法仿真

图 6 和图 7 是端口内部 WRR 调度算法仿真工作波形.仿真时设置了 32 个权重配置寄存器,其写入的值分别为:{0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7},即在各个优先级之间平均分配调度带宽.仿真时,在交换结构的输入端连续输入 500 个长度为两个信元、优先级随机均匀分布的变长分组,观察队列深度计数器(depth0-depth7)的数值变化.可以看出各个队列清空时间近似相同并随机分布.图 7 是按照固定优先级配置调度器时的队列深度变化仿真结果,各个队列按照优先级由高到底依次清空(0 为最高优先级).

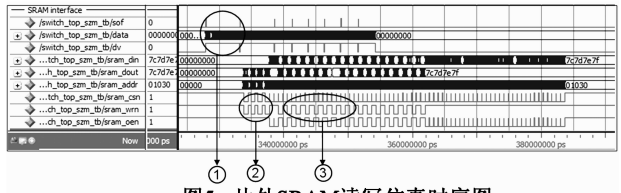

图5 片外SRAM读写仿真时序图

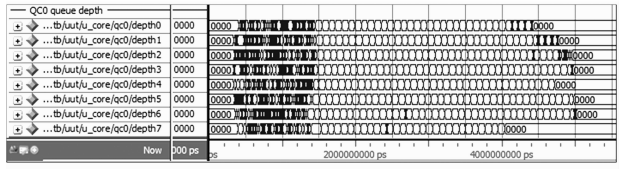

图6 按照相同优先级调度时的仿真波形

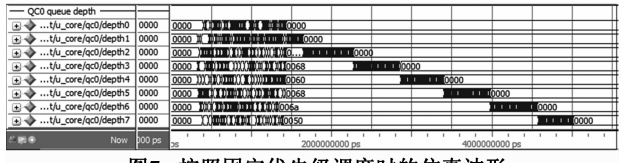

图7 按照固定优先级调度时的仿真波形

5.3 综合结果分析

共享存储交换结构在 Xilinx 的 V4sx55 FPGA 上进行了实现,使用 Xilinx XST 对其进行综合,得出共享存储交换结构的综合结果,如表 1 所示.一个 8 端口交换单元共占用 3361 个 Slice 和 82 块内部 Block RAM,所有关键源消耗量都低于 FPGA 总资源量的 1/3,可以支持完全的三模冗余设计,从而满足星载设备对可靠性的要求.

整个交换结构的数据吞吐率主要受限于片外存储访问带宽.当 SRAM 数据位宽为 64,时钟频率为 100MHz(两个时钟周期一次片外存储访问)时,SRAM 的读写带宽可以达到 3.2Gbps,此时的交换结构峰值吞吐率可以达到 1.6Gbps.在采用 133MHz、位宽为 64 的 DDR(Double Data Rate)存储器时,DDR 读写峰值带宽大于 8.5Gbps,此时交换结构的峰值吞吐率可以达到 4.25Gbps(此处

考虑了 DDR 的刷新操作等带来的影响),通过多级扩展,其可用于 10Gbps 以上的交换机设计需求.

表 1　共享存储交换结构综合结果报告

Device Utilization Summary (estimated values)			
Logic Utilization	Used	Available	Utilization
Number of Slices	3361	24576	13%
Number of Slice Flip Flops	3013	49152	6%
Number of 4 input LUTs	5982	49152	12%
Number of bonded IOBs	626	640	97%
Number of FIFO16/RAMB16s	82	320	25%
Number of GCLKs	1	32	3%

目前思科公司的 Cisco 18400 空间路由器的吞吐率为 250Mbps,本设计所实现的交换接口可以满足吉比特级星载路由器的处理需求.

6　结论

本文设计实现了可应用于星载 IP 交换机的共享存储交换结构,通过对多个调度器的优化设计和有效组合,其支持多优先级队列、可编程权重配置和变长输出调度,能有效降低星上硬件资源消耗,提高系统效率和调度的灵活性.整个电路在 Xilinx 的 V4sx55 FPGA 上实现,占用资源少,可支持完全三模冗余设计,可以满足吉比特级星载路由器的设计需求.

参考文献

[1] Tonguz O K, Sunil Maloo. Internet access via LEO satellite networks: TCP/IP or ATM? [A]. Global Telecommunications Conference[C]. Piscataway: IEEE, 1999. 301 – 305.

[2] Ors T, Sun Z, Evans B G. A meshed VSAT satellite network architecture using an on-board ATM switch[A]. IEEE International Conference on Performance, Computing, and Communications[C]. Piscataway: IEEE, 1997. 208 – 214.

[3] Buster D. Towards IP for space-based communications systems: a CISCO systems assessment of a single board router [A]. Military Communications Conference[C]. Piscataway: IEEE, 2005. 1 – 7.

[4] 徐宁,余少华,汪学舜.一种新型的负载均衡 – 交叉点缓冲交换结构[J].电子学报,2012,40(12):2360 – 2366.
Xu Ning, Yu Shao-hua, Wang Xue-shun. A new type of load-balanced crosspoint-queued switch fabric[J]. Acta Electronica Sinica, 2012, 40(12): 2360 – 2366. (in Chinese)

[5] 戴艺,苏金树,孙志刚.高性能新型交换结构综述[J].电子学报,2010,38(10):2389 – 2399.
Dai Yi, Su Jin-shu, Sun Zhi-gang. A survey on high performance switch architecture[J]. Acta Electronica Sinica, 2010, 38(10): 2389 – 2399. (in Chinese)

[6] 徐扬,唐毅,等.针对高速交换结构的广义极大匹配调度算法[J].电子学报,2007,35(10):1809 – 1816.
Xu Yang, Tang Yi, et al. Extended maximal matching algorithm in high-speed switches[J]. Acta Electronica Sinica, 2007, 35(10): 1809 – 1816. (in Chinese)

[7] 伊鹏,汪斌强,郭云飞,李挥.一种可提供 QoS 保障的新型交换结构[J].电子学报,2007,35(7):1257 – 1263.
Yi Peng, Wang Bin-qiang, Guo Yun-fei, Li Hui. Providing QoS guarantees in a novel switch architecture[J]. Acta Electronica Sinica, 2007, 35(7): 1257 – 1263. (in Chinese)

[8] Choudhury A K, Hahne E L. A new buffer management scheme for hierarchical shared memory switches[J]. IEEE Networking, 1997, 5(5): 728 – 738.

[9] Jong-Seon Kim, Lee D C. Weighted round robin packet scheduler using relative service share[A]. Military Communications Conference[C]. Piscataway: IEEE, 2001. 988 – 992.

[10] Ito Y, Tasaka S. Variably weighted round robin queueing for core IP routers[A]. IEEE International Conference on Performance, Computing and Communications[C]. Piscataway: IEEE, 2002. 159 – 166.

[11] Zhang Y, Harrison P G. Performance of a priority-weighted round robin mechanism for differentiated service networks [A]. International Conference on Computer Communications and Networks[C]. Piscataway: IEEE, 2007. 1198 – 1203.

作者简介

沈泽民　男,1987 年 11 月生于江苏盐城.2013 年毕业于解放军理工大学军事装备学专业,主要从事高性能交换结构相关研究工作.
E-mail: szm198711@gmail.com

乔庐峰　男,1971 年 11 月生于河南南乐,副教授、硕士生导师,长期从事通信和计算机网络中关键芯片和电路技术研究,发表论文 20 余篇,获部级科研奖励 4 项.